

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Faculdade de Engenharia e Ciências- campus de Guaratinguetá

ELEN YANINA AGUIRRE RODRÍGUEZ

***DATA SCIENCE E MACHINE LEARNING* APLICADO NA MORTALIDADE
INFANTIL NA REGIÃO METROPOLITANA DO VALE DO PARAÍBA E LITORAL
NORTE**

Guaratinguetá

2025

ELEN YANINA AGUIRRE RODRÍGUEZ

***DATA SCIENCE E MACHINE LEARNING* APLICADO NA MORTALIDADE
INFANTIL NA REGIÃO METROPOLITANA DO VALE DO PARAÍBA E LITORAL
NORTE**

Tese, apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Engenharia e Ciências, Guaratinguetá, para obtenção do título de Doutora em Engenharia.

Área de Concentração: Produção – Otimização, simulação e tomada de decisão

Orientador: Profº Dr. Fernando Augusto Silva Marins

Coorientador:

Profº Dr. Luiz Fernando Costa Nascimento

Profº Dr. Aneirson Francisco da Silva

Guaratinguetá

2025

R696d	Rodríguez, Elen Yanina Aguirre <i>Data science e machine learning</i> aplicado na mortalidade infantil na região metropolitana do Vale do Paraíba e Litoral Norte / Elen Yanina Aguirre Rodríguez - Guaratinguetá, 2025. 261 f : il. Bibliografia: f. 169-183 Tese (Doutorado) – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2025. Orientador: Prof. Dr. Fernando Augusto Silva Marins Coorientador: Prof. Dr. Luiz Fernando Costa Nascimento Coorientador: Prof. Dr. Aneirson Francisco da Silva 1. Aprendizado do computador. 2. Armazenamento de dados. 3. Análise espacial (Estatística). 4. Mortalidade infantil. I. Título. CDU 681.3.07(043)
-------	--

Luciana Máximo
Bibliotecária/CRB-8 3595

IMPACTO POTENCIAL DESTA PESQUISA

Este trabalho apresenta um impacto significativo tanto no âmbito científico quanto social, contribuindo diretamente para o aprimoramento das políticas públicas voltadas à redução da mortalidade infantil. O uso de Sistemas de Informação Geográfica (SIG), análise espacial e a aplicação da ciência de dados, com o desenvolvimento de modelos de predição e previsão baseados em técnicas de Aprendizado de Máquina (*Machine Learning*), permitiu a criação de ferramentas robustas para subsidiar a tomada de decisão em diferentes níveis de planejamento: estratégico, tático e operacional.

No nível estratégico, a análise espacial oferece subsídios essenciais para a formulação de políticas públicas de longo prazo, possibilitando a identificação de desigualdades locais que impactam diretamente os indicadores de saúde materno-infantil. No nível tático, os modelos baseados em séries temporais permitem prever o número mensal de óbitos, facilitando a distribuição eficiente de recursos e a implementação de medidas preventivas em períodos críticos. Já no nível operacional, os modelos de classificação, tanto binários quanto multiclasse, viabilizam a identificação de recém-nascidos com alto risco de mortalidade, permitindo intervenções mais precisas e direcionadas.

Além disso, as metodologias propostas são passíveis de replicação e adaptação a diferentes contextos regionais, garantindo a escalabilidade e continuidade das estratégias preditivas e preventivas no âmbito da saúde pública.

Adicionalmente, o impacto social desta pesquisa transcende os limites técnicos e acadêmicos, refletindo diretamente na qualidade de vida das populações mais vulneráveis, ao garantir cuidados mais eficientes e direcionados aos recém-nascidos em risco. Assim, esta pesquisa está diretamente alinhada com os Objetivos de Desenvolvimento Sustentável (ODS), especialmente com o ODS 3 (Saúde e Bem-estar), ao contribuir para a redução da mortalidade infantil e melhorar o acesso a cuidados médicos de qualidade.

POTENTIAL IMPACT OF THIS RESEARCH

This study has a significant impact both in scientific and social contexts, directly contributing to the enhancement of public policies aimed at reducing infant mortality. The use of Geographic Information Systems (GIS), spatial analysis, and data science applications, with the development of predictive and forecasting models based on Machine Learning techniques, has enabled the creation of robust tools to support decision-making at different levels of planning: strategic, tactical, and operational.

At the strategic level, spatial analysis provides essential information for formulating long-term public policies, enabling the identification of local inequalities that directly impact maternal and child health indicators. At the tactical level, time-series-based models allow for the prediction of the monthly number of deaths, facilitating the efficient distribution of resources and the implementation of preventive measures. At the operational level, classification models, both binary and multiclass, enable the identification of newborns at high risk of mortality, allowing for more precise and targeted interventions.

Moreover, the proposed methodologies are replicable and adaptable to different regional contexts, ensuring the scalability and continuity of predictive and preventive strategies in public health. Additionally, the social impact of this research goes beyond technical and academic boundaries, directly reflecting on the quality of life of vulnerable populations by ensuring more efficient and targeted care for at-risk newborns. Thus, this research is directly aligned with the Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being), by contributing to the reduction of infant mortality and improving access to quality medical care.

ELEN YANINA AGUIRRE RODRÍGUEZ

***DATA SCIENCE E MACHINE LEARNING* APLICADO NA MORTALIDADE
INFANTIL NA REGIÃO METROPOLITANA DO VALE DO PARAÍBA E LITORAL
NORTE**

Tese apresentado(a) à Universidade Estadual Paulista (UNESP), Faculdade de Engenharia e Ciências, Guaratinguetá, para obtenção do título de Doutora em Engenharia.

Data de defesa: 06/02/2025

BANCA EXAMINADORA:

Documento assinado digitalmente



FERNANDO AUGUSTO SILVA MARINS

Data: 06/02/2025 19:46:54-0300

Verifique em <https://validar.iti.gov.br>

Profº Dr. Fernando Augusto Silva Marins
Orientador - UNESP

Documento assinado digitalmente



ADRIANA DE OLIVEIRA RIBEIRO DOS SANTOS

Data: 07/02/2025 16:16:12-0300

Verifique em <https://validar.iti.gov.br>

Profº Dra. Adriana de Oliveira Ribeiro dos Santos
UNITAU

Documento assinado digitalmente



FABRICIO MACIEL GOMES

Data: 10/02/2025 10:42:28-0300

Verifique em <https://validar.iti.gov.br>

Profº Dr. Fabrício Maciel Gomes
USP

Documento assinado digitalmente



MARCELA APARECIDA GUERREIRO MACHADO DE

Data: 10/02/2025 15:28:57-0300

Verifique em <https://validar.iti.gov.br>

Profº Dra. Marcela Aparecida Guerreiro Machado de Freitas
UNESP

Documento assinado digitalmente



PALOMA MARIA SILVA ROCHA RIZOL

Data: 10/02/2025 09:10:34-0300

Verifique em <https://validar.iti.gov.br>

Profº Dra. Paloma Maria Silva Rocha Rizol
UNESP

Dedico este trabalho com amor ao meu companheiro de vida, Alexander, por seu amor e apoio incondicional; e aos meus pais, Edita e Elias, cuja sabedoria e amor sempre são minha fonte de inspiração e força.

AGRADECIMENTOS

A Deus, pela vida.

A meus pais, Edita e Elias, e a meus irmãos, Richard, Cristina, Elias e Christian, pelo amor, cuidado e por me apoiarem incondicionalmente ao longo de minha trajetória de vida – pessoal, acadêmica e profissional.

Ao meu esposo, Alexander, pelo amor, companheirismo, paciência, motivação e apoio constantes em cada momento de minha vida.

Ao meu orientador, o professor Dr. Fernando Augusto Silva Marins, pela gentileza, preocupação, carinho e por me motivar a continuar o doutorado. Ao meu coorientador, Prof. Dr. Luiz Fernando Costa Nascimento, pela confiança, por acreditar em mim, pelo carinho e pelas conversas que sempre guardarão um lugar especial em meu coração. Ambos desempenharam um papel fundamental nesta etapa de minha vida, e o apoio constante deles foi essencial para o desenvolvimento desta pesquisa.

Ao meu coorientador, Prof. Dr. Aneirson Francisco da Silva, pela paciência e orientação durante o desenvolvimento deste trabalho.

À Profa. Dra. Adriana de Oliveira Ribeiro dos Santos, Profa. Dra. Marcela Aparecida Guerreiro Machado de Freitas, Profa. Dra. Paloma Maria Silva Rocha Rizol e ao Prof. Dr. Rafael de Carvalho Miranda, pela contribuição e pela participação como membros da banca examinadora.

Aos professores e funcionários da FEG-UNESP, do Programa de Pós-Graduação em Engenharias, pela contribuição essencial em minha formação acadêmica.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro para o desenvolvimento deste trabalho.

Este trabalho contou com o apoio da seguinte entidade:

CAPES - Coordenação de Aperfeiçoamento de Pessoa de Nível Superior, código de financiamento 001.

*“Só podemos ver uma curta distância pela frente,
mas podemos ver muito aí que precisa ser feito.”*

*[« We can only see a short distance ahead, but we can see
plenty there that needs to be done.»]*

Alan M. Turing, ‘Computing machinery and Intelligence’, Mind, 1950.

RESUMO

A mortalidade infantil reflete a saúde pública e a qualidade da assistência oferecida à mãe e ao recém-nascido. Com o objetivo de reduzir essa mortalidade e outros problemas de saúde, tem-se observado um crescente interesse em utilizar ferramentas como Ciência de Dados, Aprendizado de Máquina (ML), Sistemas de Informação Geográfica (SIG) e análise espacial, cuja integração pode aprimorar os processos e apoiar decisões. Este estudo teve como objetivo explorar a aplicabilidade dessas ferramentas para abordar a mortalidade infantil durante o primeiro ano de vida na Região Metropolitana do Vale do Paraíba e Litoral Norte (RMVPLN), no Estado de São Paulo. A metodologia adotada incluiu a utilização de dados do Departamento de Informática do Sistema Único de Saúde (DATASUS) e a aplicação de análise exploratória de dados, análise espacial e desenvolvimento de modelos de previsão e predição, com propostas de soluções para a tomada de decisões em diferentes níveis: estratégico, tático e operacional. No nível tático, utilizou-se os dados de 2010 a 2020, considerando as taxas de mortalidade infantil (TMI), neonatal (TMN) e pós-neonatal, além de determinantes associados. Foi utilizado o índice univariado e bivariado de Moran para medir o grau de associação espacial na RMVPLN. No nível estratégico, foram aplicadas técnicas de regressão de ML para modelos de previsão, treinados com dados de 2010 a 2020, da ocorrência mensal de óbitos neonatais, pós-neonatais e infantis, sendo validados com dados de 2021. No nível operacional, treinaram-se modelos binários e multiclasse, utilizando técnicas de classificação, com dados rotulados de 2010 a 2019, validados com dados de 2020. Em ambos os casos, o desempenho dos modelos foi avaliado por meio de métricas de erro e testes estatísticos específicos para cada abordagem, garantindo a confiabilidade, robustez e aplicabilidade prática dos resultados obtidos. Os resultados mostraram que a redução da TMI foi principalmente atribuída à diminuição dos óbitos pós-neonatais. A análise espacial revelou agrupamentos de alto risco na região Norte Central para a TMI e TMN. Os modelos de ML para previsão de óbitos apresentaram boa precisão, sendo os melhores modelos: RF ($lag = 2$, $RMSE = 6,03$) para a mortalidade infantil, GB ($lag = 2$, $RMSE = 4,16$) para a neonatal e SVM ($lag = 3$, $RMSE = 3,162$) para a pós-neonatal. Para a identificação de risco de óbito, os modelos treinados tiveram acurácia e AUC superior a 90% e 91%, respectivamente, para os modelos binários, e acima de 81% e 82% para os modelos multiclasse. Os melhores modelos foram os treinados com SVM (acurácia = 92,88%, $AUC = 0,934$) para o modelo binário, e com XGB (acurácia = 88,38%, $AUC = 0,876$) para o modelo multiclasse. Conclui-se que a utilização de SIG, análise espacial e ML pode melhorar significativamente o cuidado do nascido vivo, alinhando-se aos Objetivos de Desenvolvimento Sustentável. Além disso, as sistemáticas propostas oferecem benefícios substanciais para a tomada de decisões, permitindo otimizar o uso de recursos e intervenções, além de contribuir para a melhoria da qualidade do cuidado ao recém-nascido e garantir sua sobrevivência.

Palavras-Chave: Ciência de dados; Aprendizado de máquina; Análise espacial; Mortalidade infantil; Mortalidade neonatal; Mortalidade Pós-neonatal.

ABSTRACT

Infant mortality reflects public health and the quality of care provided to mothers and newborns. To reduce this mortality and other health issues, there has been a growing interest in utilizing tools such as Data Science, Machine Learning (ML), Geographic Information Systems (GIS), and spatial analysis, as their integration can enhance processes and support decision-making. This study aimed to explore the applicability of these tools to address infant mortality during the first year of life in the Metropolitan Region of Vale do Paraíba and Northern Coast (RMVPLN), in the State of São Paulo. The methodology adopted included the use of data from the Department of Informatics of the Unified Health System (DATASUS) and the application of exploratory data analysis, spatial analysis, and the development of prediction and forecasting models, with proposed solutions for decision-making at different levels: strategic, tactical, and operational. At the tactical level, data from 2010 to 2020 were used, considering the infant mortality rate (IMR), neonatal mortality rate (NMR), and post-neonatal mortality rate (PNMR), as well as associated determinants. The univariate and bivariate Moran index were used to measure the degree of spatial association in RMVPLN. At the strategic level, ML regression techniques were applied to forecasting models, trained with data from 2010 to 2020 on the monthly occurrence of neonatal, post-neonatal, and infant deaths, and validated with 2021 data. At the operational level, binary and multiclass models were trained using classification techniques, with labeled data from 2010 to 2019, validated with 2020 data. In both cases, the performance of the models was evaluated using error metrics and specific statistical tests for each approach, ensuring the reliability, robustness, and practical applicability of the obtained results. The results showed that the reduction in IMR was mainly attributed to the decrease in post-neonatal deaths. Spatial analysis revealed high-risk clusters in the Northern Central region for both IMR and NMR. The ML models for death prediction demonstrated good accuracy, with the best models being: RF ($lag = 2$, RMSE = 6.03) for infant mortality, GB ($lag = 2$, RMSE = 4.16) for neonatal mortality, and SVM ($lag = 3$, RMSE = 3.162) for post-neonatal mortality. For the identification of death risk, the trained models achieved accuracy and AUC greater than 90% and 91%, respectively, for binary models, and above 81% and 82% for multiclass models. The best models were those trained with SVM (accuracy = 92.88%, AUC = 0.934) for the binary model, and XGB (accuracy = 88.38%, AUC = 0.876) for the multiclass model. It is concluded that the use of GIS, spatial analysis, and ML can significantly improve live birth care, aligning with the Sustainable Development Goals. Furthermore, the proposed methodologies offer substantial benefits for decision-making, allowing for the optimization of resource use and interventions, as well as contributing to the improvement of newborn care quality and ensuring their survival.

Keywords: Data Science; Machine Learning; Spatial Analysis; Infant Mortality; Neonatal Mortality; Post-neonatal Mortality.

LISTA DE ILUSTRAÇÕES

Figura 1 – Taxa de mortalidade infantil, neonatal e pós-neonatal no Brasil, São Paulo e RMVPLN	28
Figura 2 – Principais áreas de aplicação e soluções dos SIG e ML no cuidado do recém-nascido	34
Figura 3 – Tendência de publicações e citações do análise espacial na mortalidade do nascido vivo, período 1987 – 2024	37
Figura 4 – Produção dos países avaliada pelo autor correspondente do análise espacial na mortalidade do nascido vivo, período 1987 – 2024	37
Figura 5 – Rede de colaboração internacional dos artigos sobre análise espacial na mortalidade do nascido vivo, período 1987 – 2024	38
Figura 6 – Tendência de publicações e citações do aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024	41
Figura 7 – Produção dos países avaliada pelo autor correspondente do aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024	42
Figura 8 – Rede de colaboração internacional dos artigos sobre aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024	42
Figura 9 – Tendência de publicações e citações do aprendizado de máquina para previsão na mortalidade do nascido vivo, período 2022 – 2024	43
Figura 10 – Produção dos países avaliada pelo autor correspondente do aprendizado de máquina para previsão na mortalidade do nascido vivo, período 2022 – 2024	43
Figura 11 – Desempenho dos modelos para identificar o risco de mortalidade do nascido vivo, de acordo com a AUC	48
Figura 12 – Desempenho dos modelos para prever a mortalidade, de acordo com o RMSE	49
Figura 13 – Classificação dos recém-nascidos segundo idade gestacional e peso	51
Figura 14 – Classificação do recém nascido de acordo com peso e idade gestacional	52
Figura 15 – Principais causas de morte neonatal e entre crianças menores de 5 anos em 2017	53
Figura 16 – <i>Big Data</i> e inteligência artificial na área da saúde	55
Figura 17 – Técnicas para pré-processamento de dados	57
Figura 18 – Critério de torre e rainha	62
Figura 19 – Exemplo de matriz de vizinhos de municípios da RMVPLN	63
Figura 20 – Teste de pseudo-significância estatística	64
Figura 21 – Gráfico de dispersão de Moran	65
Figura 22 – Mapa de Moran	66
Figura 23 – Autocorrelação espacial bivariada	67
Figura 24 – Classificação de técnicas de aprendizado de máquina	72

Figura 25 – Equilíbrio entre interpretabilidade e desempenho de alguns métodos de aprendizado de máquina	74
Figura 26 – Estrutura de regressão logística	75
Figura 27 – Estrutura de uma Árvore de Decisão	76
Figura 28 – Estrutura de máquina de vetores de suporte	81
Figura 29 – Estrutura de rede neural perceptron multicamada	84
Figura 30 – Métodos de otimização de hiperparâmetros	87
Figura 31 – Matriz de confusão	89
Figura 32 – Matriz de McNemar	91
Figura 33 – Processo metodológico de CRISP-DM	93
Figura 34 – Processo metodológico de KDD	94
Figura 35 – Processo metodológico de SEMMA	95
Figura 36 – Região metropolitana do Vale do Paraíba e Litoral Norte	98
Figura 37 – Indicadores sociais da região metropolitana do Vale do Paraíba e Litoral Norte	99
Figura 38 – Sistemas que integram o repositório do DATASUS	100
Figura 39 – Divisão da faixa etária da mortalidade do nascido vivo no primeiro ano de vida	102
Figura 40 – Sequência de soluções propostas com análise especial e aprendizado de máquina	104
Figura 41 – Classificação da pesquisa	105
Figura 42 – Fluxograma das etapas da pesquisa adaptada da metodologia SEMMA . . .	106
Figura 43 – Estruturação dos subconjunto de dados para cada análise	110
Figura 44 – Procedimentos aplicados no pré-processamento das variáveis	116
Figura 45 – Mapa temático da distribuição da (A) Taxa de Mortalidade Infantil (TMI), (B) Taxa de Mortalidade Neonatal (TMN), (C) Taxa de Mortalidade Pós-neonatal (TMPN), período 2010 – 2020	121
Figura 46 – Mapa Moran e gráfico de dispersão da suavização bayesiana da Taxa de Mortalidade Infantil (TMI), período 2010 – 2020	122
Figura 47 – Mapa Moran e gráfico de dispersão da suavização da Taxa de Mortalidade Neonatal (TMN), período 2010 – 2020	123
Figura 48 – Mapa Moran e gráfico de dispersão da suavização bayesiana da Taxa de Mortalidade Pós-neonatal (TMPN), período 2010 – 2020	123
Figura 49 – Intervalo de meses da validação cruzada contínua de 2 divisões	127
Figura 50 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade infantil . .	133
Figura 51 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade neonatal . .	133
Figura 52 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade pós-neonatal	134
Figura 53 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1º divisão da validação cruzada contínua da mortalidade infantil	134

Figura 54 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade infantil	135
Figura 55 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1ª divisão da validação cruzada contínua da mortalidade neonatal	135
Figura 56 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade neonatal	136
Figura 57 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1ª divisão da validação cruzada contínua da mortalidade pós-neonatal	136
Figura 58 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade pós-neonatal	137
Figura 59 – Frequência do tamanho da janela deslizante entre os modelos treinados . . .	137
Figura 60 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade infantil . . .	140
Figura 61 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade neonatal . .	141
Figura 62 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade pós-neonatal	142
Figura 63 – Série temporal original com previsões de cada modelo validado para a mortalidade mensal infantil, neonatal e pós-neonatal	144
Figura 64 – Matriz de correlação de Pearson	146
Figura 65 – Matriz de coeficiente V de Cramér	147
Figura 66 – Importância de variáveis com LASSO e RF para modelo binário	148
Figura 67 – Importância de variáveis com LASSO e RF para modelo multiclasse	149
Figura 68 – Resultados de acurácia, F1-score e AUC obtidos por meio de validação cruzada 5-fold para todas as técnicas de ML no modelo binário	151
Figura 69 – Valores médios de desempenho dos modelos binários	153
Figura 70 – Curvas ROC dos modelos de classificação binária	154
Figura 71 – Resultados de acurácia, F1-score e AUC obtidos por meio de validação cruzada 5-fold para todas as técnicas de ML no modelo multiclasse	156
Figura 72 – Valores médios de desempenho dos modelos multiclasse	158
Figura 73 – Curvas ROC dos modelos de classificação multiclasse	159
Figura 74 – Curvas ROC dos modelos binário e multiclasse na validação	162
Figura 75 – Curvas ROC por classe do modelo multiclasse na validação	162

LISTA DE QUADROS

Quadro 1 – Classificação das publicações brasileiras por faixa etária de óbito . . .	38
Quadro 2 – Informações relevantes dos artigos relacionados à aplicação de aprendi- zado de máquina na mortalidade do nascido vivo	44
Quadro 3 – Classificação da tomada de decisão e planejamento	54
Quadro 4 – Análise combinatória dos testes de estacionariedade da série temporal .	70
Quadro 5 – Variáveis do conjunto de dados do SINASC	101
Quadro 6 – Variáveis do conjunto de dados do SIM	103
Quadro 7 – Configuração de hiperparâmetros dos modelos de aprendizado de máquina	143
Quadro 8 – Configuração de hiperparâmetros dos modelos binário e multiclasse . .	161

LISTA DE TABELAS

Tabela 1 – Número de artigos encontrados na Scopus e PubMed na primeira fase de busca	36
Tabela 2 – Número de artigos encontrados na Scopus e PubMed da segunda fase de busca	40
Tabela 3 – Análise descritiva das taxas de mortalidade e das outras variáveis de análise nos municípios da RMVPLN, no período 2010-2020	119
Tabela 4 – Índice de Moran bivariado entre a mortalidade infantil, neonatal e pós-neonatal com as variáveis selecionadas da RMVPLN, no período 2010-2020	124
Tabela 5 – Estatísticas descritivas da divisão do banco de dados com a validação cruzada contínua de 2 divisões	128
Tabela 6 – Verificação da estacionariedade dos dados de treinamento utilizando o teste de Dickey-Fuller aumentado e o teste de Kwiatkowski-Phillips-Schmidt-Shin	129
Tabela 7 – Modelos ARIMA e SARIMA ajustados para prever o número mensal de óbitos infantis, neonatais e pós-neonatais, 1º divisão - validação cruzada contínua	130
Tabela 8 – Modelos ARIMA e SARIMA ajustados para prever o número mensal de óbitos infantis, neonatais e pós-neonatais, 2º divisão - validação cruzada contínua	131
Tabela 9 – Métricas de desempenho dos modelos ARIMA e SARIMA treinados na 1º divisão da validação cruzada contínua	132
Tabela 10 – Métricas de desempenho dos modelos ARIMA e SARIMA treinados na 2º divisão da validação cruzada contínua	132
Tabela 11 – Métricas de desempenho dos modelos de aprendizado de máquina treinados na 1º divisão da validação cruzada contínua	138
Tabela 12 – Métricas de desempenho dos modelos de aprendizado de máquina treinados na 2º divisão da validação cruzada contínua	139
Tabela 13 – Métricas de desempenho dos modelos com os dados de validação	144
Tabela 14 – Resultado do teste-T pareado	145
Tabela 15 – Distribuição das instâncias nos conjuntos de treino/teste e validação para os modelos binário e multiclasse	150
Tabela 16 – Valores médios de desempenho dos modelos binários	152
Tabela 17 – Eficiência média dos modelos de classificação binária	154
Tabela 18 – Resultados do teste de McNemar dos modelos de classificação binária	156
Tabela 19 – Valores médios de desempenho dos modelos multiclasse	157
Tabela 20 – Eficiência média dos modelos de classificação multiclasse	160
Tabela 21 – Resultados do teste de McNemar dos modelos de classificação multiclasse	160
Tabela 22 – Métricas de desempenho dos modelos com os dados de validação	161

LISTA DE ABREVIATURA E SIGLAS

AA	Alto – Alto
AB	Alto – Baixo
AdaBoost	Adaptive Boosting
ADF	Augmented Dickey–Fuller
AIC	Akaike Information Criterion
AIG	Adequado para a Idade Gestacional
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
AUC	Área sob a Curva, em inglês <i>Area Under the Curve</i>
AvA	All-vs-All
BA	Baixo – Alto
BB	Baixo – Baixo
BD	<i>Big Data</i>
Bnet	Bayesian network
CART	Classification and Regression Tree
catboost	Categorical Boosting
CBO	Classificação Brasileira de Ocupações
CDC	Conjunto de Dados Completo
CDS	Conjunto de Dados Simplificado
CHAID	Chi-squared Detection of Automatic Interaction
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CRISP-DM	Processo Padrão Inter-Indústrias para Mineração de Dados
DATASUS	Departamento de Informática do Sistema Único de Saúde
DT	Decision tree
DNN	Deep Neural Network
GB	Gradient Boosting
GIG	Grande para a Idade Gestacional
GLMNET	Generalized Linear Models and Elastic Nets

GPC	Gaussian Processes Classifier
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IPEA	Instituto de Pesquisa Econômica Aplicada
IPRS	Índice Paulista de Responsabilidade Social
IQR	Intervalo Interquartilico, em inglês <i>Interquartile Range</i>
IVS	Índice de Vulnerabilidade Social
KDD	Descoberta de Conhecimento em Bases de Dados
KNN	K-Nearest Neighbors
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LGBM	Light Gradient Boosting Machine
LISA	Indicador Local para Autocorrelação Espacial
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Aprendizado de Máquina, em inglês <i>Machine Learning</i>
MLP	Multilayer Perceptron
MPD	Mean Poisson Deviance
MSE	Mean Absolute Error
NB	Naive Bayes
ODM	Objetivos de Desenvolvimento do Milênio
ODS	Objetivos de Desenvolvimento Sustentável
OMS	Organização Mundial de Saúde
OvA	One-vs-All
PCA	Análise de Componentes Principais, em inglês <i>Principal Component Analysis</i>
PIB	Produto Interno Bruto
PIG	Pequeno para a Idade Gestacional
ReLU	Rectified Linear Unit
RF	Random Forest

RLog	Regressão Logística
RMSE	Root Mean Squared Error
RMVPLN	Região Metropolitana do Vale do Paraíba e Litoral Norte
ROC	Característica Operativa do Receptor, em inglês <i>Receiver Operating Characteristic</i>
RSS	Soma Residual de Quadrados, em inglês <i>Residual Sum of Squares</i>
SARIMA	Seasonal Autoregressive Integrated Moving Average
SEADE	Fundação Sistema Estadual de Análise de Dados
SEMMA	Amostra, Explorar, Modificar, Modelar e Avaliar
SIG	Sistemas de Informação Geográfica
SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informações Sobre Nascidos Vivos
SL	Super Learner
SMAPE	Symmetric Mean Absolute Percentage Error
SVM	Support Vector Machine
Tanh	Tangente Hiperbólica
TFP	Taxa de Falsos Positivos
TICs	Tecnologias de Informação e Comunicação
TMI	Taxa de Mortalidade Infantil
TMN	Taxa de Mortalidade Neonatal
TMPN	Taxa de Mortalidade Pós-neonatal
TVP	Taxa de Verdadeiros Positivos
XGB	eXtreme Gradient Boosting

LISTA DE SÍMBOLOS

$\mathbf{X}$	Matriz de variáveis.
x_j	Variável j .
x_i	Instância i .
$x_{i,j}$	Característica da linha i e variável j .
n	Número de linhas ou instâncias.
$\bar{\mathbf{x}}_j$	Média <i>média</i> ($\mathbf{x}_j$).
$Q2$	Mediana <i>mediana</i> ($\mathbf{x}_j$), segundo quartil.
$Q1$	Primeiro quartil.
$Q3$	Terceiro quartil.
$\min(x_i)$	Valor mínimo: <i>mínimo</i> ($\mathbf{x}_j$).
$\max(x_i)$	Valor máximo: <i>máximo</i> ($\mathbf{x}_j$).
S_j^2	Variância: <i>Variância</i> ($\mathbf{x}_j$).
S_j	Desvio padrão: <i>Desvio</i> ($\mathbf{x}_j$).
$Covariância(\mathbf{x}_j, \mathbf{x}_k)$	Covariância.
r	Correlação de Pearson.
r_s	Correlação de Spearman.
$\mathbf{W}$	Matriz de contiguidade.
$w_{i,j}$	Pesos espaciais.
$E(I_U)$	Valor esperado de I_U .
$Z(I_U)$	Pressuposição de normalidade de I_U .
$DP(I_U)$	Desvio padrão de I_U .
I_U	Índice de Moran global univariado.

I_{U_i}	Índice de Moran local univariado.
$lag(x)$	Média ponderada dos valores vizinhos normalizados.
I_B	Índice de Moran global bivariado.
I_{B_i}	Índice de Moran local bivariado.
$\hat{\theta}_i$	Suavização bayesiana empírico.
m_i	Taxa média da área analisada.
C_i	Constante relacionada ao tamanho da população, varia entre 0 – 1.
$x'_{i,j}$	Normalização Min – Max.
$x'_{min,j}$	Valor mínimo da variável j .
$x'_{max,j}$	Valor máximo da variável j .
$x''_{i,j}$	Padronização.
y^t	Observações da série temporal.
t	Instante no tempo.
y_t	Observação ou previsão no instante t .
T	Duração de tempo do período considerado na série temporal.
$E(y_t)$	Propriedades da média da série.
$var(y_t)$	Propriedades da variância da série.
γ_k	Propriedades da autocovariância da série.
Δy_t	Primeira diferença.
ϕ_p	Parâmetros não sazonais do modelo autoregressivo.
ε_t	Erro aleatório ou ruído branco.
d	Diferenciação não sazonal.
θ_q	Parâmetros não sazonais do modelo de média móveis.
Φ_p	Parâmetros sazonais do modelo autoregressivo.
Θ_q	Parâmetros sazonais do modelo de média móveis.

D	Diferenciação sazonal.
$\mathbf{x}$	Matriz de variáveis de entrada.
$\mathbf{y}$	Vetor de rótulos o dados de saída.
$f(.)$	Função desconhecida.
y_i	Valor real.
$\hat{f}(.)$	Função estimada ou estimador.
$\hat{y}_i$	Valor estimado.
$\mathbf{T}$	Dados de treinamento.
f^*	Estimador com os hiperparâmetros ótimos.
$\mathcal{L}(.)$	Função de custo.
TMI	Taxa de Mortalidade Infantil
TMN	Taxa de Mortalidade Neonatal
$TMPN$	Taxa de Mortalidade Pós-neonatal
$Prevalência_x$	Prevalência de casos
$Taxa_{leitos}$	Taxa de leitos pediátricos
$Taxa_{médicos}$	Taxa de médicos
G	Índice de Gini
D	Entropia

SUMÁRIO

1	INTRODUÇÃO	27
1.1	CONTEXTUALIZAÇÃO DA PESQUISA	27
1.2	QUESTÕES DE PESQUISA E OBJETIVOS	32
1.2.1	Questões de Pesquisa	32
1.2.2	Objetivo Geral e Específicos	32
1.3	JUSTIFICATIVA E IMPORTÂNCIA	33
1.4	DELIMITAÇÃO DA PESQUISA	50
1.5	ORGANIZAÇÃO DO TEXTO	50
2	FUNDAMENTAÇÃO TEÓRICA	51
2.1	RECÉM-NASCIDO E SUA VULNERABILIDADE	51
2.2	TOMADA DE DECISÃO, <i>BIG DATA</i> E INTELIGÊNCIA ARTIFICIAL NA ÁREA DE SAÚDE	54
2.3	PRÉ-PROCESSAMENTO DE DADOS	56
2.3.1	Limpeza de Dados	57
2.3.2	Integração de Dados	59
2.3.3	Seleção de Variáveis e Balanceamento	59
2.3.4	Transformação de Dados	60
2.4	TÉCNICAS DE ANÁLISE ESPACIAL	62
2.4.1	Autocorrelação Espacial Univariada	63
2.4.2	Autocorrelação Espacial Bivariada	66
2.4.3	Suavização Bayesiana Empírico	68
2.5	TÉCNICAS DE SÉRIES TEMPORAIS	68
2.6	TÉCNICAS DE APRENDIZADO DE MÁQUINA	71
2.6.1	Regressão Logística	75
2.6.2	Métodos Baseados em Árvores de Decisão	76
2.6.3	Máquinas de Vetores de Suporte	81
2.6.4	Rede Neural Perceptron Multicamada	84
2.7	OTIMIZAÇÃO DE HIPERPARÂMETROS E AVALIAÇÃO DOS MODELOS	86
2.7.1	Otimização de Hiperparâmetros	86
2.7.2	Métricas de Desempenho	88
2.8	PROCESSOS METODOLÓGICOS EM APRENDIZADO DE MÁQUINA	92
2.8.1	Processo Padrão Inter-Indústrias para Mineração de Dados	92
2.8.2	Descoberta de Conhecimento em Bases de Dados	93
2.8.3	Amostra, Explorar, Modificar, Modelar e Avaliar	94

3	MATERIAIS E MÉTODOS	97
3.1	LOCAL DE ESTUDO	97
3.2	BANCO DE DADOS E MALHA CARTOGRÁFICA	100
3.3	PROCEDIMENTO METODOLÓGICO	104
3.4	CONSIDERAÇÕES ÉTICAS	118
4	RESULTADOS E DISCUSSÃO	119
4.1	ANÁLISE ESPACIAL DAS TAXAS DE MORTALIDADE	119
4.1.1	Mapeamento das taxas de mortalidade e determinantes	119
4.1.2	Autocorrelação espacial univariado e bivariado	121
4.2	MODELO PARA PREVISÃO DA MORTALIDADE COM SÉRIE DE TEMPO	127
4.2.1	Avaliação dos modelos clássicos e de aprendizado de máquina	127
4.2.1.1	Modelos clássicos	128
4.2.1.2	Modelos de aprendizado de máquina	132
4.2.1.3	Comparação de modelos com métricas de avaliação	140
4.2.2	Validação dos resultados	143
4.3	MODELO PARA IDENTIFICAÇÃO DE RISCO DE ÓBITO	145
4.3.1	Seleção e importância das características	146
4.3.2	Avaliação do desempenho dos modelos de classificação	151
4.3.3	Validação dos resultados	161
5	CONCLUSÕES E RECOMENDAÇÕES PARA PESQUISAS FUTURAS	165
5.1	CONCLUSÕES E VERIFICAÇÃO DAS QUESTÕES E OBJETIVOS . . .	165
5.2	PROPOSTA PARA FUTURAS PESQUISAS	169
	REFERÊNCIAS	170
	APÊNDICE A – DISTRIBUIÇÃO DAS TAXAS DE MORTALIDADE	
	INFANTIL GLOBAL	185
	APÊNDICE B – TENDÊNCIA DA ANÁLISE ESPACIAL E APREN-	
	DIZADO DE MÁQUINA NO CUIDADO DA SAÚDE	186
	APÊNDICE C – CONCEITOS E TÉCNICAS PARA ANÁLISE EX-	
	PLORATÓRIA DE DADOS	189
	APÊNDICE D – CÓDIGO PARA CONVERSÃO DE ARQUIVOS DBC	
	PARA CSV	197
	APÊNDICE E – SCRIPTS PARA VERIFICAÇÃO, CONSOLIDAÇÃO	
	E PREPARAÇÃO DO BANCO DE DADOS	198

APÊNDICE F – IMPLEMENTAÇÃO PARA REFORMATÇÃO EM SUBCONJUNTOS DE DADOS	205
APÊNDICE G – CÓDIGOS PARA EXPLORAÇÃO DOS SUBCONJUNTOS DE DADOS	206
APÊNDICE H – EXPLORAÇÃO E VERIFICAÇÃO DA QUALIDADE DOS DADOS	208
APÊNDICE I – EXPLORAÇÃO DOS NASCIDOS VIVOS E ÓBITOS NA REGIÃO	215
APÊNDICE J – PADRÃO TEMPORAL DA MORTALIDADE DO NASCIDO VIVO	220
APÊNDICE K – EXPLORAÇÃO DAS CARACTERÍSTICAS MATERNAS, PATERNAS E FÍSICAS DOS NASCIDOS VIVOS E ÓBITOS	230
APÊNDICE L – PROCEDIMENTOS COMPUTACIONAIS PARA ANÁLISE ESPACIAL E MODELAGEM COM TÉCNICAS DE MACHINE LEARNING	245
APÊNDICE M – MAPA MORAN BIVARIADO ENTRE A MORTALIDADE INFANTIL, NEONATAL E PÓS-NEONATAL COM AS VARIÁVEIS SELECIONADAS	247
ANEXO A – INDICADORES SOCIAIS DA REGIÃO METROPOLITANA DO VALE DO PARAÍBA E LITORAL NORTE . .	248
ANEXO B – DECLARAÇÃO DE NASCIDO VIVO	249
ANEXO C – DECLARAÇÃO DE ÓBITO	250
ANEXO D – DESCRIÇÃO DAS VARIÁVEIS DO CONJUNTO DE DADOS DO SISTEMA DE INFORMAÇÕES SOBRE NASCIDOS VIVOS	251
ANEXO E – DESCRIÇÃO DAS VARIÁVEIS DO CONJUNTO DE DADOS DO SISTEMA DE INFORMAÇÕES SOBRE MORTALIDADE	254
ANEXO F – TABELA DA DISTRIBUIÇÃO QUI-QUADRADO (χ^2) .	257

DADOS CURRICULARES	258
-------------------------------------	------------

1 INTRODUÇÃO

Este Capítulo apresenta a contextualização da pesquisa, a formulação das questões de pesquisa, objetivos gerais e específicos, a delimitação, a justificativa e importância da pesquisa e a estrutura do documento.

1.1 CONTEXTUALIZAÇÃO DA PESQUISA

A mortalidade infantil é o termo utilizado para descrever o óbito do nascido vivo que ocorre durante o primeiro ano de vida, e frequentemente classificado nas faixas etárias de: óbito neonatal precoce (0 a 6 dias de vida), óbito neonatal tardio (7 a 27 dias de vida) e óbito pós-neonatal (28 a 364 dias de vida) (Wise, 2017). Além disso, a mortalidade do recém-nascido é considerada um problema de saúde pública que reflete o estado e a qualidade da assistência recebida pela mãe e pelo recém-nascido, bem como o reflexo do desenvolvimento do país.

Num contexto cronológico, a mortalidade infantil mundial tem diminuído substancialmente, e uma das principais causas são os objetivos estabelecidos por diferentes países. No ano de 2000, os estados membros das Nações Unidas e a Organização Mundial de Saúde (OMS) estabeleceram nos Objetivos de Desenvolvimento do Milênio (ODM) o objetivo de reduzir a mortalidade infantil até 2015, visto que em 1990 a taxa mundial de mortalidade infantil era de 90 óbitos por mil nascidos vivos (UN, 2015a). Como resultado, em 2015 a Taxa de Mortalidade Infantil (TMI) foi reduzida em mais da metade dos óbitos, com 43 óbitos por mil nascidos vivos (UN, 2015a; Marinho *et al.*, 2020), porém, não foi suficiente para atingir o objetivo dos ODM.

Já no ano de 2015 foi determinada uma nova agenda com os Objetivos de Desenvolvimento Sustentável (ODS), na qual, especificamente no Objetivo 3 de Saúde e Bem-estar, foi estabelecida a meta de reduzir a mortalidade neonatal para pelo menos 12 por mil nascidos vivos e reduzir a mortalidade em crianças menores de 5 anos para 25 óbitos por mil nascidos vivos até 2030 (UN, 2015b).

Assim, de modo geral e de acordo com a OMS, o número de mortes de recém-nascidos em todo o mundo diminuíram de 5 para 2,4 milhões entre o período de 1990 a 2019, porém ainda há regiões onde as taxas de mortalidade ainda são altas (WHO, 2020). Não obstante, aquela diminuição de mortalidade foi refletida principalmente na diminuição da mortalidade pós-neonatal, na qual o número de mortes neonatais não teve a mesma redução (WHO, 2014; Lawn *et al.*, 2014), o que indica que o recém-nascido apresenta maior risco de óbito nos primeiros 28 dias de vida (Hug *et al.*, 2019).

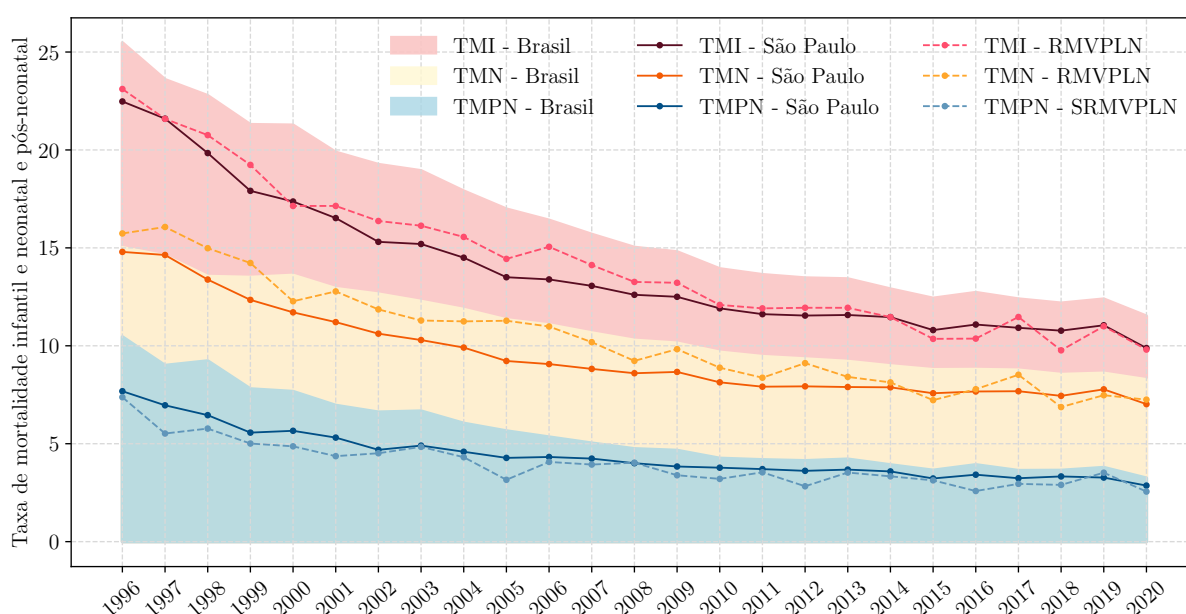
Além disso, considerando apenas os óbitos de 2019, 47% das perdas em menores de 5 anos ocorreram nos primeiros 28 dias, e aproximadamente três quartos desse valor ocorreram na primeira semana de vida (WHO, 2020). Deste modo, é evidente que reduzir o número de óbitos neonatais ainda representa um grande desafio, visto que sua diminuição é essencial para a

minimização da mortalidade infantil, além da redução da mortalidade em menores de 5 anos (Hug *et al.*, 2019).

No contexto brasileiro, as taxas de mortalidade há mais de 8 décadas eram bastante elevadas. Especificamente em 1940 a TMI era de 146,6 óbitos por mil nascidos vivos (IBGE, 2017). Para 1996, a TMI apresentou uma queda significativa, com 25,5 por mil nascidos vivos, e na divisão de Taxa de Mortalidade Neonatal (TMN) e Taxa de Mortalidade Pós-neonatal (TMPN) as taxas foram de 15,0 e 10,5 óbitos por mil nascidos vivos, respectivamente.

Em 2015, a meta dos ODM foi atingida com uma redução da mortalidade em todas as regiões do Brasil, principalmente uma queda significativa na região Nordeste (Marinho *et al.*, 2020). Naquele ano, a TMI foi de 12,4 por mil nascidos vivos (Marinho *et al.*, 2020; MS-SVE, 2019), com uma reduzida diminuição subsequente nos próximos anos, até em 2020 alcançar 11,5 óbitos por mil nascidos vivos, conforme observado na Figura 1.

Figura 1 – Taxa de mortalidade infantil, neonatal e pós-neonatal no Brasil, São Paulo e RMVPLN



TMI: Taxa de mortalidade infantil, TMN: Taxa de mortalidade neonatal, TMPN: Taxa de mortalidade pós-neonatal,
RMVPLN: Região Metropolitana do Vale do Paraíba e Litoral Norte.

Fonte: Elaborado pela autora (2025), com dados do DATASUS (2023).

A redução do número de óbitos infantis no período 1996 – 2020 (Figura 1) foi devida, principalmente, à diminuição da mortalidade pós-neonatal, pois perto de 70% dos óbitos foram perdas neonatais, e 54% ocorreram na primeira semana de vida (Leal *et al.*, 2018; Marinho *et al.*, 2020).

Além disso, de acordo com o relatório do Ministério da Saúde (divulgado em 2019), a redução dos óbitos no período infantil está relacionada a diversos fatores associados, como os avanços na cobertura vacinal, melhora na atenção da mãe e do recém-nascido, melhora da educação materna, aumento da renda da per capita, além do acesso a programas de saúde pública,

como o Programa Bolsa Família e Programa de Saúde da Família (Lourenço *et al.*, 2014; Victora *et al.*, 2011; MS-SVE, 2019).

Por outro lado, embora no Brasil a TMI até 2020 tenha diminuído substancialmente, em países como França, Alemanha, Israel, Itália, Japão, Holanda, Portugal e Reino Unido suas taxas de mortalidade foram inferiores a 3,7 óbitos por mil nascidos vivos, e em países do continente americano como Cuba, Canadá, Estados Unidos, Chile, Costa Rica e Argentina sua TMI foi menor a 6,8 óbitos por mil nascidos vivos (UNICEF, 2023), (no Apêndice A ver mapa da distribuição da TMI no mundo).

No caso do Estado de São Paulo, a mortalidade infantil também apresentou um decréscimo ao longo dos anos, sendo que no período 1996 – 2015 a queda foi de 22,5 óbitos por mil nascidos vivos para 10,8 óbitos por mil nascidos vivos. A partir de 2015 foi observada uma redução desacelerada, alcançado em 2020 uma TMI de 9,9 óbitos por mil nascidos vivos (DATASUS, 2023). Mesmo assim, essa taxa de mortalidade ainda é maior quando comparada a algumas taxas de países desenvolvidos e de países vizinhos (Ortiz; Oushiro, 2008; Areco; Konstantyner; Taddei, 2016; UNICEF, 2023).

Do mesmo modo, o Estado de São Paulo seguiu o mesmo padrão observado no nível nacional, visto que a principal redução dos óbitos infantis deveu-se à diminuição da TMPN (Ortiz; Oushiro, 2008), como foi mostrado na Figura 1. Do qual, 71% dos óbitos infantis do 2020 ocorreram no período neonatal (DATASUS, 2023).

Deste jeito, a distribuição das altas taxas de mortalidade neonatal no Estado de São Paulo localizaram-se principalmente na região Sudoeste do Estado e na região do Vale do Paraíba (Nascimento; Almeida; Gomes, 2014; Almeida; Gomes; Nascimento, 2014). Nesse contexto, destacam-se, com taxas elevadas os municípios de Itapeva, Capão Bonito, Santos, Itanhaém da região Sudoeste (Almeida; Gomes; Nascimento, 2014), e os municípios de Paraibuna, Caraguatatuba, Guaratinguetá e Campos do Jordão, pertencente à Região Metropolitana do Vale do Paraíba e Litoral Norte (RMVPLN) (Nascimento *et al.*, 2007; Almeida; Gomes; Nascimento, 2014).

Por outro lado, a mortalidade infantil na RMVPLN tem apresentado maior variação ao longo do tempo (Figura 1), mas com queda semelhante à do Estado de São Paulo. Em 1996 a TMI da região foi 23,1 óbitos por mil nascidos vivos, diminuindo em 2015 para 10,4 óbitos por mil nascidos vivos, e em 2020 com 9,8 óbitos por mil nascidos vivos (DATASUS, 2023). De forma semelhante, na região, 68% das mortes ocorreram no primeiro mês de vida.

Deste modo, fica evidente que reduzir o número de óbitos na RMVPLN é um grande desafio, e a redução do número de óbitos neonatais é a chave, visto que o número de perdas é maior que os ocorridos no período pós-neonatal (Lawn *et al.*, 2014).

No entanto, sob outra perspectiva, alguns estudos na literatura sugerem que os óbitos infantis podem ser evitados, mesmo que sejam recém-nascidos prematuros (<37 semanas) ou com baixo peso ao nascer (<2.500 g), pois, estão associados a outros fatores que podem ser controlados ou minimizados (Silva *et al.*, 2010; Saugstad, 2011; Lau *et al.*, 2013; Paixao *et al.*, 2021; Bernardino *et al.*, 2022; Rodríguez *et al.*, 2023). Assim, além da prematuridade e o baixo peso,

o óbito infantil está associado a outros fatores biológicos como anomalias congênitas, doenças infecciosas, asfixia (Hug *et al.*, 2019), além de estar relacionado com características maternas, fatores socioeconômicos e características relacionadas ao acesso e qualidade do serviço de saúde recebidos pela mãe e pelo recém-nascido (Schoeps *et al.*, 2007; Ortiz; Oushiro, 2008; Areco; Konstantyner; Taddei, 2016; Nascimento; Machin; Santos, 2017; Marinho *et al.*, 2020).

Neste cenário, graças aos avanços da ciência, a digitalização das informações da área de saúde tem gerado uma grande quantidade de dados estruturados e não estruturados, também conhecidos como *Big Data* (BD), que podem ser utilizadas para extrair informações valiosas para melhorar a qualidade da assistência prestada (Raghupathi; Raghupathi, 2014; Chiavegatto, 2015; Javaid *et al.*, 2022).

Nos últimos anos tem-se observado um interesse crescente na utilização e integração de diferentes tecnologias da área de informática, entre elas os Sistemas de Informação Geográfica (SIG) e a Inteligência Artificial (IA), num esforço de desenvolver ferramentas para auxiliar os especialistas na tomada de decisões em problemas de saúde pública. Entre as principais aplicações dessas ferramentas estão a identificação de áreas de risco, previsão de eventos futuros, identificação de nascidos vivos com risco de mortalidade ou morbidade, ou diagnóstico de doenças, etc.

Assim, essas ferramentas podem ser utilizadas em qualquer um dos níveis de planejamento (estratégico, tático e operacional) da investigação epidemiológica que, além de auxiliar na tomada de decisões, podem gerar um diferencial na assistência prestada ao recém-nascido, além de melhorar o serviço de saúde prestado permitindo otimizar a alocação de recurso (Sin, 2020).

Especificamente, os SIG são ferramentas de geoprocessamento de dados que permitem a aplicação de métodos estatísticos para análise e visualização de informações espaço-temporais, com o objetivo de medir a existência de relações espaciais (Nykiforuk; Flaman, 2011). Por conseguinte, a análise espacial pode permitir a identificação de aglomerados espaciais e áreas de risco, onde uma possível intervenção prioritária poderia ser importante para minimizar consequências.

A literatura evidencia que os SIG têm sido aplicados a diferentes problemas de saúde, e a análise espacial da mortalidade de nascidos vivos não tem sido exceção, tanto em áreas geográficas em nível internacional (Kazembe; Kandala, 2015; Insaf; Talbot, 2016), como em nível do Brasil (Nascimento *et al.*, 2007; Gonçalves; Costa; Braga, 2011; Almeida; Gomes; Nascimento, 2014; Nascimento; Almeida; Gomes, 2014).

Por outro lado, a IA pode ser definida como a capacidade das máquinas de simular a inteligência dos seres humanos, ou seja, de perceber, aprender e compreender seu ambiente (Rong *et al.*, 2020). Além disso, a IA é multidisciplinar e compõe-se da subárea de Aprendizado de Máquina, em inglês *Machine Learning* (ML), que está relacionada com a Ciência de Dados.

A Ciência de Dados pode ser definida como uma sequência de processos que combina a aplicação de métodos estatísticos e programação para a extração de conhecimento de um conjunto de dados (Raghupathi; Raghupathi, 2014; Javaid *et al.*, 2022). Dessa forma, pode-se

considerar que a análise espacial faz parte da Ciência de Dados, também conhecida como *spatial data science*, por meio da utilização de SIG e dados espaciais para a extração de informações relevantes. Já o ML pode ser definido como o estudo de métodos computacionais para identificar padrões complexos em milhões de dados a fim de construir modelos preditivos (Jordan; Mitchell, 2015).

Assim, o ML e Ciência de Dados podem ser utilizados para processar uma grande quantidade de dados e desenvolver uma ampla variedade de aplicações no setor de saúde (Javaid *et al.*, 2022), pelo qual vêm ganhando uma grande popularidade na área (Kumar *et al.*, 2020).

Além do mais, nos últimos anos observou-se um interesse crescente na utilização das técnicas de ML na modelagem preditiva de eventos futuros por sua capacidade de capturar as relações complexas nos dados (Doupe; Faghmous; Basu, 2019; Kumar *et al.*, 2020). Assim, as técnicas de ML podem ser aplicadas a problemas relacionados com predição (*prediction*) e previsão (*forecasting*), levando em consideração a diferença entre os dois termos.

Portanto, é importante destacar que, ao longo deste trabalho, o uso dos termos predição e previsão foram usados de forma diferente, considerando que predição refere-se à estimativa de eventos com base em dados não vistos (desconhecidos), e previsão refere-se à previsão do futuro, considerando o componente temporal e suas variações (Bares; Bosse, 2022).

Nesta continuidade, em conformidade com a literatura publicada, ainda são poucos os trabalhos relacionados com a aplicação de ML para predição ou previsão em problema de mortalidade infantil. Grande parte desses trabalhos da literatura têm como abordagem principal o desenvolvimento de modelos para predizer (ou identificar) o risco de mortalidade em recém-nascidos (Lee *et al.*, 2021; Batista *et al.*, 2021; Brahma; Mukherjee, 2022; Teji *et al.*, 2022), seguido pela aplicação de ML para previsão de eventos futuros com séries de tempo (Rodriguez *et al.*, 2022; Ventura *et al.*, 2023).

Entretanto, é importante evidenciar que ainda são poucos os estudos em que o ML foi aplicado na previsão de eventos de saúde (Soyiri; Reidpath, 2013), e a quantidade é ainda menor quando se trata de mortalidade.

Já no cenário brasileiro, a aplicação de ML e Ciência de Dados ainda está em desenvolvimento, sabendo-se estas tecnologias podem contribuir significativamente no suporte da tomada de decisão (Santos, 2018; Doupe; Faghmous; Basu, 2019; Javaid *et al.*, 2022), no planejamento de políticas públicas de saúde e na melhoria da prestação de serviços de saúde e cuidados preventivos (Soyiri; Reidpath, 2013; Silva *et al.*, 2021).

Neste contexto, o presente trabalho visa a exploração das técnicas de análise espacial e ML para a visualização da distribuição e identificação de áreas de risco da mortalidade do nascido vivo na RMVPLN, além do desenvolvimento de modelos para prever a mortalidade e identificar nascidos vivos com risco de óbito. A área geográfica de estudo da RMVPLN foi selecionada por ser a quarta região do estado de São Paulo com maior número de habitantes (Mello; Teixeira; Mello, 2016).

1.2 QUESTÕES DE PESQUISA E OBJETIVOS

Com base no descrito em 1.1, as questões de pesquisa e objetivos deste trabalho são definidos a seguir.

1.2.1 Questões de Pesquisa

As questões de pesquisa que encaminharam o desenvolvimento desta investigação foram definidos conforme descrito a seguir.

- É possível utilizar SIG e técnicas de ML em dados de mortalidade de nascidos vivos para produzir análises e identificar padrões significativos?
- Como os SIG e a análise espacial podem ser empregados para compreender a distribuição geográfica da mortalidade e identificar áreas de maior risco?
- A análise dos dados disponíveis permite identificar padrões de associação entre a mortalidade do recém-nascido e determinantes associados?
- De que forma as técnicas de ML podem ser aplicadas para a previsão de mortalidade de nascidos vivos na região estudada e qual é a técnica mais adequada para o desenvolvimento do modelo?
- De que forma as técnicas de ML podem ser aplicadas para a predição de risco de mortalidade, qual é a técnica mais adequada e quais variáveis devem ser priorizadas como determinantes-chave?
- Qué técnicas usar para validar a eficácia dos modelos desenvolvidos e garantir sua robustez e aplicabilidade prática?

1.2.2 Objetivo Geral e Específicos

O objetivo principal deste trabalho é explorar e integrar técnicas de análise espacial utilizando SIG e aplicar técnicas de ML em dados da mortalidade de nascidos vivos na RMVPLN, com o intuito de identificar áreas de risco, desenvolver modelos de previsão e identificação da mortalidade do nascido vivo, a fim de auxiliar especialistas em saúde na tomada de decisões para a prevenção e redução de óbitos.

Para atingir o objetivo principal desta pesquisa, os objetivos específicos são apresentados a seguir:

- Analisar os dados disponíveis da RMVPLN, com ênfase nas informações relacionadas aos nascidos vivos no período de 2010 a 2020.
- Integrar, transformar e preparar os dados, garantindo sua consistência e adequação ao uso em SIG e ML.

- Identificar a distribuição espacial e as áreas de maior risco de mortalidade de nascidos vivos, considerando as faixas etárias de óbito (neonatal, pós-neonatal e infantil), e correlacionar a mortalidade com os principais determinantes associados, destacando os fatores mais relevantes na área de estudo.
- Desenvolver e avaliar modelos de ML para previsão do número mensal de óbitos por faixa etária, priorizando a seleção de modelos com melhor desempenho.
- Analisar os determinantes utilizados no treinamento dos modelos de classificação, destacando sua relevância na predição do risco de mortalidade.
- Implementar e testar modelos de ML baseados em técnicas de classificação para identificar nascidos vivos com maior probabilidade de risco de óbito, selecionando as técnicas mais eficazes.
- Validar os resultados obtidos pelos modelos preditivos e garantir sua funcionalidade para aplicação prática em cenários de apoio à decisão.

1.3 JUSTIFICATIVA E IMPORTÂNCIA

Os recém-nascidos correm o risco de várias complicações de saúde, e aqueles que são prematuros ou com baixo peso ao nascer são ainda mais vulneráveis, apresentando um alto risco de óbito neonatal, pós-neonatal ou infantil (Reyes; Manalich, 2005). Além disso, em relação à evitabilidade de óbitos, alguns mortes podem ser eventos evitáveis devido a sua associação com a qualidade da assistência recebida pela gestante e pelo recém-nascido no pré-natal, durante e após o parto (Bernardino *et al.*, 2022), além de sua associação com outros determinantes (Nascimento; Machin; Santos, 2017; Hug *et al.*, 2019; Marinho *et al.*, 2020).

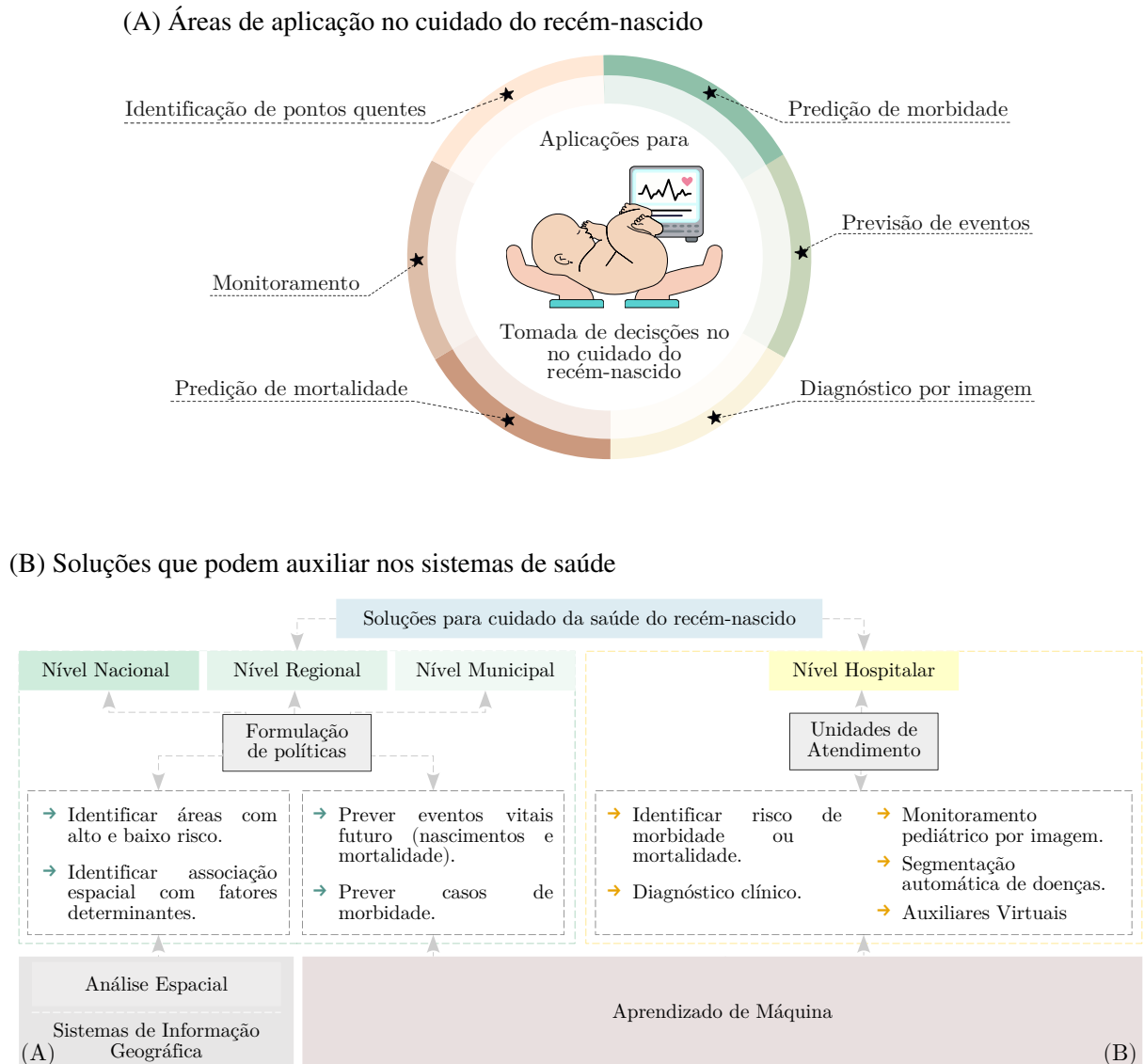
Neste cenário, observou-se na literatura um interesse crescente no uso de diferentes tecnologias, como os SIG e ML, por sua capacidade de fornecer informações úteis que podem ser utilizadas no planejamento de políticas públicas e na tomada de decisões para o cuidado preventivo e melhoria dos serviços de saúde (Soyiri; Reidpath, 2013; Silva *et al.*, 2021). (no Apêndice B apresenta-se uma análise da tendência no uso da análise espacial e ML no cuidado da saúde)

Deste jeito, conforme observado em estudos, aquelas tecnologias podem ser aplicadas a diferentes problemas no cuidado do recém-nascido (Matsushita; Krebs; Carvalho, 2022; Malhotra *et al.*, 2023), aproveitando a geração e o armazenamento massivo de dados de saúde, para melhorar a qualidade dos serviços (Chiavegatto, 2015).

A Figura 2 (A) apresenta as principais áreas de aplicação dos SIG e ML no cuidado do recém-nascido. Portanto, entre os principais campos de aplicação dos SIG e ML no cuidado do recém-nascido, destaca-se o desenvolvimento de modelos para predição de morbidade, mortalidade (Matsushita; Krebs; Carvalho, 2022), previsão de eventos de saúde com séries

de tempo (Javaid *et al.*, 2022), diagnóstico com processamento de imagem, monitoramento pediátrico (Malhotra *et al.*, 2023) e identificação de pontos quentes para o gerenciamento de riscos (Javaid *et al.*, 2022).

Figura 2 – Principais áreas de aplicação e soluções dos SIG e ML no cuidado do recém-nascido



Fonte: Elaborado pela autora (2025).

Sob essa perspectiva, e com foco no atendimento ao recém-nascido, especificamente no cuidado da saúde do nascido vivo, a Figura 2 (B) mostra uma sequência de soluções, indicando alguns aspectos sobre como os SIG e o ML podem auxiliar os sistemas de saúde na mitigação de riscos tanto a nível nacional, regional ou municipal, como a nível hospitalar.

Assim, com base na Figura 2 (B), os SIG e estudos de distribuição espacial podem ajudar na identificação de aglomerados municipais com alto de risco em eventos vitais e morbilidade (Nascimento *et al.*, 2007; Gonçalves; Costa; Braga, 2011; Nascimento; Almeida; Gomes, 2014; Kazembe; Kandala, 2015), permitindo a obtenção de informações importantes que podem ajudar

na formulação de estratégias de intervenção ou ações nas políticas públicas para melhora do cuidado da mãe e do recém-nascido (Nascimento *et al.*, 2007).

Por outro lado, o ML pode ser aplicado no desenvolvimento de ferramentas tanto a nível nacional, regional ou municipal, como também a nível hospitalar (Figura 2 (B)). A aplicação destas técnicas na área de saúde ainda é considerada um campo emergente (Doupe; Faghmous; Basu, 2019), porém, com um interesse crescente em diferentes problemas relacionados ao cuidado do recém-nascido.

Deste modo, o ML é muito popular por sua capacidade de processar informações, possibilitando o desenvolvimento de modelos para podem fornecer suporte aos especialistas (Javaid *et al.*, 2022), que juntamente com os SIG podem servir de apoio na tomada de decisões nos diferentes níveis de planejamento das diferentes áreas do setor (estratégico, tático e operacional), auxiliando na priorização de ações, além de auxiliar no monitoramento de problemas (Sin, 2020).

Assim, relacionando as informações apresentadas na Figura 2 (B) com o problema da mortalidade do nascido vivo deste trabalho, os SIG e ML podem ser usados como ferramentas de apoio tanto para a tomada de decisões em nível regional com a identificação de padrões espaciais e previsão de eventos futuros como a mortalidade (Javaid *et al.*, 2022), quanto em nível hospitalar para a identificação de nascidos vivos com risco de óbito (Matsushita; Krebs; Carvalho, 2022).

Nesta sequência, para destacar a contribuição desta pesquisa, foi realizado um estudo bibliométrico descritivo-retrospectivo das pesquisas existentes na literatura com o intuito de mapear e analisar os trabalhos relacionados com este estudo. Especificamente, o foco principal da análise foi sobre a utilização de métodos de análise espacial e aplicação de técnicas de ML para previsão eventos e identificação de nascidos vivos com risco em problemas relacionados com a mortalidade (e suas faixas etárias).

Os artigos analisados foram obtidos das bases de dados da PubMed e Scopus, tendo em vista que fornecem acesso a um grande número de periódicos, e especialmente PubMed disponibiliza uma vasta gama de documentos na área da medicina (Falagas *et al.*, 200).

A busca na literatura foi dividida em duas fases, sendo que a primeira fase envolveu os trabalhos relacionados à análise espacial da mortalidade do recém-nascido, sendo utilizados os termos de: "*spatial analysis*", "*healthcare*", "*infant*", "*neonatal*", "*postneonatal*", "*mortality*", "*death*", "*Brazil*", "*Paraíba valley*" e "*Paraíba Valley and North Coast*".

Já a segunda fase foi da análise dos artigos relacionados com ML na mortalidade do recém-nascido, mapeando os trabalhos focados na identificação de recém-nascidos com risco de óbito e na previsão de mortalidade de nascidos vivos. Os termos utilizados na busca foram: "*machine learning*", "*time series*", "*forecast*", "*healthcare*", "*infant*", "*neonatal*", "*postneonatal*", "*mortality*", "*death*", "*Brazil*", "*Paraíba valley*" e "*Paraíba Valley and North Coast*".

Em cada fase, foi apresentada a estrutura da busca com as palavras-chaves consideradas, além de mostrar a quantidade de documentos encontrados em cada base de dados. A busca de artigos nas bases de dados foi refinada incluindo apenas artigos originais e artigos de anais de eventos, publicados até 3 de novembro de 2024. Na Scopus a busca de documentos foi

baseada em 'TITLE' e 'Article title, abstract, keywords', e no PubMed foi baseada em 'Title' e 'Title/Abstract'.

Por outro lado, como o Scopus e o PubMed foram utilizados na busca de artigos científicos, os resultados mostraram quantidades diferentes em cada base de dados, indicando a necessidade de uma análise e filtragem. A análise dos documentos foi realizada por metadados e por leitura de artigos, na qual foram encontrados artigos semelhantes, o que justificou a realização de uma depuração. Após a filtragem de documentos, obteve-se o conjunto final de artigos analisados para cada abordagem de estudo.

Adicionalmente, considerando um nível de significância estatística de $p\text{-valor} < 0,05$, foi utilizado o coeficiente de correlação de Spearman (r_s) para avaliar a associação entre o número de publicações e o número total de citações, que indica a existência de associação negativa (-1) ou positiva ($+1$) (Demir *et al.*, 2020).

Fase I – Análise espacial:

A Tabela 1 apresenta a estrutura e resultados da primeira fase de busca, mostrando a quantidade de artigos encontrados em cada base de dados e a quantidade final de documentos.

Tabela 1 – Número de artigos encontrados na Scopus e PubMed na primeira fase de busca

Palavras-chave	Scopus	PubMed	Artigos
TITLE-ABS-KEY("spatial analysis"AND ("health care"OR "healthcare"))**	2.476*	521*	2.577*
TITLE-ABS-KEY("spatial analysis") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	74	35	76
TITLE-ABS-KEY("spatial analysis"AND "Brazil") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	25	12	26
TITLE-ABS-KEY("spatial analysis"AND "Paraíba valley") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	1	1	1
TITLE-ABS-KEY("spatial analysis"AND "Paraíba Valley and North Coast") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	0	0	0

(**) No Apêndice B encontra-se a análise da tendência dessas publicações.

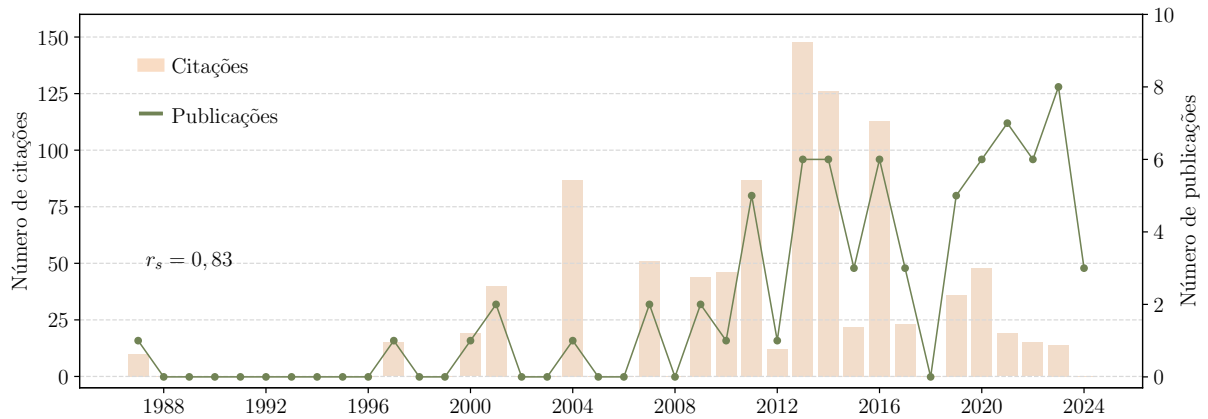
Fonte: Elaborado pela autora (2025).

Os resultados mostram um número considerável de publicações (Tabela 1), com 2.577 documentos, o que indica a presença de um interesse crescente da comunidade científica na aplicação da análise espacial em estudos relacionados à saúde (ver Apêndice B).

Em contrapartida, esse mesmo comportamento, mas em menor intensidade, foi observado nas publicações relacionadas à mortalidade de nascidos vivos, conforme ilustrado na Figura 3. Além disso, a busca dos artigos publicados com foco na análise espacial aplicada na exploração de padrões espaciais de mortalidade do recém nascido resultou em 74 documentos encontrados na Scopus e 35 documentos em PubMed.

Na filtragem dos metadados coletados foram encontrados 33 artigos duplicados, resultando em 76 artigos únicos com 975 citações acumuladas no período 1987 – 2024 (Figura 3), com uma taxa de crescimento anual de publicações de 4,01%.

Figura 3 – Tendência de publicações e citações do análise espacial na mortalidade do nascido vivo, período 1987 – 2024

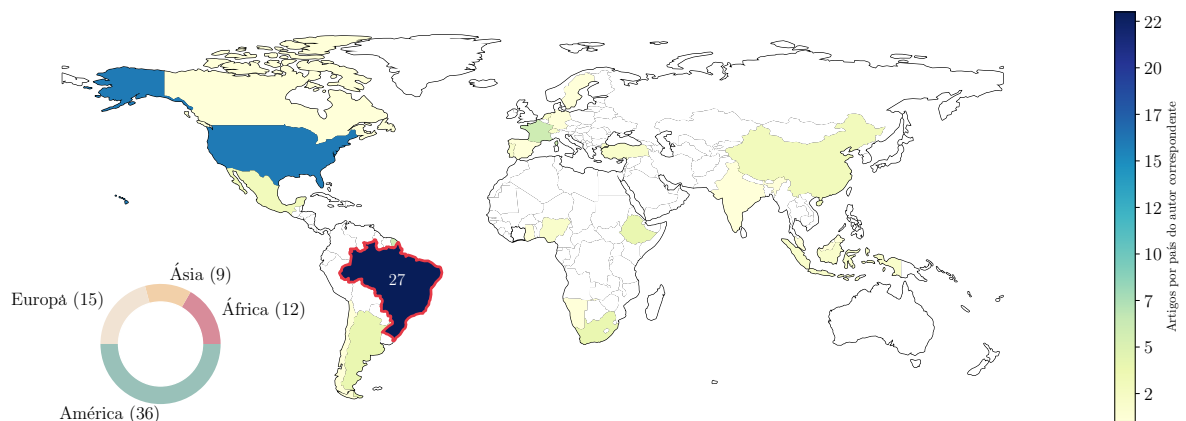


Fonte: Elaborado pela autora (2025).

Os resultados indicaram a existência de uma correlação positiva significativa entre o número total de publicações e o número total de citações ($r_s = 0,83$, $p\text{-valor} < 0,001$, Figura 3), indicando que o aumento do número de publicações nesta área de estudo está relacionado ao aumento do número de citações. Em particular, observou-se que 83,33% (65) dos artigos foram publicados nos últimos quatorze anos (2011 – 2024), com 663 (68,00%) citações acumuladas.

Na análise da contribuição por país, avaliado pelo autor correspondente, foi identificado que os artigos examinados foram publicados por 23 países. A Figura 4 apresenta aqueles países interessados na aplicação de análise espacial na mortalidade do nascido vivo, mostrando o número de documentos publicados.

Figura 4 – Produção dos países avaliada pelo autor correspondente do análise espacial na mortalidade do nascido vivo, período 1987 – 2024

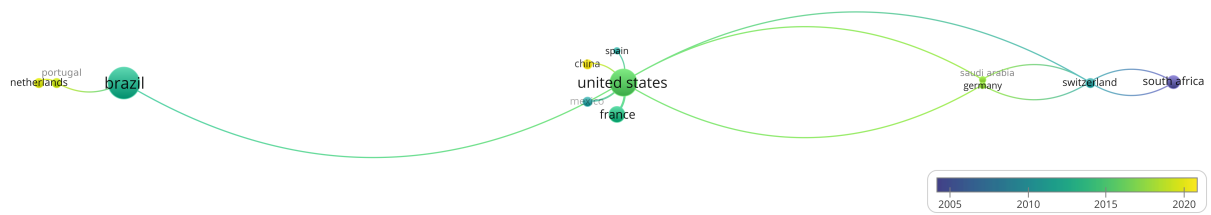


Fonte: Elaborado pela autora (2025).

Na avaliação de documentos por continente, identificou-se maior interesse nesta área por países pertencentes ao continente americano (36), seguidos pelos países dos continentes europeu (15), africano (12) e asiático (9). No entanto, os três países com maior contribuição científica foram Brasil (27), Estados Unidos (16) e França (6).

Por outro lado, a Figura 5 ilustra a rede de colaboração internacional entre os países interessados nesta área de estudo, mostrando que a colaboração entre países vem aumentando, principalmente a partir de 2005. A rede de colaboração internacional foi gerada no VOSviewer¹ (Eck; Waltman, 2010).

Figura 5 – Rede de colaboração internacional dos artigos sobre análise espacial na mortalidade do nascido vivo, período 1987 – 2024



Fonte: Elaborado pela autora (2025).

A rede da Figura 5 apresenta a participação de 12 países que publicaram mais de um documento, refletindo a participação e seu grau de colaboração entre os países. Assim, o tamanho do círculo mostra o número de documentos publicados, a espessura da linha reflete a força da colaboração entre países, e as cores são o ano médio de colaboração (2005–2020) (Eck; Waltman, 2010). Além disso, sendo o Brasil o país com maior número de artigos publicados, sua rede de colaboração internacional incluiu principalmente uma colaboração com os Estados Unidos, seguido por Portugal e Países Baixos.

Por outro lado, a análise destes artigos possibilitou a identificação dos estudos realizados no Brasil, facilitando sua revisão e classificação sistemática de acordo com a região ou local de estudo e a faixa etária do óbito, conforme apresentado no Quadro 1.

Quadro 1 – Classificação das publicações brasileiras por faixa etária de óbito

(continua)

Local de estudo	Área	Faixa etária do óbito					Artigos
		N ^a	NP ^b	NT ^c	PN ^d	I ^e	
Brasil	País	✓					Schramm & Szwarcwald (2000)
Brasil	País					✓	Buhler <i>et al.</i> (2014), Melo <i>et al.</i> (2023), Justino <i>et al.</i> (2021) e Dilélio <i>et al.</i> (2024)
Brasil	País		✓	✓			Rodrigues <i>et al.</i> (2016)

Faixa etária (a) N: neonatal, (b) NP: neonatal precoce, (c) NT: neonatal tardio, (d) PN: pós-neonatal e (e) I: infantil.
Fonte: Elaborado pela autora (2025).

¹ VOSviewer é um software para a construção de mapas de redes bibliométricas.

Quadro 1 – Classificação das publicações brasileiras por faixa etária de óbito

(conclusão)

Local de estudo	Área	Faixa etária do óbito					Artigos
		N ^a	NP ^b	NT ^c	PN ^d	I ^e	
Rio de Janeiro	Município	✓					Andrade & Szwarcwald (2001)
Rio de Janeiro	Estado	✓	✓	✓			Leal & Szwarcwald (1997)
Rio de Janeiro	Município	✓			✓	✓	Dutra <i>et al.</i> (2022)
São Paulo	Estado	✓					Nascimento, Almeida & Gomes (2014) e Costa-Nobre <i>et al.</i> (2021)
São Paulo	Estado	✓	✓	✓			Almeida, Gomes & Nascimento (2014)
São Paulo	Estado	✓					Marinonio <i>et al.</i> (2022)
São Paulo	Estado	✓			✓	✓	Rodríguez <i>et al.</i> (2023)
Pernambuco	Estado					✓	Rodrigues <i>et al.</i> (2014) e Rodrigues <i>et al.</i> (2013)
Pernambuco	Estado	✓			✓	✓	Silva <i>et al.</i> (2023)
Minas Gerais	Estado					✓	Faria & Santana (2016) e Silva & Silva ()
Espírito Santo	Estado					✓	Dias <i>et al.</i> (2019)
Ceará	Estado	✓			✓		Filho <i>et al.</i> (2007)
Goiânia	Município	✓			✓		Neto <i>et al.</i> (2001)
Salvador	Município	✓					Gonçalves, Costa & Braga (2011)
Recife	Município					✓	Bonfim <i>et al.</i> (2020)
Recife	Município	✓			✓	✓	Barros <i>et al.</i> (2019)
Vale do Paraíba	Região	✓	✓	✓			Nascimento <i>et al.</i> (2007)
Paraná	Região					✓	Broday & Kluthcovsky (2021)

Faixa etária (a) N: neonatal, (b) NP: neonatal precoce, (c) NT: neonatal tardio, (d) PN: pós-neonatal e (e) I: infantil.

Fonte: Elaborado pela autora (2025).

Os resultados da revisão de artigos mostraram que a mortalidade do recém-nascido foi estudada espacialmente em diferentes áreas geográficas de Brasil (país, região, estado e município), além de ter sido explorada nas diferentes faixas da etárias, como mortalidade neonatal, neonatal precoce, neonatal tardio, pós-neonatal e infantil (Quadro 1).

Além disso, foi possível identificar que, além de ser analisada a distribuição espacial da mortalidade do nascido vivo e identificação de áreas de risco (Schramm; Szwarcwald, 2000; Andrade; Szwarcwald, 2001; Neto *et al.*, 2001; Nascimento *et al.*, 2007; Rodrigues *et al.*, 2013; Rodrigues *et al.*, 2014; Almeida; Gomes; Nascimento, 2014; Faria; Santana, 2016; Rodrigues *et al.*, 2016; Dutra *et al.*, 2022), também foram analisadas outras variáveis de estudo.

Neste sentido, a mortalidade do nascido vivo foi analisada espacialmente por causas evitáveis (Nascimento; Almeida; Gomes, 2014; Dias *et al.*, 2019; Bonfim *et al.*, 2020), sua associação com o baixo peso ao nascer (Rodríguez *et al.*, 2023), prematuridade, síndrome respiratória (Marinonio *et al.*, 2022), toxoplasmose congênita (Melo *et al.*, 2023), diarreia infantil (Buhler *et al.*,

2014). Outrossim, a mortalidade também foi avaliada com outras variáveis como características demográficas, socioeconômicos, atenção à saúde, características maternas e do parto (Leal; Szwarcwald, 1997; Filho *et al.*, 2007; Gonçalves; Costa; Braga, 2011; Barros *et al.*, 2019; Silva; Silva, ; Broday; Kluthcovsky, 2021; Costa-Nobre *et al.*, 2021; Justino *et al.*, 2021; Silva *et al.*, 2023; Dilélio *et al.*, 2024).

Por outro lado, é importante destacar que, na análise de documentos publicados, foi identificado um artigo de análise espacial na região do Vale do Paraíba (Quadro 1). Especificamente, o artigo investigou a mortalidade neonatal em suas divisões precoce e tardia na mesorregião do Vale do Paraíba Paulista, identificando aglomerações espaciais significativas com altos valores de mortalidade neonatal, principalmente na região Central (Nascimento *et al.*, 2007).

Fase II – ML para predição e previsão:

A Tabela 2 apresenta a estrutura de busca e a quantidade de documentos encontrados na segunda fase de análise. Como apresentado na Tabela 2, na literatura há um grande número de publicações sobre ML aplicado em temas de saúde, e o mesmo é observado ao se filtrar a busca por abordagens de previsão de séries temporais com ML (ver Apêndice B).

Tabela 2 – Número de artigos encontrados na Scopus e PubMed da segunda fase de busca

		(continua)		
	Palavras-chave	Scopus	PubMed	Artigos
ML abordagem geral	TITLE-ABS-KEY("machine learning"AND ("health care"OR "healthcare"))*	24.107*	10.010*	25.830*
	TITLE-ABS-KEY("machine learning") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	47	24	49
	TITLE-ABS-KEY("machine learning"AND "Brazil") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	5	2	6
	TITLE-ABS-KEY("machine learning"AND "Paraíba valley") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	0	0	0
	TITLE-ABS-KEY("machine learning"AND "Paraíba Valley and North Coast") AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	0	0	0
	TITLE-ABS-KEY("machine learning"AND ("time series" OR "forecast") AND ("health care"OR "healthcare"))*	1.231*	294*	1.339*
ML para previsão	TITLE-ABS-KEY(("time series" OR "forecast")) AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	153	94	161
	TITLE-ABS-KEY("machine learning"AND ("time series" OR "forecast")) AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))	3	1	4

(*) No Apêndice B encontra-se a análise da tendência dessas publicações.

Fonte: Elaborado pela autora (2025).

Tabela 2 – Número de artigos encontrados na Scopus e PubMed na segunda fase de busca (conclusão)

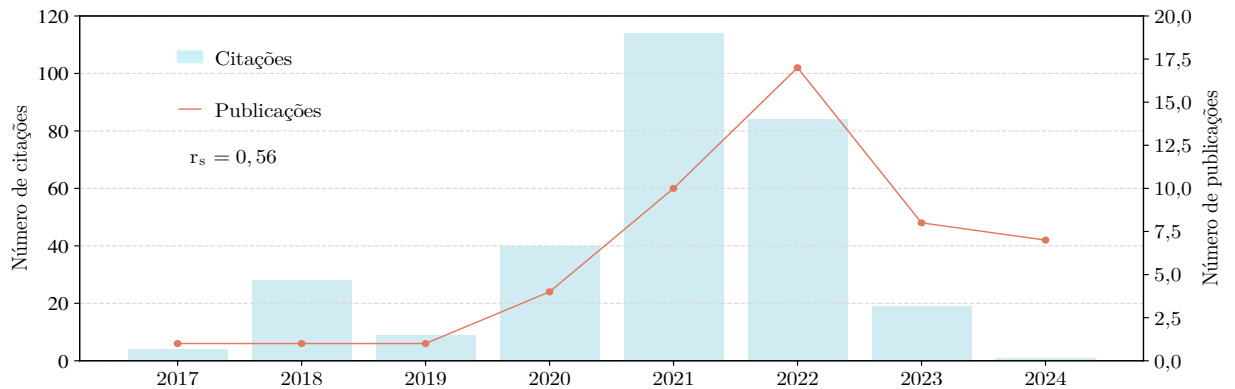
	Palavras-chave	Scopus	PubMed	Artigos
ML para previsão	<i>TITLE-ABS-KEY("machine learning"AND "Brazil AND ("time series" OR "forecast")) AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))</i>	1	0	1
	<i>TITLE-ABS-KEY("machine learning"AND "Paraíba valley AND ("time series" OR "forecast")) AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))</i>	0	0	0
	<i>TITLE-ABS-KEY("machine learning"AND "Paraíba Valley and North Coast AND ("time series" OR "forecast")) AND TITLE(("infant"OR "neonatal"OR "postneonatal") AND ("mortality"OR "death"))</i>	1	0	1

Fonte: Elaborado pela autora (2025).

Em primeiro lugar, a busca foi refinada tentando encontrar os documentos onde o ML foi aplicado à mortalidade do recém-nascido de forma geral, e, segundo, a busca foi feita para aqueles artigos cujo foco principal foi a aplicação do ML para previsão de séries temporais.

Assim, com respeito à análise geral dos documentos relacionados com a aplicação de ML na mortalidade do recém-nascido, foram observadas poucas publicações, porém, com uma tendência crescente, conforme apresentado na Figura 6.

Figura 6 – Tendência de publicações e citações do aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024



Fonte: Elaborado pela autora (2025).

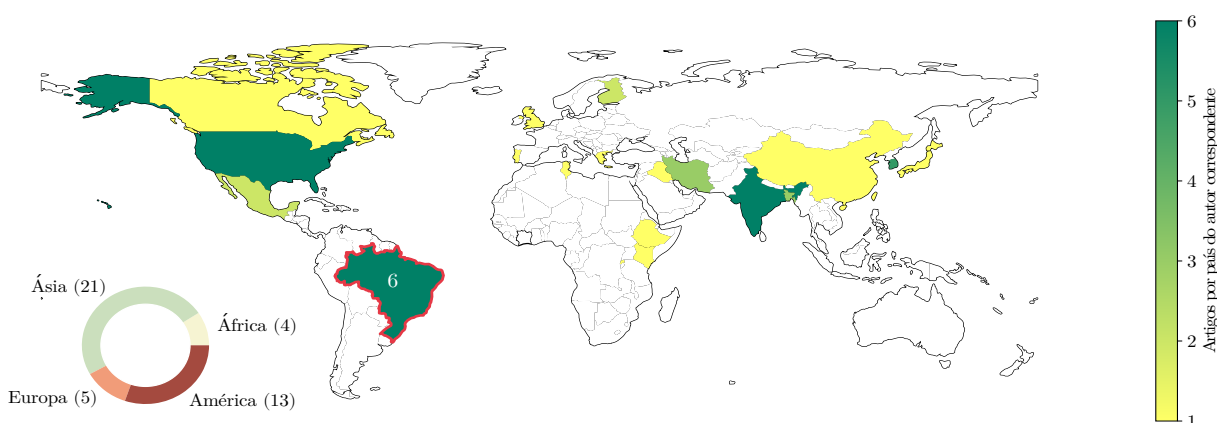
Na análise individual de artigos, foram identificados 47 documentos no banco de dados da Scopus e 24 no banco de dados da PubMed. Após a filtragem de artigos duplicados, foram identificados 49 documentos únicos com 299 citações acumuladas no período 2017 – 2024, e uma taxa de crescimento anual de 6,85%.

No período analisado (Figura 6), o número de publicações e citações apresentou um coeficiente de correlação r_s igual a 0,56, no entanto, o valor obtido não foi significativo, com p -valor = 0,15. Outrossim, foi observado que 93,88% (46) dos artigos foram publicados nos últimos quatro anos (2020 – 2024), com 258 (86,29%) citações acumuladas.

Esses resultados mostraram que, embora o uso de ML na área da saúde já seja realizado há alguns anos com um número substancial de artigos publicados (ver Apêndice B), sua aplicação em problemas relacionados à mortalidade do recém-nascido ainda é incipiente, mas está em crescimento.

A análise da contribuição por país, avaliado pelo autor correspondente, mostrou que os artigos analisados foram publicados por 20 países (Figura 7), sendo que os países pertencentes aos continentes de Ásia (21) e América (13) demonstraram maior interesse devido ao maior número de publicações neste campo de estudo. Os países com maior concentração de pesquisas foram Brasil (6), Estados Unidos (6) e Índia (6).

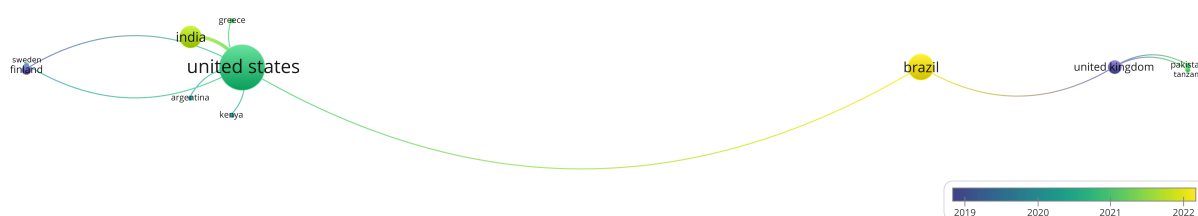
Figura 7 – Produção dos países avaliada pelo autor correspondente do aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024



Fonte: Elaborado pela autora (2025).

Na avaliação da rede de colaboração internacional foi identificada uma crescente colaboração internacional desde 2019 em diante, conforme ilustrado na Figura 8. A colaboração analisada foi entre 11 países com mais de um documento publicado, sendo os Estados Unidos o país com maior rede de colaboração, entre os países de Brasil, Suécia, Finlândia, Índia, Grécia, Argentina e Quênia. Além disso, o Brasil também apresentou colaboração com o Reino Unido.

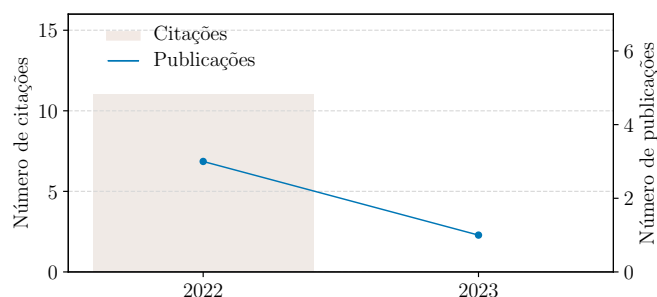
Figura 8 – Rede de colaboração internacional dos artigos sobre aprendizado de máquina na mortalidade do nascido vivo, período 2017 – 2024



Fonte: Elaborado pela autora (2025).

Sob outra perspectiva, a análise individual dos 49 artigos permitiu identificar a presença de 4 artigos relacionados às palavras-chave “time series” e “forecast” (Tabela 2). Esses artigos foram publicados entre 2022 e 2023, com 12 citações acumuladas, conforme a Figura 9.

Figura 9 – Tendência de publicações e citações do aprendizado de máquina para previsão na mortalidade do nascido vivo, período 2022 – 2024



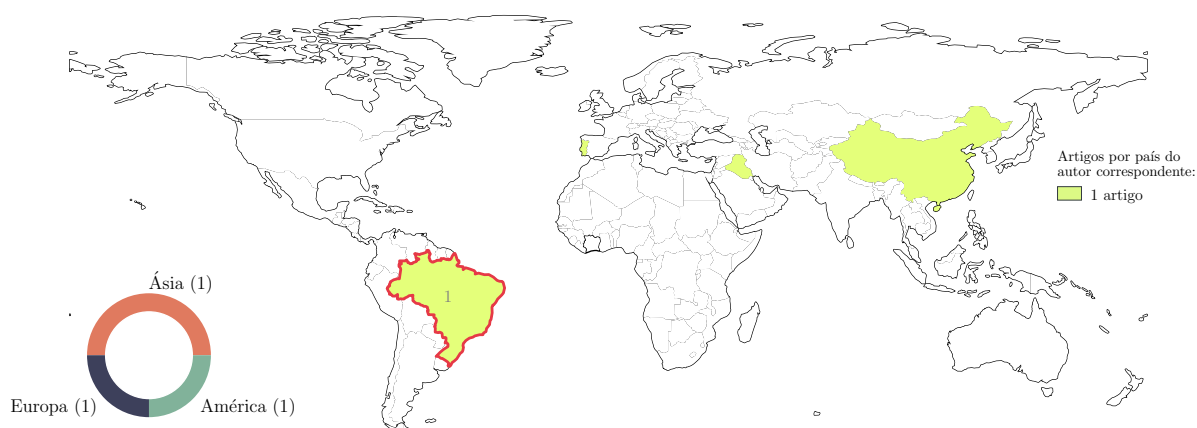
Fonte: Elaborado pela autora (2025).

Embora o número de publicações da Figura 9 ainda seja pequeno, na literatura observou-se que o ML para previsão tem sido aplicado a outros problemas de saúde, com um número substancial de publicações (ver Apêndice B).

Em relação à previsão de séries temporais na mortalidade de nascidos vivos, existem estudos publicados onde métodos tradicionais têm sido utilizados (Tabela 2), sendo indicado que essas técnicas podem ser uma poderosa ferramenta para a tomada de decisão (Silva *et al.*, 2021).

Assim, as quatro pesquisas com a abordagem de ML para previsão foram publicadas pelos países Portugal, Iraque, China e Brasil, conforme apresentado na Figura 10.

Figura 10 – Produção dos países avaliada pelo autor correspondente do aprendizado de máquina para previsão na mortalidade do nascido vivo, período 2022 – 2024



Fonte: Elaborado pela autora (2025).

Complementarmente, para aprofundar a análise dos 49 artigos, cada documento foi classificado de acordo com sua abordagem principal (identificação, previsão, revisão e associação), foram identificadas as técnicas de ML utilizadas, a faixa etária dos dados utilizados e o país de publicação. Os resultados desta análise foram apresentados no Quadro 2.

Quadro 2 – Informações relevantes dos artigos relacionados à aplicação de aprendizado de máquina na mortalidade do nascido vivo

(continua)

Artigo	País	Abordagem	Faixa etária			Técnicas de aprendizado de máquina
			N ^a	PN ^b	I ^c	
Batista <i>et al.</i> (2021)	Brasil	Identificação	✓			Regressão Logística (RLog), Artificial Neural Network (ANN), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), Categorical Boosting (catboost).
Matsushita, Krebs & Carvalho (2022)	Brasil	Identificação	✓			RLog, XGB.
Soares <i>et al.</i> (2022)	Brasil	Identificação			✓	Random Forest (RF), Support Vector Machine (SVM), C4.5, Multilayer Perceptron (MLP).
Rodriguez <i>et al.</i> (2022)	Brasil	Previsão	✓			RF, SVM, XGB, Gradient Boosting (GB), LGBM, catboost, MLP, Seasonal Autoregressive Integrated Moving Average (SARIMA).
Colodette <i>et al.</i> (2021)	Brasil	Associação			✓	XGB.
Garcia <i>et al.</i> (2023)	Brasil	Associação	✓			Classification and Regression Tree (CART).
Cooper, Minneci & Deans (2018)	Estados Unidos	Identificação	✓			Super Learner (SL).
Lee <i>et al.</i> (2021)	Estados Unidos	Identificação			✓	RF.
Teji <i>et al.</i> (2022)	Estados Unidos	Identificação	✓		✓	Adaptive Boosting (AdaBoost), XGB, RF, RLog.
Yoneda <i>et al.</i> (2023)	Estados Unidos	Identificação	✓			RLog.
Shukla & Carlo (2022)	Estados Unidos	Identificação	✓			GB.
Shukla <i>et al.</i> (2024)	Estados Unidos	Associação			✓	Autoregressive Integrated Moving Average (ARIMA) e ML.
Deepika & Chowhan (2020)	Índia	Revisão	✓			–
Sivarajan, Aditya & Sivasankar (2022)	Índia	Identificação	✓			Ensemble

Faixa etária (a) N: neonatal, (b) PN: pós-neonatal e (c) I: infantil.

Fonte: Elaborado pela autora (2025).

Quadro 2 – Informações relevantes dos artigos relacionados à aplicação de aprendizado de máquina na mortalidade do nascido vivo
(continua)

Artigo	País	Abordagem	Faixa etária			Técnicas de aprendizado de máquina
			N ^a	PN ^b	I ^c	
Shaw, Pachpor & Sankaranarayanan (2023)	Índia	Identificação			✓	RF, CART, SVM, RLog, ANN, Naive Bayes (NB), Deep Neural Network (DNN), XGB, GB, LGBM.
Brahma & Mukherjee (2022)	Índia	Identificação	✓		✓	–
Zhang, Xiao & Luo (2023)	Índia	Identificação			✓	RF, RLog, ANN, Linear Discriminant Analysis (LDA) e XGB.
Iqbal <i>et al.</i> (2023)	Índia	Identificação	✓			RLog
Rinta-Kosk <i>et al.</i> (2017)	Finlândia	Identificação			✓	Gaussian Processes Classifier (GPC)
Jaskari <i>et al.</i> (2020)	Finlândia	Identificação			✓	RF, CART, SVM, RLog.
Villagrana-Bañuelos <i>et al.</i> (22)	México	Associação	✓			Agoritmo genético
Villagrana-Bañuelos <i>et al.</i> (2022)	México	Identificação			✓	RF, SVM, RLog, NB.
Rahman <i>et al.</i> (2021)	Bangladesh	Identificação			✓	RF, CART, SVM, RLog.
Rahman <i>et al.</i> (2022)	Bangladesh	Identificação			✓	RF, CART, SVM e RLog.
Sayeed <i>et al.</i> (2024)	Bangladesh	Identificação			✓	SVM e RLog.
Sheikhtaheri <i>et al.</i> (2021)	Irã	Identificação	✓			RF, SVM, ANN, C5.0, Chi-squared Detection of Automatic Interaction (CHAID), Bayesian network (Bnet), Ensemble.
Hajipour <i>et al.</i> (2021)	Irã	Identificação			✓	RF, SVM, RLog, ANN, K-Nearest Neighbors (KNN), NB, LGBM.
Yang <i>et al.</i> (2024)	Irã	Identificação	✓			RF, RLog, ANN, KNN, GB e Decision tree (DT).
Kefi, Aloui & Naceur (2019)	Tunísia	Identificação	✓			RF, CART, SVM, RLog, LDA, KNN, NB.

Faixa etária (a) N: neonatal, (b) PN: pós-neonatal e (c) I: infantil.

Fonte: Elaborado pela autora (2025).

Quadro 2 – Informações relevantes dos artigos relacionados à aplicação de aprendizado de máquina na mortalidade do nascido vivo
(continua)

Artigo	País	Abordagem	Faixa etária			Técnicas de aprendizado de máquina
			N ^a	PN ^b	I ^c	
Ogallo <i>et al.</i> (2020)	Quênia	Identificação	✓			GB.
Hsu <i>et al.</i> (2021)	Taiwan	Identificação	✓			RF, SVM, ANN, DNN, XGB, bagging, Generalized Linear Models and Elastic Nets (GLMNET).
Shariat <i>et al.</i> (2024)	Taiwan	Identificação	✓			KNN, NB, DT, RLog, RF, Bagging, GB e SVM.
Khurshid <i>et al.</i> (2021)	Canadá	Identificação	✓		✓	RF, SVM, RLog, ANN, Ensemble, KNN.
Saravanou <i>et al.</i> (2021)	Grécia	Identificação			✓	GB, NB, ANN, SVM, RLog.
Kovacs <i>et al.</i> (2021)	Reino Unido	Identificação	✓			GLMNET.
Dereje <i>et al.</i> (2021)	Etiópia	Identificação	✓		✓	Bagging, AdaBoost
Mfateneza <i>et al.</i> (2022)	Ruanda	Identificação			✓	RF, CART, SVM, RLog.
Choi <i>et al.</i> (2022)	Coreia do Sul	Identificação	✓			RF, CART, RLog, CHAID, XGB.
Do, Moon & Jin (2022)	Coreia do Sul	Identificação	✓			RF, SVM, RLog, ANN.
Na <i>et al.</i> (2023)	Coreia do Sul	Identificação	✓			RF, SVM, RLog, <i>Ensemble</i> , AdaBoost e ANN.
Kim, Kim & Park (2024)	Coreia do Sul	Identificação	✓			RF, RLog e XGB.
Oh <i>et al.</i> (2024)	Coreia do Sul	Identificação	✓			RF, GB e LGBM.
Salih <i>et al.</i> (2022)	Iraque	Previsão			✓	ANN.
Chen, He & Wang (2022)	China Coreia Mongólia Japão Taiwan	Previsão		✓		ANN.
Yoneda <i>et al.</i> (2023)	Japão	Identificação			✓	RLog.
Ventura <i>et al.</i> (2023)	Portugal	Previsão	✓			ANN, Autoregressive Moving Average (ARMA), Fuzzy model.

Faixa etária (a) N: neonatal, (b) PN: pós-neonatal e (c) I: infantil.

Fonte: Elaborado pela autora (2025).

Quadro 2 – Informações relevantes dos artigos relacionados à aplicação de aprendizado de máquina na mortalidade do nascido vivo (conclusão)

Artigo	País	Abordagem	Faixa etária			Técnicas de aprendizado de máquina
			N ^a	PN ^b	I ^c	
Lourenço <i>et al.</i> (2024)	Portugal	Associação			✓	Least Absolute Shrinkage and Selection Operator (LASSO) e KNN.
Espinola-Sánchez <i>et al.</i> (2023)	Perú	Identificação	✓			SVM, RLog, <i>Ensemble</i> , GB e DT.
Alsubaie, Omar & Riaz (2023)	Arábia Saudita	Identificação			✓	RLog

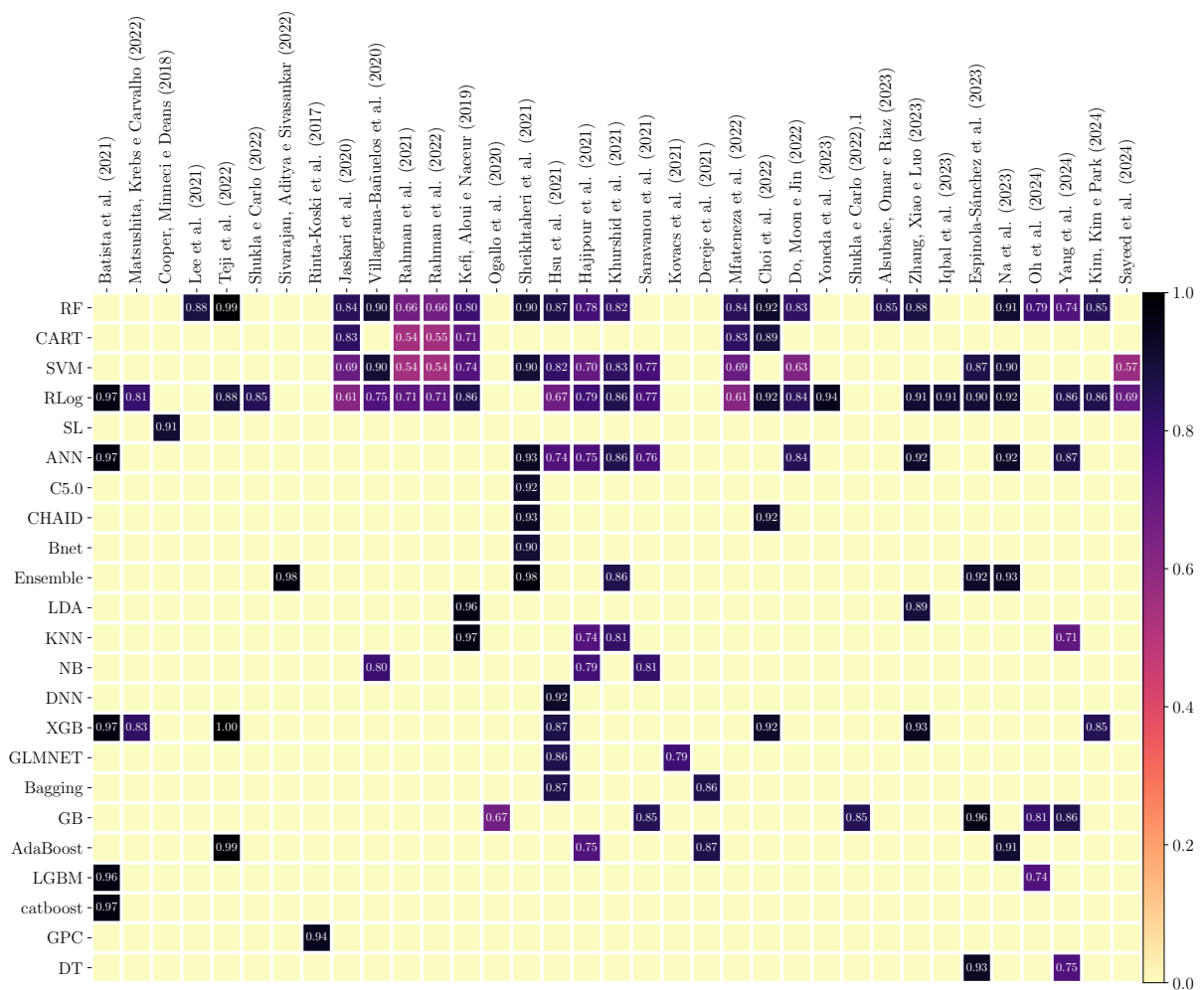
Faixa etária (a) N: neonatal, (b) PN: pós-neonatal e (c) I: infantil.

Fonte: Elaborado pela autora (2025).

Analisando-se o Quadro 2, identificou-se que dos 49 artigos analisados, um artigo teve como foco a revisão da literatura sobre ML (Deepika; Chowhan, 2020), e cinco tiveram como foco principal a análise de fatores associados à mortalidade de nascidos vivos por meio de árvore de decisão (Colodette *et al.*, 2021; Garcia *et al.*, 2023), ARIMA (Shukla *et al.*, 2024), LASSO, KNN (Lourenço *et al.*, 2024), e algoritmo genético (Villagrana-Bañuelos *et al.*, 22). Dos 44 artigos restantes, 40 focaram principalmente no desenvolvimento de modelos de predição e 4 em modelo para previsão.

Dos 40 artigos orientados ao desenvolvimento e avaliação de modelos de ML para identificação de recém-nascidos com risco de óbito, 29 foram orientados para o óbito neonatal, 1 para o óbito pós-neonatal e 23 para o óbitos infantil (Quadro 2). Além disso, a análise permitiu identificar as técnicas testadas e suas métricas de desempenho, conforme está apresentado na Figura 11.

Figura 11 – Desempenho dos modelos para identificar o risco de mortalidade do nascido vivo, de acordo com a AUC



AUC: Área sob a Curva (*Area Under the Curve*).

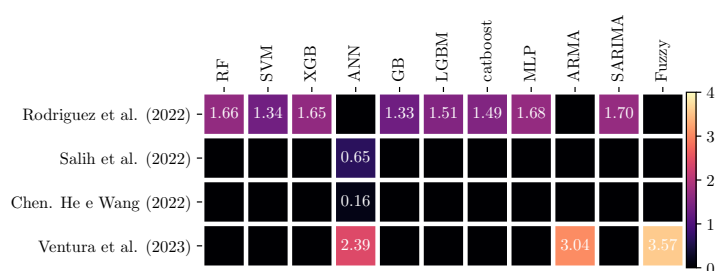
Fonte: Elaborado pela autora (2025).

Conforme ilustrado na Figura 11, foram testadas diferentes técnicas de ML para identificação

de recém-nascidos com risco de óbito, e o desempenho dos modelos ficou na faixa de 0,54 a 0,99. Além disso, identificou-se que as técnicas de RLog, RF, SVM, DT e CART apresentaram o maior número de vezes que foram testadas.

Por outro lado, nos artigos publicados sobre previsão de mortalidade do nascido vivo foram testadas técnicas clássicas de séries de tempo, bem como técnicas de ML (Rodriguez *et al.*, 2022; Salih *et al.*, 2022; Chen; He; Wang, 2022; Ventura *et al.*, 2023), conforme mostrado na Figura 12.

Figura 12 – Desempenho dos modelos para prever a mortalidade, de acordo com o RMSE



RMSE: Raiz do erro quadrático Médio (Root Mean Squared Error)

Fonte: Elaborado pela autora (2025).

De acordo com os resultados revisados nas publicações, os valores obtidos na métrica de erro variaram entre 0,16 e 3,57, sendo a técnica mais utilizada as baseadas em redes neurais, conforme apresentado na Figura 12.

Especificamente, Salih *et al.* (2022) utilizaram redes neurais para prever trimestralmente o número de óbitos por 7 faixas etárias diferentes e suas causas, incluindo previsões de menores de um ano a maiores de 80 anos no Norte do Iraque. Outrossim, em Chen, He & Wang (2022) desenvolveram um modelo híbrido com redes neurais e ARIMA para prever a incidência e mortalidade do vírus da imunodeficiência humana durante o período pós-neonatal em nascidos vivos infectados no Leste Asiático.

Em contrapartida, Ventura *et al.* (2023) testaram os modelos ARMA, fuzzy e redes neurais para prever taxas mensais de mortalidade neonatal em Portugal. Além disso, em Rodriguez *et al.* (2022) testaram as técnicas clássicas ARIMA e SARIMA com algumas técnicas de ML para prever as taxas mensais de mortalidade neonatal na região metropolitana do Vale do Paraíba e Litoral Norte no Brasil.

Assim, as informações obtidas na análise da literatura permitiram visualizar o panorama geral das pesquisas que aplicaram análise espacial e ML na área da saúde. Portanto, fica evidente que há uma grande oportunidade de pesquisa nesta área de estudo, tanto na aplicação de ML para prever a mortalidade quanto para identificar recém-nascidos em risco de morte.

Finalmente, com base nos artigos relacionados com esta pesquisa, analisados na fase I e II desta Seção, pode-se afirmar que no contexto brasileiro a aplicação de tecnologias como SIG e ML na mortalidade do recém nascido ainda é um campo pouco explorado, e a oportunidade de pesquisa é maior na região RMVPLN.

1.4 DELIMITAÇÃO DA PESQUISA

Entre as limitações deste estudo, destaca-se a utilização de técnicas de aprendizado supervisionado, portanto, outras abordagens de ML que poderiam ser exploradas para análises complementares não foram abordadas.

Os dados analisados foram extraídos de domínios públicos do Ministério da Saúde, que fornece uma grande quantidade de dados que auxiliam positivamente na gestão e melhoria do Sistema Único de Saúde. Além disso, os registros utilizados abrangeram o período de 2010 a 2020 para informações relacionadas aos nascidos vivos e de 2010 a 2021 para dados de óbitos.

Os resultados foram validados especificamente para a previsão de séries temporais de mortalidade infantil e a identificação precoce de recém-nascidos com risco de óbito na RMVPLN. Dessa forma, as conclusões obtidas não podem ser generalizadas para outros problemas de saúde, nem para outras regiões do estado de São Paulo ou do Brasil. Contudo, a abordagem metodológica adotada pode ser adaptada e aplicada em contextos semelhantes, fornecendo uma base robusta para investigações futuras e potenciais aplicações em cenários relacionados.

Por último, como limitações técnicas, todo o processo de análise, processamento e desenvolvimento do modelo foi realizado utilizando ferramentas de código aberto, como Python v.3.9.7, Jupyter-Notebook v.6.4.6.

1.5 ORGANIZAÇÃO DO TEXTO

Esta tese está estruturada em mais quatro Capítulos, além deste introdutório. No Capítulo 1 é apresentada a contextualização da pesquisa, bem como a formulação das questões de pesquisa, os objetivos gerais e específicos, a justificativa e importância da pesquisa e, por fim, a delimitação da pesquisa.

No Capítulo 2 é apresentado a fundamentação teórica, na qual são descritos os principais aspectos mais importantes relacionados à mortalidade infantil, análise espacial, técnicas de ML, métricas de avaliação, entre outros.

O Capítulo 3 descreve os materiais e métodos utilizados, detalhando o procedimento metodológico adotado nesta investigação, a definição da área geográfica de estudo e a apresentação do banco de dados.

O Capítulo 4 é dedicado para a apresentação e discussão dos resultados, que estão organizados nas seções de análise espacial das taxas de mortalidade, modelo para previsão da mortalidade com séries de tempo e modelo para identificação de risco de óbito.

No Capítulo 5, são apresentadas as conclusões deste trabalho, com a verificação das questões e objetivos da pesquisa, conforme os resultados obtidos. Além disso, este capítulo também propõe sugestões para pesquisas futuras, bem como lista as produções científicas geradas durante o desenvolvimento deste trabalho.

Finalmente estão os Anexos, Apêndices e Bibliografia consultada para a elaboração deste texto.

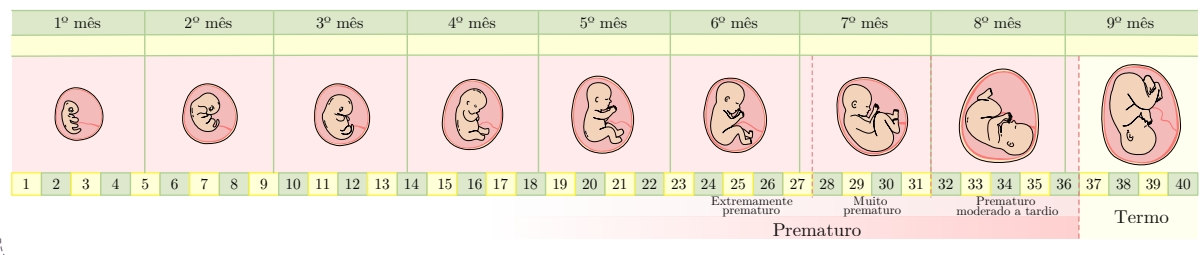
2 FUNDAMENTAÇÃO TEÓRICA

2.1 RECÉM-NASCIDO E SUA VULNERABILIDADE

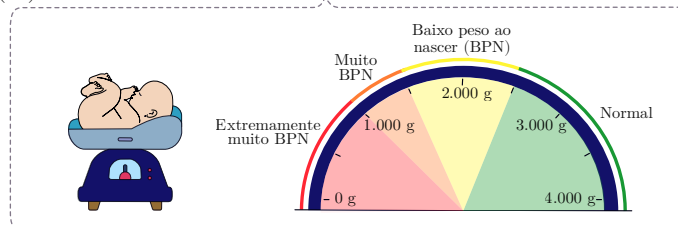
Recém-nascidos com baixo peso ao nascer, prematuros (pré-termo) e Pequeno para a Idade Gestacional (PIG) são considerados vulneráveis, com alto risco de complicações durante a gestação ou durante o parto, além de apresentar risco de infecções (Hsu *et al.*, 2021), morbidade como condições crônicas relacionadas com a obesidade ou diabetes (UNICEF, 2019), e até mortalidade, principalmente no período neonatal (Paixao *et al.*, 2021; Rodriguez *et al.*, 2022). A Figura 13 ilustra a classificação dos recém-nascidos segundo idade gestacional e peso ao nascer.

Figura 13 – Classificação dos recém-nascidos segundo idade gestacional e peso

(A) Idade gestacional



(B) Peso ao nascer



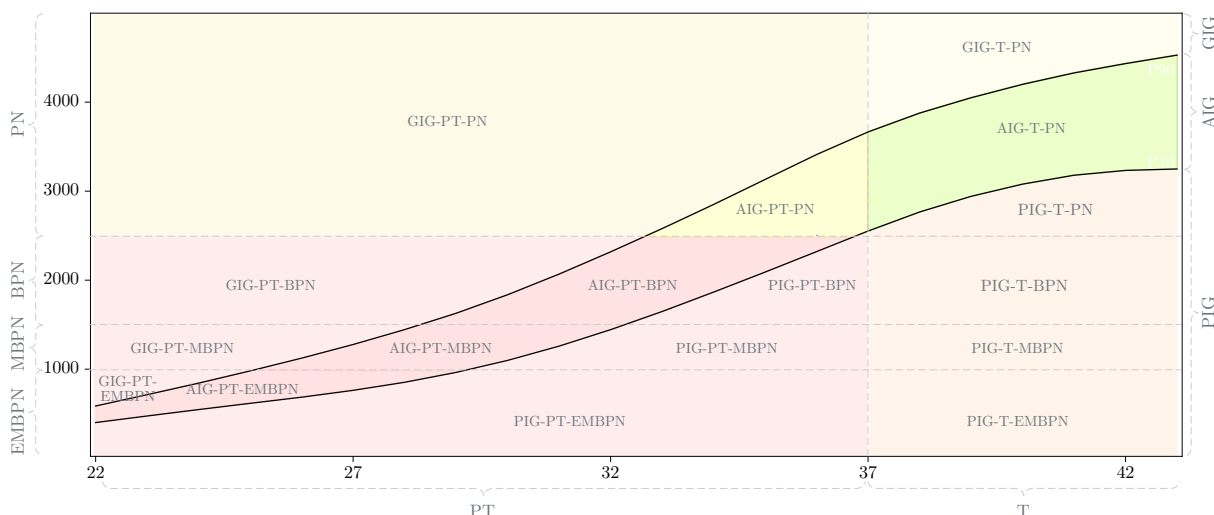
Fonte: Elaborado pela autora (2025).

De acordo com estimativas globais, aproximadamente 20,5 milhões de recém-nascidos a cada ano apresentam baixo peso ao nascer, representando cerca de 15,5% do total de nascidos vivos em todo o mundo (UNICEF, 2019). Além disso, perto de 15 milhões de nascidos vivos são prematuros e cerca de 23 milhões apresentam tamanho inadequado para a idade gestacional (Ashorn *et al.*, 2020). Ademais, aproximadamente 80% dos óbitos neonatais apresentam baixo peso, e nascidos vivos prematuros ou pequenos para a idade gestacional têm maior risco de morte do que os recém-nascidos a termo e com tamanho Adequado para a Idade Gestacional (AIG) (Paixao *et al.*, 2021).

O recém-nascido também pode ser classificado como AIG, PIG e Grande para a Idade Gestacional (GIG), conforme ilustrado na curva de Battaglia & Lubchenco (1967) da idade gestacional e peso do nascido vivo, apresentada na Figura 14.

De acordo com a Figura 14, apenas a categoria de cor verde na ilustração corresponde aos nascidos vivos que não apresentam vulnerabilidade.

Figura 14 – Classificação do recém nascido de acordo com peso e idade gestacional



AIG: Adequado para a idade gestacional, PIG: Pequeno para a idade gestacional, GIG: Grande para a idade gestacional, PT: Pré-termo, T: Termo, PN: Peso normal, BPN: Baixo peso ao nascer, MBPN: Muito baixo peso ao nascer, EMBPN: Extremadamente muito baixo peso ao nascer.

Fonte: Elaborado pela autora (2025).

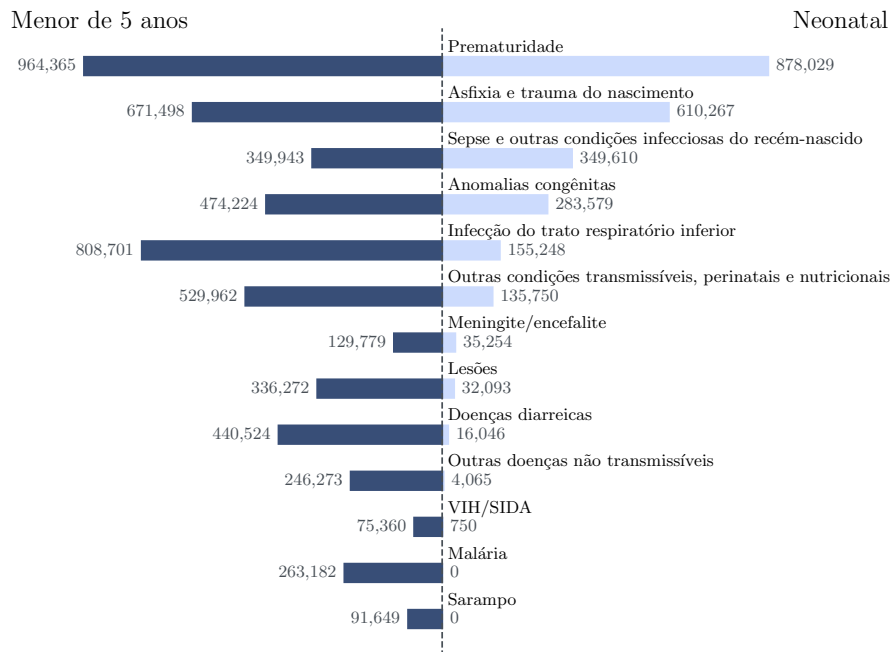
Em adição, a Figura 14 apresenta a combinação da vulnerabilidade do recém-nascido, identificando a presença de baixo peso e prematuridade, permitindo a identificação de outros grupos de vulnerabilidade. Portanto, o grupo não vulnerável de nascidos vivos são aqueles não prematuros, com peso normal e AIG, pois, no caso de considerar apenas os nascidos vivos com baixo peso ao nascer, excluem-se os nascidos vivos prematuros com peso ≥ 2.500 gramas, que também constituem um grupo vulnerável (Ashorn *et al.*, 2020; Paixao *et al.*, 2021).

Por outro lado, além do peso e da prematuridade, a morte do recém-nascido também pode ser decorrente de outras causas durante ou após o parto, que podem ser evitadas por meio de medidas preventivas, intervenções e tratamentos adequados para reduzir os riscos (Hsu *et al.*, 2021; Paixao *et al.*, 2021; Bernardino *et al.*, 2022; Shaw; Pachpor; Sankaranarayanan, 2023).

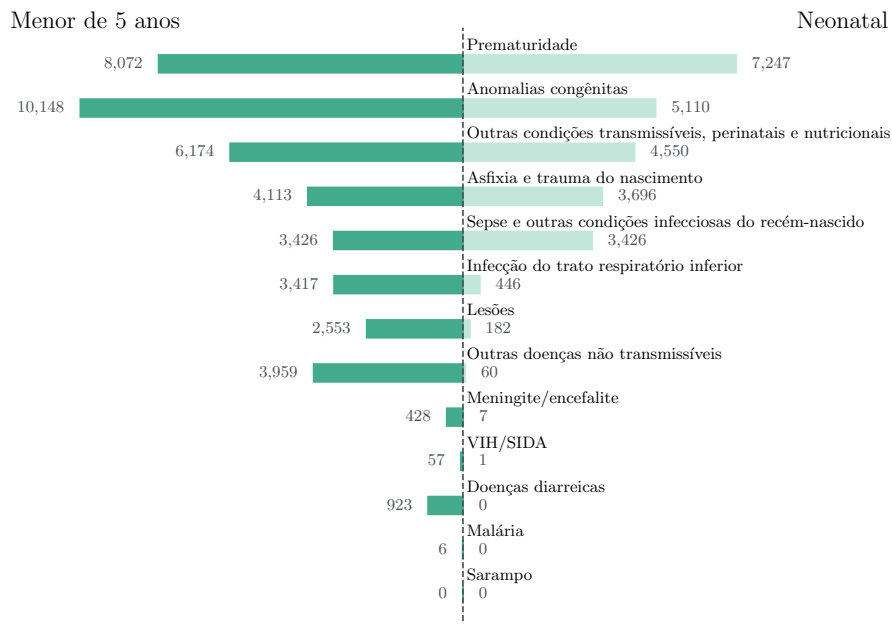
De acordo com a WHO (2018), a nível global, entre as principais causas de morte em crianças menores de cinco anos estão: infecção do trato respiratório, asfixia, trauma no nascimento, anomalias congênitas, doenças diarreicas, sepses e outras condições infecciosas, etc. (WHO, 2018), como apresentado na Figura 15 (A). Considerando as causas de morte neonatal (ocorridas antes dos 28 dias após o nascimento) em todo o mundo, além da prematuridade, as cinco principais causas são a asfixia, trauma no nascimento, sepse, anomalias congênitas, infecção do trato respiratório e condições transmissíveis, perinatais e nutricionais (WHO, 2018).

Figura 15 – Principais causas de morte neonatal e entre crianças menores de 5 anos em 2017

(A) Principais causas a nível mundial em 2017



(B) Principais causas em Brasil em 2017



Fonte: Elaborado pela autora (2025), com dados de WHO (2018).

Adicionalmente, a Figura 15 (B) apresenta as principais causas de óbito neonatal e principais causas de morte em crianças menores de 5 anos no Brasil. A anomalia congênita é a principal causa de morte entre crianças menores de 5 anos no Brasil, além de ser a segunda causa de morte neonatal, seguida pela asfixia e sepses (WHO, 2018).

2.2 TOMADA DE DECISÃO, *BIG DATA* E INTELIGÊNCIA ARTIFICIAL NA ÁREA DE SAÚDE

Numa perspectiva geral, a tomada de decisão consiste no processo de resolver um problema, selecionando a alternativa mais adequada entre várias opções disponíveis. Com frequência, as decisões podem ser tomadas por meio de métodos intuitivos baseados na experiência, o que em muitos casos pode levar a decisões inadequadas, principalmente quando se trata de problemas complexos ou dinâmicos que exigem uma análise mais detalhada, além da utilização de modelos matemáticos (Vercellis, 2011).

Além disso, no contexto dos serviços de saúde, podem ser utilizadas informações relacionadas aos diferentes processos da área para o planejamento estratégico, gerando conhecimentos valiosos para a tomada de decisões (Sin, 2020).

Em conformidade com Vercellis (2011), existe uma relação entre a tomada de decisão e o planejamento, pois um planejamento adequado pode auxiliar positivamente a tomada de decisão, permitindo que os diferentes recursos e/ou tecnologia disponíveis sejam utilizados de forma mais eficiente. Assim, tanto o planejamento quanto as decisões podem ser classificados como estratégicos, táticos e operacionais, conforme descrito no Quadro 3.

Quadro 3 – Classificação da tomada de decisão e planejamento

Tipo	Descrição
Estratégico	• Caracteriza-se por um horizonte de planejamento de longo prazo. • Determina a direção que a entidade seguirá em relação ao problema analisado. • Influencia os objetivos e a formulação de políticas, sendo decisões geralmente tomadas por níveis organizacionais superiores.
Tático	• O horizonte de planejamento é de médio prazo. • Relaciona-se às decisões estratégicas e operacionais, buscando estratégias de planejamento mais específicas. • Decisões são tomadas com base na avaliação de cenários por níveis organizacionais intermediários.
Operacional	• O horizonte de planejamento é de curto prazo, com decisões detalhadas e específicas. • Está diretamente relacionado às atividades realizadas pelos operadores da entidade e ao acompanhamento das ações executadas. • Decisões são tomadas por níveis organizacionais inferiores.

Fonte: Vercellis (2011) e Sin (2020).

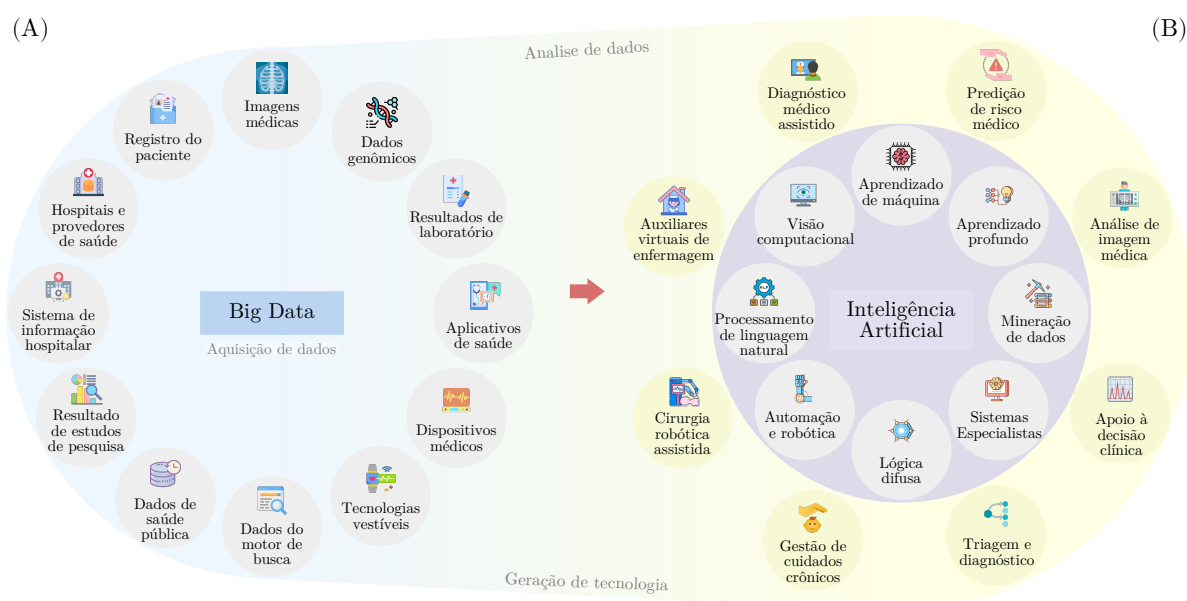
Desta forma, as diferentes tecnologias emergentes podem ser integradas aos diversos processos organizacionais, proporcionando suporte à tomada de decisão em diferentes níveis (estratégico, tático ou operacional) (Vercellis, 2011), aproveitando as informações que são armazenadas em diferentes domínios do serviço de saúde (Sin, 2020).

Neste contexto, a digitalização e integração das Tecnologias de Informação e Comunicação (TICs) em diferentes processos tem gerado um aumento na geração e produção de dados,

resultando em um grande volume de dados brutos, tanto estruturados quanto não estruturados, provenientes de diferentes domínios.

Essa grande geração e armazenamento de dados também são conhecidos como BD, e caracterizando-se pelo seu volume, velocidade, variedade, variabilidade, veracidade e complexidade na vinculação, captura, compartilhamento e análise computacional, com o objetivo de identificar padrões, associações e/ou tendências (Wang; Kung; Byrd, 2018; Ristevski; Savoska, 2021), conforme ilustrado na Figura 16.

Figura 16 – *Big Data* e inteligência artificial na área da saúde



Fonte: Elaborado pela autora (2025).

A geração desse grande volume de dados tem originado um interesse crescente na utilização desses recursos para análise e desenvolvimentos tecnológicos (Benke; Benke, 2018; Nascimento *et al.*, 2021; Matsushita; Krebs; Carvalho, 2022), que possam dar suporte aos especialistas, tornando suas decisões mais oportunas e eficazes (Vercellis, 2011; Sin, 2020).

Na área de saúde, o BD ainda representa um grande desafio, visto que sua análise e utilização em ambientes hospitalares está começando a se tornar realidade, auxiliando na tomada de decisões (Chiavegatto, 2015; Matsushita; Krebs; Carvalho, 2022). Assim, a análise e processamento do BD podem revolucionar a assistência médica e os processos hospitalares, de modo que podem apoiar provedores, médicos especialistas, membros do governo, entre outros, no planejamento e na implementação de medidas de intervenção (Sin, 2020; Nascimento *et al.*, 2021).

Em conformidade com WHO (2021), a análise do BD pode gerar diversos benefícios na assistência médica, como com a exploração e extração de informações relevantes, o desenvolvimento de modelos de previsão de eventos clínicos, o diagnóstico de doenças, entre outros (Figura 16), possibilitando um impacto positivo tanto para os pacientes quanto para os sistemas de saúde.

Em relação aos impactos sociais, observam-se resultados mais significativos quando são tecnologias relacionadas à IA são utilizadas no processamento desses grandes conjuntos de dados

(Benke; Benke, 2018). A IA, portanto, é uma área interdisciplinar que combina habilidades em ciência da computação, com conhecimentos em lógica, matemática e estatística no desenvolvimento de algoritmos complexos com capacidade de aprender com a experiência, simulando a inteligência humana (Benke; Benke, 2018; Rong *et al.*, 2020).

O rápido avanço da tecnologia permitiu a divisão da IA em diferentes ramos de estudo, como mineração de dados, ML, aprendizado profundo, processamento de linguagem natural, visão computacional, sistemas especialistas, lógica difusa e robótica (Rong *et al.*, 2020), conforme ilustrado na Figura 16. Como resultado, a aplicação dessas tecnologias em diversos ambientes hospitalares podem permitir uma medicina personalizada, melhorando a qualidade do atendimento, com redução de custos e aprimoramento dos resultados (Nascimento *et al.*, 2021).

Neste cenário, a OMS incentiva a pesquisa com BD e IA, para o desenvolvimento de ferramentas que possam ser incorporadas aos diferentes processos da medicina diária, compartilhando a descrição dos estudo e o processo metodológico seguido no desenvolvimento (WHO, 2021). Deste modo, no contexto dos ambientes hospitalares pediátricos, e de acordo com estudos da literatura, o uso de tecnologias de BD e IA em aplicações neonatais ou pediátricas não tem sido exceção, embora ainda seja um campo pouco explorado (Matsushita; Krebs; Carvalho, 2022; Malhotra *et al.*, 2023).

2.3 PRÉ-PROCESSAMENTO DE DADOS

O conjunto de dados pode estar constituído por um conjunto de observações também conhecidas como variáveis ou atributos. Esse conjunto é comumente representado por uma matriz de variáveis $\mathbf{X}_{n \times d}$, sendo n o número de linhas (instâncias) do conjunto de dados e d corresponde ao número de variáveis ou à dimensionalidade da matriz (Faceli *et al.*, 2021; Mohr; Wilson; Freund, 2021). A Equação 2.1 descreve a matriz $\mathbf{X}_{n \times d}$, sendo $x_{i,j}$ uma característica da variável j e da linha i :

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & \cdots & x_{n,d} \end{bmatrix} \quad (2.1)$$

As variáveis da matriz $\mathbf{X}_{n \times d}$ podem conter informações não numéricas, também conhecidas como variáveis qualitativas ou categóricas, que podem ser classificadas como nominais ou ordinais; caso as variáveis contenham informações numéricas são conhecidas como variáveis quantitativas, que podem ser classificadas como discretas ou contínuas (Mohr; Wilson; Freund, 2021). O Apêndice C apresenta os fundamentos da análise exploratória de dados.

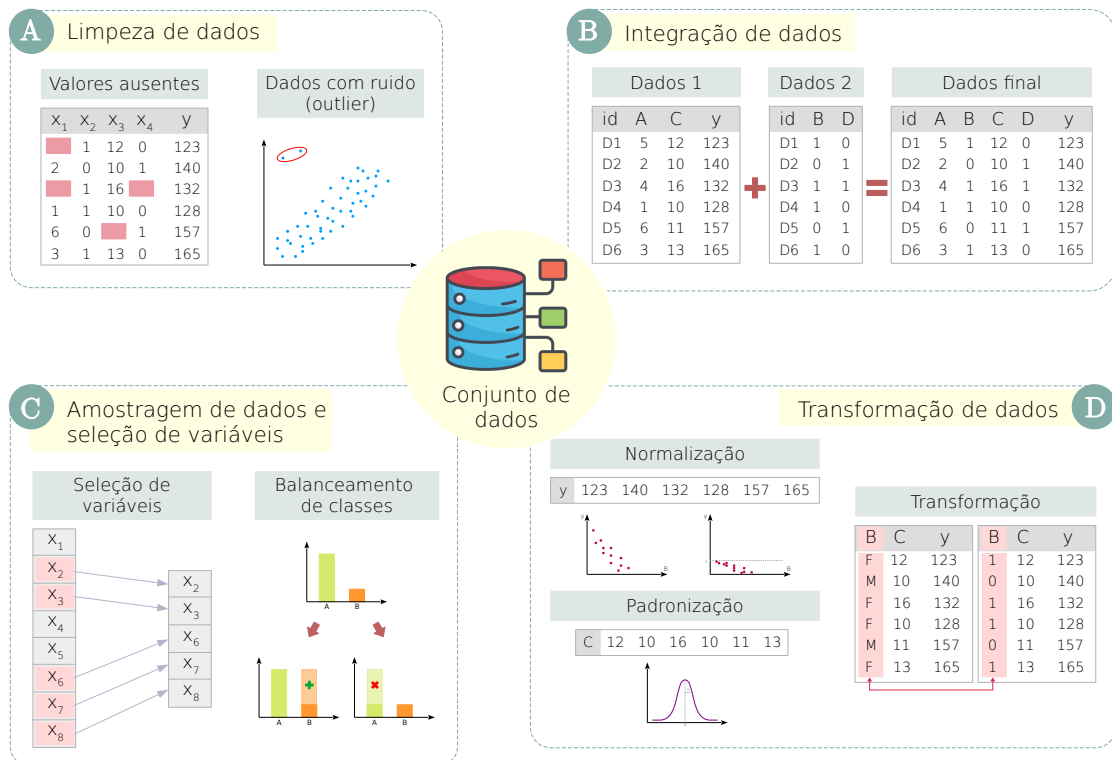
Extraír conhecimento do conjunto de dados não requer apenas da visualização e compreensão do problema que está sendo estudado, também é necessário da verificação da qualidade dos dados

coletados, pois dados de baixa qualidade podem levar a modelos de baixa qualidade (Faceli *et al.*, 2021; Han; Pei; Tong, 2022).

O processo de melhoria da qualidade dos dados, também conhecido como pré-processamento, abrange diversas técnicas descritas na literatura, que visam aprimorar a qualidade dos dados, facilitar seu uso e aumentar a eficiência dos modelos (Han; Pei; Tong, 2022).

Como observado na Figura 17, as técnicas de pré-processamento podem ser classificadas em quatro grupos, que não são mutuamente exclusivos.

Figura 17 – Técnicas para pré-processamento de dados



Fonte: Elaborado pela autora (2025).

2.3.1 Limpeza de Dados

Dados coletados de processos reais podem apresentar valores ausentes ou valores ruidosos (Figura 17 (A)), frequentemente causados por problemas na coleta, armazenamento ou preenchimento das informações, o que compromete sua qualidade.

- **Valores incompletos (ausentes):**

A presença de dados incompletos em algumas variáveis do conjunto de dados é frequentemente observada em bancos de dados reais. O processamento de conjuntos de dados com valores ausentes é essencial, visto que, existem técnicas de ML que podem apresentar erros quando lidam com variáveis incompletas (Faceli *et al.*, 2021). Assim, dentre as ações que podem ser tomadas para lidar com esse problema nos dados, destacam-se as seguintes alternativas:

- Eliminar linhas (se necessário variáveis) que possuem valores ausentes, embora essa abordagem não seja recomendada para pequenos conjuntos de dados.
- Identificar e preencher as células que possuem valores ausentes, mas não é altamente recomendado quando o conjunto de dados é pequeno ou o número de valores ausentes é muito grandes.
- Preencher células vazias utilizando métodos ou heurísticas, como:
 - * Inserir um valor que indique explicitamente a ausência de informação;
 - * Utilizar medidas estatísticas como média, mediana ou moda;
 - * Aplicar técnicas de ML, como regressão ou árvores de decisão, para estimar os valores ausentes, entre outros.
- **Valores com ruído:** A presença de ruído no conjunto de dados pode gerar inconsistência nas informações e comprometer os resultados obtidos. A análise de dados por meio de estatística descritiva ou técnicas de visualização pode ajudar a identificar a possível presença de ruído (Han; Pei; Tong, 2022). Assim, além da análise estatística e da visualização de dados, existem outros métodos para a identificação de valores *outliers*, como:

- Análise baseada no Intervalo Interquartílico, em inglês *Interquartile Range* (IQR), sendo uma medida robusta de dispersão que utiliza os quartis para definir limites (Han; Pei; Tong, 2022). O IQR, ou amplitude interquartílica, é calculado com a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1). Dessa forma, valores que excedem os limites inferiores ou superiores são identificados como *outliers*. Esses limites são determinados pela Equação 2.2:

$$\begin{aligned} \text{Limite Inferior} &= Q_1 - 1,5 \times IQR \\ \text{Limite Superior} &= Q_3 + 1,5 \times IQR \end{aligned} \quad (2.2)$$

- Análise de percentis, frequentemente aplicada em contextos onde os dados apresentam distribuições assimétricas (Faceli *et al.*, 2021; Han; Pei; Tong, 2022). Essa medida estatística oferece pontos de corte ao dividir os dados ordenados em 100 partes iguais. Da mesma forma, valores que excedem os limites inferior ou superior, como o 1º e o 99º percentil, são identificados como *outliers* (Han; Pei; Tong, 2022). Os limites são calculados pela Equação 2.3.

$$\begin{aligned} \text{Limite Inferior} &= P_1 \\ \text{Limite Superior} &= P_{99} \end{aligned} \quad (2.3)$$

Os percentis P_1 e P_{99} são calculados pela Equação 2.4:

$$P_k = \frac{k}{100} \times n + 1 \quad (2.4)$$

Assim, após a identificação dos valores ruidosos, dentre as alternativas para lidar com este problema, destaca-se a depuração das linhas ou colunas, no entanto esta abordagem pode levar à perda de informações valiosas (Faceli *et al.*, 2021; Han; Pei; Tong, 2022). Outras alternativas são a transformação dos dados ou em alguns casos a imputação de valores utilizando medidas como média, mediana, moda ou uso de técnicas de regressão ou classificação para estimar o valor que substituirá o ruído (Faceli *et al.*, 2021). No caso de imputação, se os valores selecionados para imputar *outliers* não forem representativos, pode levar a modelos tendenciosos.

2.3.2 Integração de Dados

Os dados podem ser coletados de diferentes fontes, dando a possibilidade de combiná-los em um único conjunto através da identificação de um atributo identificador (Faceli *et al.*, 2021), conforme ilustrado na Figura 17 (B). Além disso, na integração de dados é importante considerar e detectar a presença de redundância de variáveis que podem ser identificados através da técnica de correlação (Han; Pei; Tong, 2022).

2.3.3 Seleção de Variáveis e Balanceamento

A eficiência dos modelos de ML está, em grande parte, relacionada ao número de instâncias e variáveis do conjunto de dados. A seleção de variáveis e balanceamento de classes, da Figura 17 (C), são processos importantes, que podem ser executados para melhorar a qualidade do conjunto de dados durante o treinamento dos modelos (Faceli *et al.*, 2021).

- **Seleção de variáveis:**

Existem casos em que as técnicas de ML lidam melhor com conjuntos de dados que possuem um número não tão grande variáveis. Portanto, uma forma de lidar com esse problema é eliminando as variáveis menos relevantes ou combinando informações entre variáveis, também conhecidas como seleção de variáveis e agregação respectivamente (Faceli *et al.*, 2021).

De acordo com Faceli *et al.* (2021) e Han, Pei & Tong (2022), a agregação e seleção de variáveis são descritas abaixo:

- O método de agregação, com a Análise de Componentes Principais, em inglês *Principal Component Analysis* (PCA), é uma das técnicas mais conhecidas para a combinação de variáveis. Esta técnica correlaciona as informações de cada variável, eliminando a presença de redundância e reduzir a dimensionalidade.
- O método de seleção de variáveis é frequentemente empregado em bancos de dados reais devido à presença de variáveis redundantes, irrelevantes, ou com alto nível de ruído ou dados ausentes. Portanto, para identificar e selecionar as variáveis mais importantes, podem ser utilizados métodos como filtragem, embutido e *wrapper*.

Especificamente, no método de filtragem, as variáveis são ordenadas de acordo com um valor de relevância, como a análises de correlação ou associação, para selecionar aquelas com maior importância. No método embutido, a seleção de variáveis é incorporada ao processo de treinamento do modelo, com técnicas como DT ou regressão linear (com regularização $L1$, como na regressão LASSO) permitem a seleção de variáveis durante o treinamento. Do mesmo modo, no método *wrapper*, a seleção de variáveis também é incorporada ao processo de treinamento, avaliando subconjuntos de variáveis com técnicas de ML como caixa preta. O método *wrapper* geralmente envolve a utilização de técnicas de otimização com busca pelo melhor subconjunto de atributos, mas pode apresentar um alto custo computacional.

- ***Balanceamento de classes:***

Em problemas reais de classificação, é comum encontrar classes desequilibradas, ou seja, um número significativamente diferente de observações em cada classe da variável alvo.

Para ajustar a distribuição das classes e reduzir o viés, podem ser utilizadas técnicas de amostragem como sub-amostragem (*undersampling*) e sobre-amostragem (*oversampling*), que possuem comportamento equivalente no processo de balancear as classes (Bekkar; Alitouche, 2013). As técnicas são descritas a seguir, conforme Bekkar & Alitouche (2013) e Faceli *et al.* (2021):

- A sub-amostragem consiste em realizar uma seleção aleatória das observações da classe majoritária, a fim de emparelhar a quantidade de instâncias entre todas as classes com a classe minoritária.
- A sub-amostragem, de forma equivalente, as classes da variável alvo são emparelhadas duplicando aleatoriamente algumas observações da classe minoritária. No entanto, essa técnica pode levar ao overfitting, pois o modelo pode memorizar as observações duplicadas.

Além disso, é importante destacar que a escolha entre as técnicas de sub-amostragem e sobre-amostragem depende do problema de estudo e do conjunto de dados.

2.3.4 Transformação de Dados

Existem algumas técnicas de transformação como normalização, padronização ou transformação categórica (Figura 17 (D)), que podem ser utilizadas para manipular os dados, considerando o tipo de variável (categórico ou numérico) do conjunto de dados (Mohr; Wilson; Freund, 2021; Faceli *et al.*, 2021).

- **Normalização Min – Max:**

A técnica de normalização Min – Max é utilizada para redimensionar os dados de uma variável dentro do intervalo $[0 - 1]$. Os dados normalizados da variável j são calculados pela Equação 2.5:

$$x'_{i,j} = \frac{x_{i,j} - x_{\min,j}}{x_{\max,j} - x_{\min,j}} \quad (2.5)$$

sendo $x'_{i,j}$ o valor normalizado entre 0 e 1 de $x_{i,j}$, $x_{\min,j}$ valor mínimo e $x_{\max,j}$ valor máximo da variável (Han; Pei; Tong, 2022). Por outro lado, a inversão da normalização da variável j é calculado pela Equação 2.6.

$$x_{i,j} = \left[x'_{i,j} \times (x_{\max,j} - x_{\min,j}) \right] + x_{\min,j} \quad (2.6)$$

na qual, os valores de $x_{\min,j}$ e $x_{\max,j}$ são os mesmos valores usados no processo de normalização da Equação 2.5.

- **Padronização:**

A padronização de dados, também conhecida como normalização $z - score$, é utilizada para o reescalamento dos dados, garantindo a propriedade da distribuição normal de $\bar{x}_j = 0$ e $S_j = 1$ (Faceli *et al.*, 2021). A padronização de dados é estimada utilizando a Equação 2.7:

$$x''_{i,j} = \frac{x_{i,j} - \bar{x}_j}{S_j} \quad (2.7)$$

sendo $x''_{i,j}$ o valor padronizado de $x_{i,j}$. Além disso, $\bar{x}_j$ e S_j são o valor médio e desvio padrão da variável (Faceli *et al.*, 2021; Han; Pei; Tong, 2022). Da mesma forma, caso seja necessária a inversão da padronização da variável j , pode ser calculada conforme a Equação 2.8:

$$x_{i,j} = (x''_{i,j} \times S_j) + \bar{x}_j \quad (2.8)$$

sendo os valores de $\bar{x}_j$ e S_j os mesmos do processo de padronização da Equação 2.7.

- **Transformação categórica:**

O tipo de transformação aplicada na variável categórica depende se é do tipo de variável nominal ou ordinal. No caso de variáveis categóricas nominais, as informações são transformadas utilizando-se o esquema de codificação *one-hot*, também conhecida como transformação *one-of-k*, que cria variáveis binárias para cada uma das categorias da variável (Faceli *et al.*, 2021).

Nas variáveis categóricas ordinais, as informações são transformadas utilizando-se a ordem hierárquica da variável, transformando as informações com o esquema de codificação de 0 a $k - 1$ categorias (Han; Pei; Tong, 2022).

2.4 TÉCNICAS DE ANÁLISE ESPACIAL

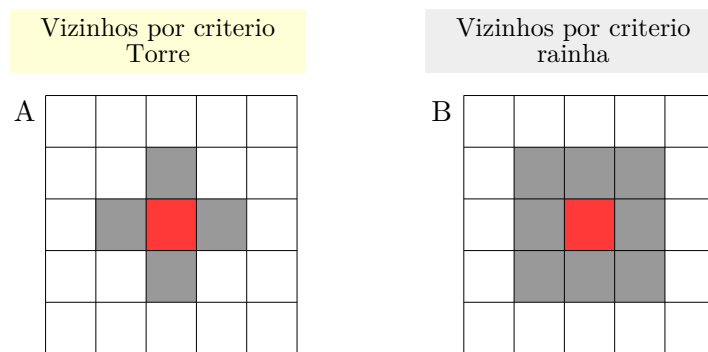
O termo dependência espacial refere-se ao grau de autocorrelação espacial existente entre valores independentes observados em uma determinada área geográfica (Cupido; Jevtic; Paez, 2020). Com a análise de autocorrelação espacial é possível analisar o comportamento de uma variável de estudo em um espaço geográfico a fim de identificar padrões espaciais existentes, analisar seu comportamento e distribuição, mensurar até que ponto as subáreas vizinhas podem ser afetadas, além de identificar aglomerações ou pontos atípicos (Almeida, 2012).

Assim, as subáreas vizinhas podem ser definidas como os municípios que compartilham limites, e com a análise de autocorrelação é possível modelar suas relações espaciais (Anselin; Rey, 2014; Cupido; Jevtic; Paez, 2020). Para isso, é utilizada a matriz de vizinhança, também conhecida como matriz de ponderação ou matriz de contiguidade, que pode ser representada conforme a matriz 2.9 (**W**) :

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{bmatrix} \quad w_{i,j} = \begin{cases} w_{i,j} > 0, & \text{Se } i \text{ e } j \text{ são vizinhos.} \\ w_{i,j} = 0, & \text{Caso contrário.} \end{cases} \quad (2.9)$$

sendo **W** uma matriz $n \times n$ e $w_{i,j}$ são os pesos espaciais. Os valores de $w_{i,j}$ são diferentes de zero quando i e j são vizinhos, de modo que, os $w_{i,j}$ diagonais são iguais a zero (Anselin; Rey, 2014). No entanto, a identificação dos vizinhos da matriz de contiguidade pode ser calculada pelos critérios de torre (*rook*) e rainha (*queen*), conforme observado na Figura 18.

Figura 18 – Critério de torre e rainha



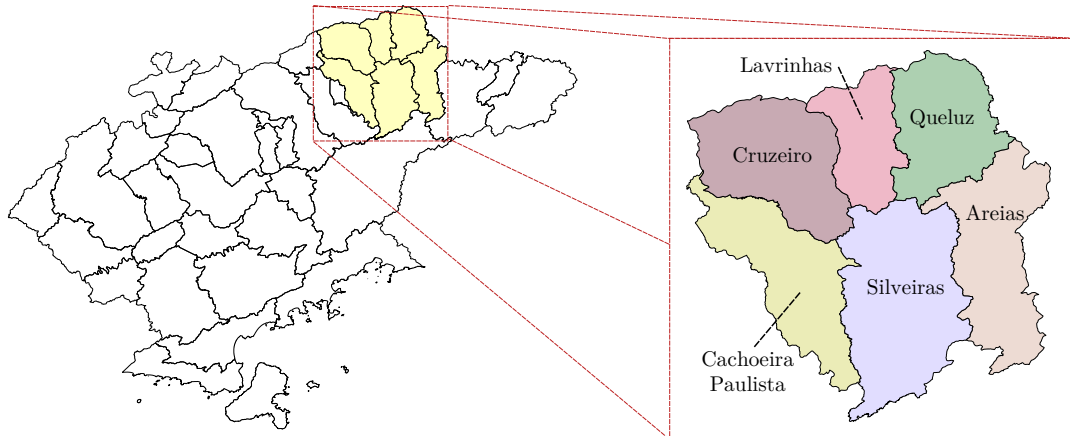
Fonte: Elaborado pela autora (2025).

Conforme a Figura 18, com o critério torre (A), são considerados vizinhos de um município central aqueles municípios que compartilham uma aresta comum; pelo contrário, com o critério rainha (B), todos os municípios que compartilham uma aresta ou vértice com o município central são considerados vizinhos (Anselin; Rey, 2014).

Assim, a Figura 19 apresenta um exemplo de cálculo de matriz de contiguidade normalizada $[0 - 1]$, mostrando a matriz de pesos espaciais de seis municípios da RMVPLN. No caso do

município de Cruzeiro, seus municípios vizinhos e seus pesos normalizados foram: Lavrinhas (0,33), Cachoeira Paulista (0,33) e Silveiras (0,33). Note-se que a soma dos pesos espaciais dos vizinhos de cada município é igual a um.

Figura 19 – Exemplo de matriz de vizinhos de municípios da RMVPLN



	Cruzeiro	Lavrinhas	Queluz	Cachoeira Paulista	Silveiras	Areias
Cruzeiro	0	0,33	0	0,33	0,33	0
Lavrinhas	0,33	0	0,33	0	0,33	0
Queluz	0	0,33	0	0	0,33	0,33
Cachoeira Paulista	0,50	0	0	0	0,5	0
Silveiras	0,20	0,20	0,20	0,20	0	0,20
Areias	0	0	0,50	0	0,50	0

Fonte: Elaborado pela autora (2025).

Por outro lado, na literatura existem diferentes indicadores estatísticos que podem ser utilizados para medir o grau de associação espacial existente em uma área estudo. Um dos indicadores mais populares de autocorrelação espacial é o índice de Moran, que usa a medida de autocovariância dos dados na forma cruzada ($x_i \times x_j$), e que pode ser avaliado da forma univariada e bivariada, além de poder ser avaliada em uma perspectiva global e local (Almeida, 2012), apresentados em 2.4.1 e 2.4.2. Adicionalmente, em 2.4.3 foi apresentada a suavização bayesiana empírico, utilizado para suavizar as taxas.

2.4.1 Autocorrelação Espacial Univariada

O índice de Moran global univariado (I_U) é um indicador estatístico utilizado para medir o grau de associação espacial na área analisada, testando a hipótese nula ($H_0 = 0$) de independência espacial, ou seja, a ausência de autocorrelação espacial (Cliff; Ord, 1981; Cupido; Jevtic; Paez, 2020). Assim, I_U pode ser definida conforme apresentado na Equação 2.10:

$$I_U = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (2.10)$$

sendo n o número total de municípios da área analisada, $\bar{x}$ é a média global da variável de análise x normalizada, e w_{ij} é o peso espacial que reflete a proximidade espacial entre os municípios i e j (Anselin; Rey, 2014; Bivand; Wong, 2018).

Além disso, o valor do I_U varia entre -1 e $+1$, e, valores de I_U maiores ou menores que o valor esperado ($E(I_U) = -[1/(n-1)]$) indicam a existência de correlação direta e inversa, respectivamente (Bivand; Wong, 2018). Portanto $I_U > E(I_U)$ evidencia uma autocorrelação espacial positiva, $I_U < E(I_U)$ indica uma autocorrelação espacial negativa, e $I_U = E(I_U)$ indica ausência de autocorrelação ou aleatoriedade (Cliff; Ord, 1981; Bivand; Wong, 2018).

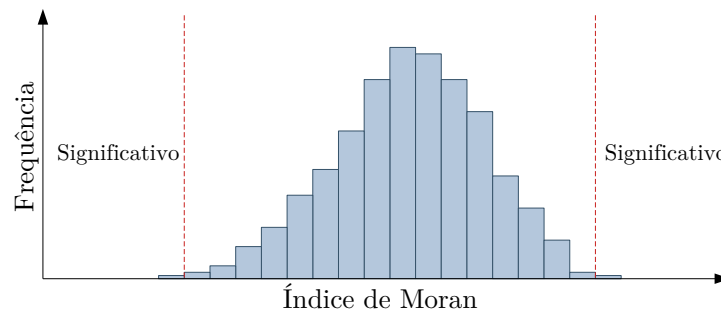
Em relação à significância estatística de I_U , a forma mais simples é pela pressuposição de normalidade de $Z(I_U)$, com variância unitária ($\sigma^2 = 1$) e média igual a zero ($\mu = 0$). O $Z(I_U)$ é calculado conforme a Equação 2.11:

$$Z(I_U) = \frac{I_U - E(I_U)}{DP(I_U)} \quad (2.11)$$

sendo $E(I_U)$ o valor esperado e $DP(I_U)$ o desvio padrão de I_U (Almeida, 2012).

Outra forma comumente utilizada para avaliar a significância estatística do I_U é por meio do teste de pseudo-significância estatística. Na pseudo-significância, são geradas m permutações dos valores analisados que são associados aleatoriamente aos municípios de tal forma que m valores de I_U são gerados, dando a possibilidade de avaliar o valor original de I_U com a distribuição dos m valores (Cliff; Ord, 1981; Anselin; Rey, 2014), conforme mostrado na Figura 20.

Figura 20 – Teste de pseudo-significância estatística



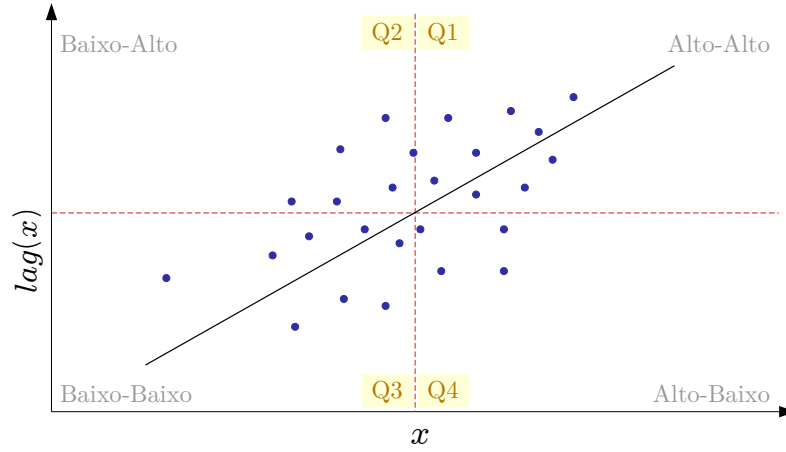
Fonte: Elaborado pela autora (2025).

Portanto, o valor de I_U original é não significativo quando está fora da região crítica de aceitação especificada previamente (Almeida, 2012; Anselin; Rey, 2014) (Figura 20). Usualmente é utilizado $p\text{-valor} = 0,001$ com 999 permutações.

Outrossim, por meio do gráfico de dispersão de Moran é possível visualizar a existência de autocorrelação espacial, além de identificar municípios com valores atípicos (Anselin, 1995; Almeida, 2012). O gráfico de dispersão de Moran, como ilustrado na Figura 21, mostra a relação entre a variável de análise x de cada município normalizada, com o valor médio dos municípios

vizinhos da variável de análise normalizada, ou seja, a média ponderada dos valores vizinhos normalizados ($lag(x)$) (Almeida, 2012).

Figura 21 – Gráfico de dispersão de Moran



Q1: quadrante 1, Q2: quadrante 2, Q3: quadrante 3, Q4: quadrante 4.

Fonte: Elaborado pela autora (2025), com informação de Almeida (2012).

Conforme a Figura 21, o gráfico de dispersão de Moran é dividido em quatro quadrantes: Alto – Alto (AA), Baixo – Alto (BA), Baixo – Baixo (BB), Alto – Baixo (AB) (Anselin, 1995; Almeida, 2012). Os quadrantes um e três estão relacionados com a autocorrelação espacial positiva, indicando a presença de agrupamentos de municípios vizinhos com valores semelhantes (AA e BB). Por outro lado, os quadrantes dois e quatro (BA e AB) estão relacionados com a autocorrelação espacial negativa, indicando a existência de aglomerações espaciais de municípios com alto ou baixo valor na variável de análise, com municípios vizinhos de baixo ou alto valor, respectivamente. No entanto, é importante destacar que nem todos os pontos do gráfico apresentam significância estatística (Anselin, 1995).

Por outro lado, sob outra perspectiva, o índice de Moran local univariado (I_{U_i}), também conhecido como Indicador Local para Autocorrelação Espacial (LISA), permite a análise da variável de interesse na área de estudo de forma individual (por município), ao contrário da análise de autocorrelação global que mostra o grau de associação em toda a área de análise (Anselin, 1995; Almeida, 2012; Bivand; Wong, 2018).

Assim, o I_{U_i} permite identificar padrões de associação espacial estatisticamente significativos, como pontos atípicos ou aglomerações (Anselin, 1995; Bivand; Wong, 2018), sendo calculado a partir da Equação 2.12:

$$I_{U_i} = \frac{(x_i - \bar{x}) \sum_j^n w_{ij}(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2} \quad (2.12)$$

onde a soma dos n valores locais de Moran é igual à autocorrelação global, dada pela Equação 2.13:

$$I_U = \frac{\sum_i^n I_{U_i}}{n} \quad (2.13)$$

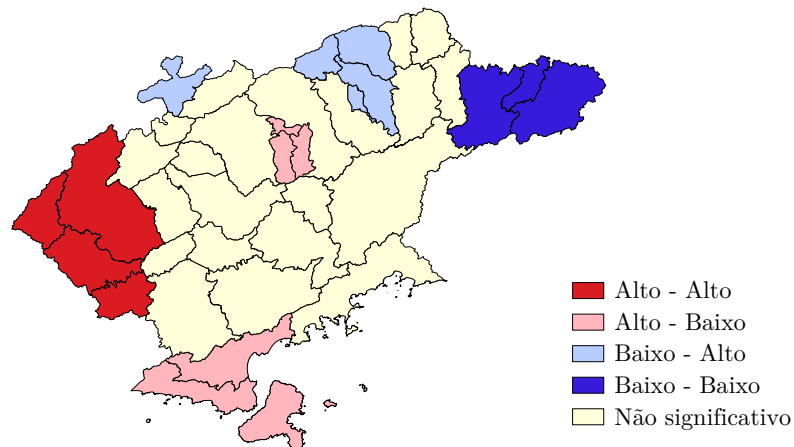
Da mesma forma que na análise global, os valores de I_{U_i} superiores ao valor esperado ($E(I_{U_i})$, Equação (2.14)) indicam a presença aglomerações de municípios com valores semelhantes, e valores inferiores indicam a presença de aglomerações de municípios com risco de transição (Anselin, 1995; Bivand; Wong, 2018). Do mesmo modo, a significância estatística pode ser obtida pela pressuposição de normalidade ou pelo teste de pseudo-significância.

$$E(I_{U_i}) = \frac{-\sum_j^n w_{ij}}{n-1} \quad (2.14)$$

Além disso, as informações dos quadrantes do gráfico de dispersão de Moran e a significância local de I_{U_i} podem ser utilizados na elaboração do mapa temático de Moran, torna-se possível a identificação de aglomerações espaciais (Anselin, 1995; Almeida, 2012), conforme ilustrado na Figura 22.

O mapa temático de Moran permite a identificação de aglomerados espaciais estatisticamente significativos (Figura 22), na divisão das categorias de AA, BB, AB e BA (Almeida, 2012), conforme os quadrantes do gráfico de dispersão de Moran (Figura 21).

Figura 22 – Mapa de Moran



Fonte: Elaborado pela autora (2025), com informação de Anselin (1995) e Almeida (2012).

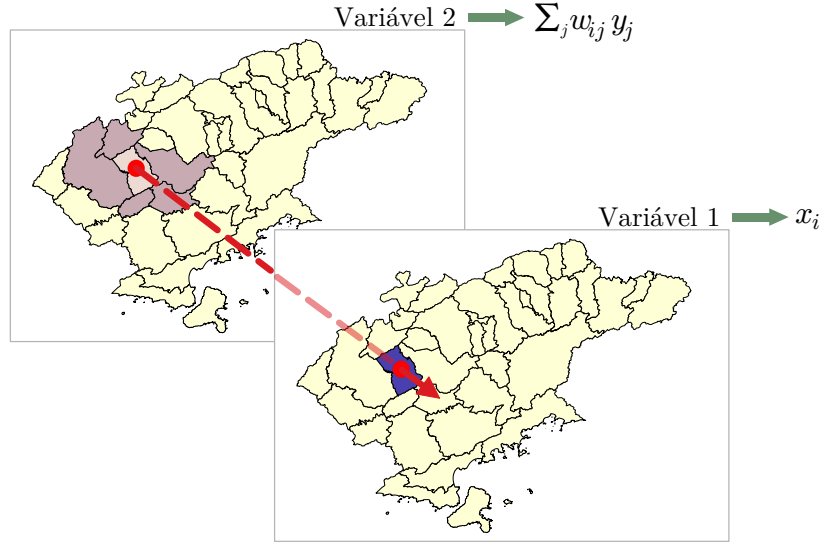
Adicionalmente, é importante destacar que a ausência de significância no índice global I_U não indica ausência de aglomerações locais estatisticamente significativas (Anselin, 1995; Almeida, 2012).

2.4.2 Autocorrelação Espacial Bivariada

De forma similar ao índice univariado, o índice de Moran global bivariado (I_B) permite analisar o comportamento espacial da variável de interesse, com a diferença que no I_B é analisada sua relação com uma segunda variável. Isto é, conforme observado na Figura 23, a análise com o Moram bivariado permite medir o grau de associação espacial da variável de interesse (x_i)

em relação com os municípios vizinhos de uma segunda variável ($\sum_j^n w_{ij}y_j$) (Anselin; Syabri; Smirnov, 2002).

Figura 23 – Autocorrelação espacial bivariada



Fonte: Elaborado pela autora (2025), com informação de Anselin, Syabri & Smirnov (2002).

Deste modo, o I_B pode ser calculado conforme apresentado na Equação (2.15):

$$I_B = \frac{n}{\sum_i^n \sum_j^n w_{ij}} \times \frac{\sum_i^n \sum_j^n w_{ij} (x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \sqrt{\sum_j^n (y_j - \bar{y})^2}} \quad (2.15)$$

sendo que, x_i e y_i são as duas variáveis analisadas espacialmente (Anselin; Syabri; Smirnov, 2002). Além disso, a significância estatística é estimada da mesma forma que no índice univariado (Anselin, 1995; Anselin; Syabri; Smirnov, 2002; Anselin; Syabri; Kho, 2005).

No caso do gráfico de dispersão de Moran, sua elaboração segue a mesma seqüência do diagrama univariado (Figura 21), com a diferença de que a média ponderada dos valores vizinhos é da outra variável em análise normalizada ($lag(y)$). Além disso, o gráfico de dispersão bivariado também é dividido nos quatro quadrantes de AA, BA, BB e AB (Anselin, 1995).

Portanto, os quadrantes um e três (AA e BB), relacionam-se à presença de valor positivo no I_B , indicando existência de aglomerados municipais com valores semelhantes entre as duas variáveis, ou seja, municípios com valores altos na variável de interesse tendem a estar rodeadas por municípios vizinhos com valores altos na outra variável de análise (AA) (Anselin; Syabri; Smirnov, 2002; Almeida, 2012), sendo o mesmo comportamento no caso de valores baixos nas duas variáveis (BB).

Por sua vez, os quadrantes dois e quatro (BA e AB), referem-se à presença de valor negativo no I_B , indicando existência de aglomerados municipais com valores baixos na variável principal, cercados por municípios com elevados valores na outra variável de análise (BA); ou municípios

com altos valores na variável principal são vizinhos de municípios com baixos valores na outras variável (AB) (Anselin; Syabri; Smirnov, 2002; Anselin; Syabri; Kho, 2005).

Já no índice de Moran local bivariado (I_{B_i}), a Equação 2.16 é uma adaptação do índice local univariado (Equação 2.12), na qual é integrada a forma bivariada:

$$I_{B_i} = \frac{(x_i - \bar{x}) \sum_j^n w_{ij} (y_j - \bar{y})}{\frac{\sum_i^n (x_i - \bar{x})^2}{n}} \quad (2.16)$$

Os valores significativos de I_{B_i} e os resultados dos quadrantes do gráfico de dispersão de Moran bivariado são utilizados no mapa temático de Moran, similarmente à Figura 21 apresentada.

2.4.3 Suavização Bayesiana Empírico

Em muitos casos, a variável de análise pode apresentar instabilidade, com alta variabilidade e presença de *outliers*, frequentemente esse comportamento é observado na avaliação da taxa, pelo qual, alguns métodos de suavização podem ser utilizados para corrigi-los (Costa-Nobre *et al.*, 2021; Marinonio *et al.*, 2022).

Um método de suavização amplamente utilizado na literatura é a suavização bayesiana empírica, que se caracteriza por estabilizar e reduzir a alta variabilidade observada nas taxas brutas (Marshall, 1991; Devine; Louis; Halloran, 1994). Assim, a Equação 2.17 mostra como estimar a taxa suavizada com o método bayesiana empírico:

$$\hat{\theta}_i = m_i + C_i(x_i) - m_i \quad (2.17)$$

sendo m_i é a taxa média da área analisada, x_i a taxa bruta do município i , e C_i depende do tamanho da população da área com valores entre 0 e 1 (Marshall, 1991; Costa-Nobre *et al.*, 2021).

2.5 TÉCNICAS DE SÉRIES TEMPORAIS

Geralmente, os termos predição (*prediction*) e previsão (*forecasting*) são frequentemente usados para descrever o mesmo evento, dado que os termos parecem ser quase idênticos, no entanto, existem diferenças significativas entre os dois termos pois podem ser utilizados em diferentes situações (Bares; Bosse, 2022; Awotunde *et al.*, 2023).

A previsão é considerada um ramo da predição, visto que as previsões são baseadas em dados atuais e passados, organizados cronologicamente, para fornecer informações futuras (Bares; Bosse, 2022). Pelo contrário, a predição não está diretamente ligada à componente temporal, sendo um modelo ajustado com base em dados relacionados a um conjunto de características (Awotunde *et al.*, 2023). Portanto, a principal diferença entre previsão e predição é a consideração de dados com dimensão temporal.

Uma série temporal é uma sequência de observações aleatórias listadas cronologicamente, das quais as observações passadas coletadas em intervalos regulares podem ser usadas para gerar previsões futuras. Portanto, as observações da série temporal de y^t podem ser denotados como o vetor 2.18:

$$y^t = \begin{bmatrix} y_t \\ y_{t+1} \\ y_{t+3} \\ \vdots \\ y_T \end{bmatrix} \quad (2.18)$$

na qual y_t é uma observação no instante t , satisfazendo a restrição de $1 \leq t \leq T$, e T é a duração de tempo do período considerado relacionada com o número total de observações (Parmezan; Souza; Batista, 2019).

Em estudos de previsão com séries temporais utilizando métodos tradicionais, as séries temporais devem atender à característica de estacionariedade para que as previsões possam ser generalizadas para outros períodos (Makridakis; Wheelwright; Hyndman, 2008; Gujarati, 2022).

Uma série temporal é estacionária quando é estável ao longo do tempo, ou seja quando qualquer alteração no tempo não causa mudança na forma de sua distribuição (Makridakis; Wheelwright; Hyndman, 2008). Assim, a série temporal estacionária deve atender às propriedades de que a média, variância e autocovariância não mudam com o tempo:

$$E(y_t) = \mu \quad (2.19)$$

$$var(y_t) = E(y_t - \mu)^2 = \sigma^2 \quad (2.20)$$

$$\gamma_k = E[(y_t - \mu)(y_{t+k} - \mu)] \quad (2.21)$$

sendo γ_k la autocovariância no atraso k (ou períodos) entre os valores de y_t e y_{t+k} (Gujarati, 2022).

Na prática, muitos dos dados coletados de diversos processos de estudo não apresentam um padrão temporal estacionário, de forma que uma prática comum é transformar dados não estacionários em dados estacionários por meio de diferenças sucessivas (Parmezan; Souza; Batista, 2019). Geralmente, a primeira diferença Δy_t da série é suficiente para converter a série temporal não estacionária em estacionária, conforme apresentada na Equação 2.22:

$$\Delta y_t = y_t - y_{t-1} \quad (2.22)$$

Deste modo, existem testes estatísticos como o teste de Dickey-Fuller aumentado, em inglês Augmented Dickey–Fuller (ADF), e o teste de Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

que podem ser utilizados para verificar a estacionariedade da série temporal (Kwiatkowski *et al.*, 1992; Fuller, 2009).

No teste de ADF a hipótese nula ($H_0 : |\rho| = 0$) afirma que a série temporal é não estacionária, indicando a existência de raiz unitária (Fuller, 2009). Por outro lado, no teste de KPSS a hipótese nula ($H_0 : |\rho| < 0$) verifica que a série temporal não tem raiz unitária, mostrando que a série é estacionária (Kwiatkowski *et al.*, 1992). Desta forma, caso o p-valor resultante do teste seja inferior ao nível de significância (α) considerado, a hipótese nula de cada teste pode ser rejeitada (Kwiatkowski *et al.*, 1992; Fuller, 2009).

O Quadro 4 apresenta a descrição da análise combinatória dos testes de ADF e de KPSS, conforme descrito por Nusair (2003), para uma decisão mais precisa sobre a estacionariedade da série temporal.

Quadro 4 – Análise combinatória dos testes de estacionariedade da série temporal

KPSS	ADF	
	Aceitar H_0	Rejeitar H_0
Aceitar H_0	Decisão inconclusiva (Informação insuficiente)	Decisão conclusiva (Estacionária)
Rejeitar H_0	Decisão conclusiva (Não estacionário)	Decisão inconclusiva (Integração fracionária)

Fonte: Adaptação de Kwiatkowski *et al.* (1992) e Fuller (2009).

Em relação aos modelos clássicos de previsão, na literatura, as técnicas bastante conhecidas em séries temporais são ARIMA e SARIMA, sendo acrônimos do modelo Auto-regressivo Integrado de Médias Móveis e do modelo Auto-regressivo Integrado de Médias Móveis com Sazonalidade.

O modelo ARIMA é uma classe de modelo linear, conhecido como $ARIMA(p, d, q)$, que inclui os processos estatísticos de autorregressão (AR) de ordem p e o processo de média móveis (MA) de ordem q . Além disso, o modelo ARIMA adiciona a integração (I) da série, que indica o número de diferenças d necessárias para garantir a estacionariedade da série temporal (Makridakis; Wheelwright; Hyndman, 2008).

Assim, descrevendo separadamente o processo AR, a previsão do valor de y_t é calculada utilizando a combinação linear dos p valores passados da variável. O modelo $AR(p)$ é definido conforme a Equação 2.23:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (2.23)$$

sendo ϕ_p os parâmetros do modelo, e ε_t o erro aleatório também conhecido como ruído branco.

No processo MA a previsão do valor y_t é calculada utilizando os q erros de previsão anteriores. O modelo $MA(q)$ é definido conforme a Equação 2.24:

$$y_t = \theta_0 - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - \dots - \theta_q \varepsilon_{t-q} \quad (2.24)$$

sendo θ_q os parâmetros do modelo e ε_{t-q} os erros.

Deste modo, no modelo ARIMA as previsões do valor y_t são calculadas aplicando-se a combinação do modelo AR e do modelo MA, com o pressuposto de estacionaridade (Gujarati, 2022). Portanto, se a série temporal não for estacionária, ela deverá ser diferenciada d vezes para torná-la estacionária. O modelo ARIMA(p, d, q) é definido conforme a Equação 2.25:

$$y_t = \overbrace{\phi_0 + \phi_1 y_{t-1}^d + \phi_2 y_{t-2}^d + \dots + \phi_p y_{t-p}^d}^{AR} + \underbrace{\theta_0 - \theta_1 \varepsilon_{t-1}^d - \theta_2 \varepsilon_{t-2}^d - \dots - \theta_q \varepsilon_{t-q}^d}_{MA} \quad (2.25)$$

sendo y_{t-p}^d a série de tempo diferenciada de ordem d (Makridakis; Wheelwright; Hyndman, 2008; Gujarati, 2022).

Por outro lado, no caso do modelo SARIMA, o modelo integra na previsão a presença de sazonalidade dos dados. Este modelo é conhecido como SARIMA(p, d, q)(P, D, Q) $_s$, sendo s o período de sazonalidade da série, e P, D e Q são, respectivamente, os procedimentos AR(P), I(D) e MA(Q) da parte sazonal (Parmezan; Souza; Batista, 2019).

O modelo SARIMA(p, d, q)(P, D, Q) $_s$ é definido conforme a Equação 2.26:

$$y_t = \overbrace{\overbrace{\phi_0 + \phi_1 y_{t-1}^d + \dots + \phi_p y_{t-p}^d}^{AR} + \overbrace{\theta_0 - \theta_1 \varepsilon_{t-1}^d - \dots - \theta_q \varepsilon_{t-q}^d}^{MA}}^{\text{Parte não sazonal}} + \underbrace{\overbrace{\Phi_0 + \Phi_1 y_{t-1}^D + \dots + \Phi_p y_{t-p}^D}^{AR} + \overbrace{\Theta_0 - \Theta_1 \varepsilon_{t-1}^D - \dots - \Theta_q \varepsilon_{t-q}^D}^{MA}}_{\text{Parte sazonal}} \quad (2.26)$$

sendo ϕ_p e θ_q os parâmetros não sazonais do modelo AR e MA, Φ_p e Θ_q os parâmetros sazonais do modelo AR e MA, s o período de sazonalidade, d a diferenciação não sazonal e D a diferenciação da parte sazonal (Petropoulos *et al.*, 2022).

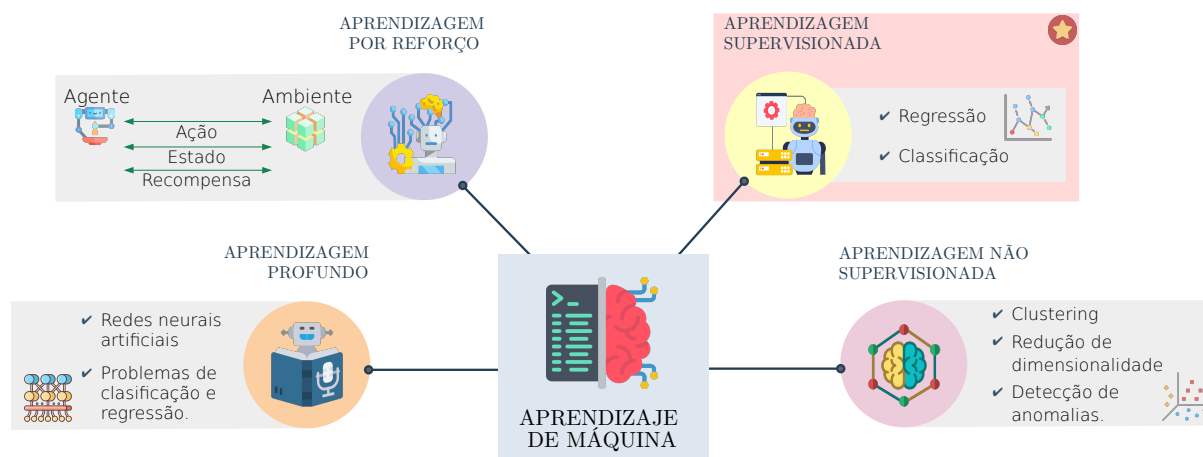
2.6 TÉCNICAS DE APRENDIZADO DE MÁQUINA

ML é um subcampo da IA, que pode ser definido como a capacidade de identificar padrões complexos nos dados, a fim de realizar previsões futuras ou auxiliar na compreensão do processo em estudo (Alpaydin, 2021; Pugliese; Regondi; Marini, 2021). De maneira geral, as técnicas baseadas em IA começaram a se difundir com maior intensidade a partir de 1970 (Alpaydin, 2021), e, especificamente, o termo de ML foi cunhado por Arthur Lee Samuel (1959), com seu algoritmo para o jogo de damas. No entanto, foi a partir de 1990 que o ML passou a ser estudado como um campo de pesquisa separado.

Dessa forma, o ML fundamenta-se na estatística para o desenvolvimento de modelos matemáticos que possibilitam a realização de inferências a partir de uma amostra de dados (Jordan;

Mitchell, 2015; Alpaydin, 2021). Além disso, no desenvolvimento de modelos, é possível testar diferentes técnicas de ML, as quais podem ser classificadas de acordo com seu foco principal em quatro grupos ou métodos principais, conforme ilustrado na Figura 24.

Figura 24 – Classificação de técnicas de aprendizado de máquina



Fonte: Adaptação de Alpaydin (2021) e Pugliese, Regondi & Marini (2021).

Cada um dos métodos apresentados na Figura 24 é descrito a seguir:

- O aprendizado supervisionado, em inglês *Supervised Learning*, está relacionado com o conjunto de técnicas utilizadas para estimar valores contínuos ou valores discretos, como é o caso das técnicas de regressão e classificação, respectivamente (Jordan; Mitchell, 2015; Pugliese; Regondi; Marini, 2021). O principal objetivo dessas técnicas é identificar relações em conjuntos de dados rotulados (entrada-saída), para estimar valores desconhecidos com base nessas relações (Alpaydin, 2021).
- O aprendizado não supervisionado, em inglês *Unsupervised Learning*, tem como objetivo principal identificar padrões e estruturas existentes em conjuntos de dados sem orientação explícita, ou seja, sem especificação a priori de rótulos dos dados (Faceli *et al.*, 2021; Alpaydin, 2021). A clusterização (*clustering*) e redução de dimensionalidade são abordagens bem representativas usadas na aprendizagem não supervisionada (Pugliese; Regondi; Marini, 2021).
- O aprendizado por reforço, em inglês *Reinforcement Learning*, tenta aprender padrões mediante um processo de *feedback* sequencial, onde os principais componentes de agente e ambiente se destacam pela capacidade de perceber seu ambiente, tomar decisões e aprender por meio do processo de tentativa e erro (Pugliese; Regondi; Marini, 2021).
- No aprendizado profundo, em inglês *Deep Learning*, as técnicas utilizadas podem ser supervisionadas ou não supervisionadas, sendo amplamente aplicadas a dados de alta dimensão, como processamento de texto, voz e imagens (Doupe; Faghmous; Basu, 2019).

Os algoritmos de aprendizado profunda baseiam-se principalmente em redes neurais artificiais, caracterizadas por possuir uma arquitetura multicamadas destinada a simular o comportamento do cérebro humano (Alpaydin, 2021).

No entanto, considerando que o principal objetivo deste estudo é desenvolver modelos de predição utilizando aprendizado supervisionado, a seguir descreve-se o método, incluindo suas abordagens de regressão e classificação.

Deste modo, a principal finalidade das técnicas de ML supervisionado é aprender uma função desconhecida $y = f(\mathbf{x}) + \epsilon$, para predizer um valor aproximado $\hat{y} = \hat{f}(\mathbf{x})$ (Faceli *et al.*, 2021; James *et al.*, 2023). Essa função é construída com base nas relações entre as variáveis de entrada e saída do conjunto de dados. Em outras palavras, as relações entre os valores de $\mathbf{x}$ e $\mathbf{y}$ na matriz $\mathbf{X}$ (Equação 2.1) podem ser representadas como $\mathbf{X} = (\mathbf{x}, \mathbf{y})$, conforme mostrado na matriz e vetor da Equação 2.27:

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & \cdots & x_{n,d} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad (2.27)$$

sendo $\mathbf{x}$ a matriz de d variáveis de entrada, $x_{i,j}$ é o atributo j da variável i , e cada vetor $\mathbf{x}_i$ possui um rótulo y_i (James *et al.*, 2023). Assim, os dados de treinamento $\mathbf{Tr}$ para modelos de ML são compostos por uma sequência finita de pares de observações $\mathbf{x}_i \times y_i$, conforme mostrado na Equação 2.28:

$$\mathbf{Tr} = \{(\mathbf{x}_1, y_1)_1, (\mathbf{x}_2, y_2)_2, (\mathbf{x}_3, y_3)_3, \dots, (\mathbf{x}_i, y_i)_n\} \quad (2.28)$$

Em um cenário de dados com observações de uma série temporal (Equação 2.18), o conjunto de dados de treinamento $\mathbf{Tr}$ é representado por uma subsequência de observações $\mathbf{x}_i \times y_i$, conforme apresentado na Equação 2.29:

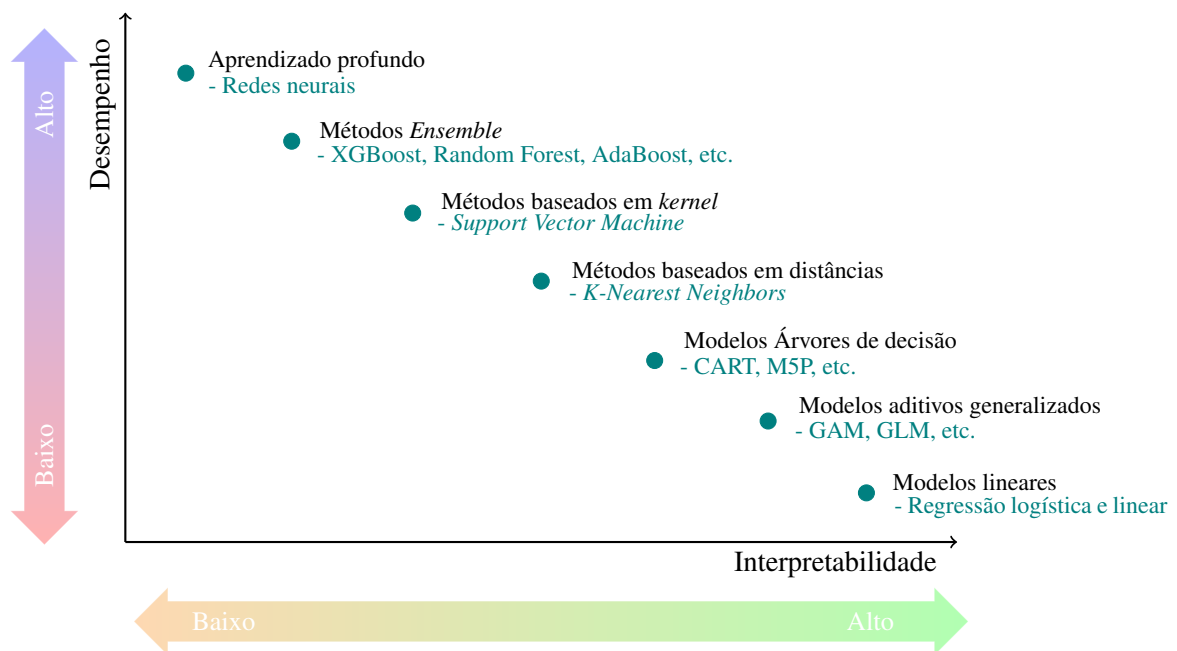
$$\mathbf{x} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,d} \\ y_{2,2} & y_{2,3} & \cdots & y_{2,d+1} \\ y_{3,3} & y_{3,4} & \cdots & y_{3,d+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,d+(n-d)} & y_{n,d+(n-2)} & \cdots & y_{n,d+(n-1)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_{1,d+1} \\ y_{2,d+2} \\ y_{3,d+3} \\ \vdots \\ y_{n,d+n} \end{bmatrix} \quad (2.29)$$

sendo $\mathbf{x}$ a matriz de variáveis de entrada do padrão temporal de dimensão d , e $\mathbf{y}$ é o vetor de rótulos. Neste caso, d é o tamanho da janela deslizante, ou seja, o número de valores anteriores

considerados para prever o valor subsequente de y_{d+1} .

Além disso, a função desconhecida $f(\cdot)$ pode ser de natureza contínua ou discreta. Os algoritmos utilizados para conjuntos de dados em que o rótulo (y) possui valores contínuos são conhecidos como modelos de regressão, em que o estimador treinado $f(\mathbf{x}) \in \mathbb{R}$ pode assumir um conjunto finito de valores (Alpaydin, 2021; James *et al.*, 2023). Já os algoritmos usados quando o rótulo (y) possui uma natureza discreta são conhecidos como modelos de classificação, nos quais o estimador $f(\mathbf{x}) \in K$ pode assumir um conjunto de valores discretos, também conhecidos como classes (Alpaydin, 2021; James *et al.*, 2023). Além disso, os métodos de classificação podem ser categorizados em classificação binária $K = \{0, 1\}$ ou classificação multiclasse $K = \{c_1, c_2, \dots, c_k\}$, com $k > 2$.

Figura 25 – Equilíbrio entre interpretabilidade e desempenho de alguns métodos de aprendizado de máquina



Fonte: Adaptação de James *et al.* (2023).

Outro aspecto importante dos modelos supervisionados é que podem ser categorizados como modelos de caixa branca, quando têm uma estrutura mais transparente, ou como modelos de caixa preta, quando possuem uma estrutura mais complexa e oferecem maior precisão (Abdullah; Zahid; Ali, 2021). A Figura 25 apresenta o equilíbrio entre interpretabilidade e desempenho de alguns métodos de ML.

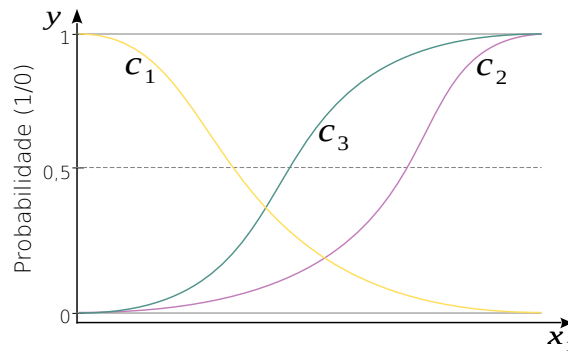
Como pode ser observado na Figura 25, os métodos mais complexos geram modelos com melhor desempenho, sendo vantajosos quando o interesse principal é a precisão. No entanto, os modelos mais simples, como aqueles gerados pelas técnicas de regressão linear, logística, árvores de decisão, etc., são mais interpretáveis, embora apresentem uma menor pontuação de precisão (James *et al.*, 2023).

Na área da saúde, observa-se uma tendência crescente pelo uso de modelos interpretáveis, em vez daqueles que priorizam a alta precisão (Doupe; Faghmous; Basu, 2019; Abdullah; Zahid; Ali, 2021). No contexto desta pesquisa, diversas técnicas de ML, as quais podem ser agrupadas em diferentes categorias: modelos lineares (como a RLog), métodos baseados em árvores de decisão (como DT, RF, GB, XGB, LGBM, Bagging e AdaBoost), SVM e redes neurais (com MLP). As características e aplicações dessas técnicas são detalhadas a seguir.

2.6.1 Regressão Logística

A RLog é uma técnica estatística clássica e amplamente utilizada entre os métodos de aprendizado supervisionado para tarefas de classificação. Essa técnica ajusta uma curva em forma de ‘S’, na qual a estimativa gerada representa a probabilidade ($p(x_i) = p(y_i = n|x_i)$) de que a observação de entrada pertença a uma das k classes (Bishop; Nasrabadi, 2006). A RLog permite a modelagem direta da probabilidade de pertencimento a uma classe específica, conforme ilustrado na Figura 26.

Figura 26 – Estrutura de regressão logística



Fonte: Elaborado pelo autor (2025).

A Figura 26 ilustra a curva sigmoidal (‘S’), característica da função logística, que é empregada para calcular a probabilidade de classes. A função preditora logística $f_{RLog}(x)$, quando aplicada a problemas de classificação binária $\{0, 1\}$, é expressa pela Equação 2.30:

$$f_{RLog}(x) = p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}} \quad (2.30)$$

sendo β_0 e β_1 os coeficientes do modelo, e x_i representa as variáveis de entrada (Marsland, 2015; Alpaydin, 2021). Esses coeficientes são estimados por meio do método de máxima verossimilhança, garantindo que os parâmetros encontrados maximizem a probabilidade dos dados observados.

Para problemas de classificação multiclasse ($k > 2$), a RLog é estendida para a versão

multinomial, cuja função preditora é definida pela Equação 2.31:

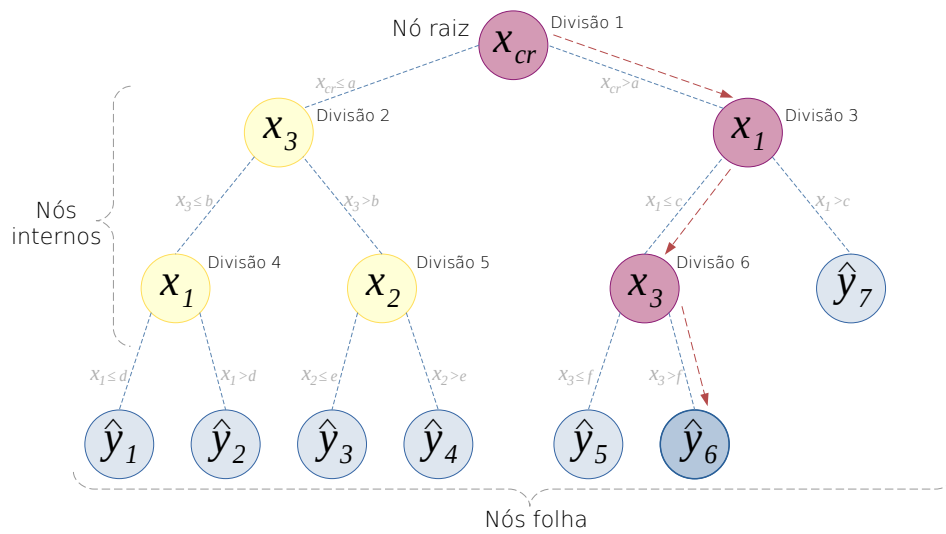
$$f_{RLog}(x) = p(x_i) = \frac{e^{\beta_{k0} + \beta_{ki}x_i}}{\sum_{l=1}^K e^{\beta_{l0} + \beta_{li}x_i}}, \quad l \in 1, \dots, K \quad (2.31)$$

A $f_{RLog}(x)$ da Equação 2.31 permite modelar diretamente a probabilidade de cada classe, assegurando que a soma das probabilidades para todas as classes seja igual a 1 (Alpaydin, 2021; James *et al.*, 2023).

2.6.2 Métodos Baseados em Árvores de Decisão

As DT são modelos amplamente utilizados em tarefas de classificação e regressão, cuja estratégia de solução consiste em decompor um problema complexo em subproblemas mais simples, a fim de facilitar o processo de tomada de decisão (Faceli *et al.*, 2021; Izicki; Santos, 2021).

Figura 27 – Estrutura de uma Árvore de Decisão



Fonte: Elaborado pelo autor (2025).

Formalmente, uma DT é um método não paramétrico, cuja estrutura hierárquica é construída a partir do conjunto de treinamento (Tr) rotulado. Essa estrutura é composta por nós internos ou nós de divisão, que representam regras de particionamento sobre as variáveis de entrada, e por nós folha, que contêm uma função responsável por atribuir as previsões finais, conforme ilustrado na Figura 27.

A divisão binária descendente do conjunto de dados em subconjuntos mais homogêneos ($\mathcal{X}^a$, $\mathcal{X}^b$) é realizada até que se obtenha uma redução substancial do erro em cada folha da árvore (Alpaydin, 2021). Além disso, os nós de divisão devem garantir que a união dos elementos de ambos os subconjuntos seja igual ao total de elementos do nó de divisão ($\mathcal{X}^a \cup \mathcal{X}^b = \mathcal{X}$), bem

como que os subconjuntos não possuam instâncias em comum ($\mathcal{X}^a \cap \mathcal{X}^b = \emptyset$). No caso dos nós folha, eles armazenam as previsões finais do modelo e podem ser interpretados como a saída da função preditora. O critério para avaliar seu desempenho depende da abordagem adotada, seja para modelos de classificação ou regressão.

No caso das árvores de regressão, o nó de divisão é selecionado considerando aquele atributo que melhor particiona o espaço de entrada e minimiza o erro na predição, resultando em uma segmentação mais eficiente dos dados. O critério mais utilizado para a divisão em árvores de regressão é a Soma Residual de Quadrados, em inglês *Residual Sum of Squares* (RSS), devido à sua capacidade de reduzir a variabilidade dos dados (Izbicki; Santos, 2021). A função de custo RSS, empregada para segmentar os dados avaliando a qualidade do nó de divisão, é definida pela Equação 2.32, a fim de minimizar a soma dos resíduos:

$$RSS(\mathcal{X}_{pai}) = RSS(\mathcal{X}^a) + RSS(\mathcal{X}^b)$$

$$RSS(\mathcal{X}_{pai}) = \left[\sum_{x_i \in \mathcal{X}^a}^n (y_i - \bar{y}^a)^2 \right] + \left[\sum_{x_i \in \mathcal{X}^b}^n (y_i - \bar{y}^b)^2 \right] \quad (2.32)$$

onde $\bar{y}_t^a$ e $\bar{y}_t^b$ representam as médias dos valores da variável de saída em cada subconjunto, sendo o ponto de corte ótimo aquele que minimiza o RSS. Além disso, as árvores de regressão assumem um modelo da forma apresentada na Equação 2.33:

$$f_{DT}(x) = \sum_{b=1}^B c_b \times 1_{\{X \in R_b\}} \quad (2.33)$$

onde o espaço de entrada é segmentado em B regiões terminais $R_1, R_2, \dots, R_b$. A predição de cada região R_b é dada pelo coeficiente c_b , que geralmente corresponde à média dos valores da variável de saída na região (James *et al.*, 2023). A função $1_{\{X \in R_b\}}$ é uma função característica que assume 1 ou 0, conforme apresentado na Equação 2.34:

$$c_b = \frac{1}{|R_b|} \sum_{x \in R_b} y_i \quad \wedge \quad 1_{(X \in R_b)} = \begin{cases} 1 & x \in R_b, \\ 0 & \text{Caso contrário.} \end{cases} \quad (2.34)$$

No caso das árvores de classificação, a seleção do nó de divisão é baseada em critérios que avaliam a pureza ou homogeneidade das classes nos subconjuntos de dados (Izbicki; Santos, 2021; Suzuki; Suzuki, 2021). Dois dos métodos mais comuns para medir a pureza são o índice de Gini (G) e a entropia (D), cujos objetivos são minimizar seus valores ao particionar os dados.

O índice de Gini mede o grau de impureza do subconjunto de dados, indicando a probabilidade de que um elemento escolhido aleatoriamente seja classificado incorretamente, considerando a distribuição das classes no subconjunto. O G é definido pela Equação 2.35, e varia entre 0 e 1, sendo que quanto menor o valor do índice ($G = 0$), mais homogêneo é o grupo; e quanto maior

o valor ($G = 1$), mais impurezas o nó apresenta.

$$G = \sum_{k=1}^K p(\mathcal{X}_k) (1 - p(\mathcal{X}_k)) \quad (2.35)$$

A probabilidade $p(\mathcal{X}_k)$ representa a chance de que um dado selecionado aleatoriamente no nó pertença à classe k (Suzuki; Suzuki, 2021).

A entropia mede o grau de impureza de um subconjunto de dados e é definida pela Equação 2.36:

$$D = - \sum_{k=1}^K p(\mathcal{X}_k) \log p(\mathcal{X}_k) \quad (2.36)$$

Assim como no cálculo do índice de Gini (G), $p(\mathcal{X}_k)$ representa a que um dado escolhido aleatoriamente no nó pertença à classe k (Faceli *et al.*, 2021). Dessa forma, tanto o índice de Gini quanto a entropia são critérios fundamentais na construção de árvores de classificação, permitindo a escolha dos nós de divisão que melhor separam as classes e maximizam a eficiência do modelo preditivo.

A função preditora de uma árvore de classificação segue uma lógica semelhante àquela da árvore de regressão apresentada na Equação 2.33, com a diferença de que c_b é a classe mais frequente na região R_b , conforme apresentado na Equação 2.37:

$$c_b = \text{moda}(y_1, y_2, \dots, y_n) \quad (2.37)$$

Por outro lado, embora a DT seja uma técnica poderosa, apresenta algumas limitações importante, como alta variância e tendência ao sobreajuste (Suzuki; Suzuki, 2021; Faceli *et al.*, 2021). Por essa razão, e com o objetivo de mitigar esses problemas, surgiram outras técnicas de ML baseadas em DT, como os métodos de *ensemble*, que permitem desenvolver modelos mais robustos.

Os métodos de *ensemble* foram desenvolvidos com o objetivo de reduzir a variabilidade e melhorar a precisão dos modelos, resultando no desenvolvimento de uma ampla variedade de abordagens (Galicia *et al.*, 2019). Embora esses métodos sejam baseados em DT, o processo de treinamento dos modelos varia conforme a técnica adotada (Faceli *et al.*, 2021).

No caso do RF e do Bagging, as DT internas são treinadas em paralelo de forma independente. Para melhorar seu poder preditivo, o processo de treinamento desses modelos envolve a construção de D árvores internas, baseadas em subconjuntos de dados gerados por *bootstrap*¹, para realizar a predição das estimativas $\hat{f}_1(x), \hat{f}_2(x), \dots, \hat{f}_d(x)$ (Izbicki; Santos, 2021).

Especificamente, o Bagging utiliza todas as variáveis do conjunto de dados para construir cada

¹ Bootstrap (*BS*): Método de amostragem estatística que gera múltiplas amostras a partir de um conjunto de dados original utilizando amostragem com reposição. Isso significa que um mesmo dado pode ser selecionado várias vezes na mesma amostra. Essa abordagem permite estimar a variabilidade de uma estatística (como a média ou a variância) ao gerar várias amostras representativas do conjunto de dados original (Faceli *et al.*, 2021).

um dos modelos base $f_d^{BS}(x)$, que são treinados com subconjuntos amostrados aleatoriamente por meio do método *bootstrap*. A função preditora do modelo Bagging para problemas de regressão é dada pela Equação 2.38, enquanto para problemas de classificação, a predição é realizada conforme a Equação 2.39:

$$f_{Bagging}(x) = \frac{1}{D} \sum_{d=1}^D f_d^{BS}(x) \quad (2.38)$$

$$f_{Bagging}(x) = \text{moda} (f_1^{BS}(x), f_2^{BS}(x), \dots, f_d^{BS}(x)) \quad (2.39)$$

O $f_{bagging}(x)$ representa, no caso de regressão, a média das previsões dos D modelos base, enquanto, para problemas de classificação, é o valor mais frequente entre as predições.

Em contrapartida, o RF representa uma extensão do Bagging, cujas funções preditoras para modelos de regressão e classificação são semelhantes às do Bagging, conforme as Equações 2.40 e 2.41:

$$f_{RF}(x) = \frac{1}{D} \sum_{d=1}^D f_d^{*BS}(x) \quad (2.40)$$

$$f_{RF}(x) = \text{moda} (f_1^{*BS}(x), f_2^{*BS}(x), \dots, f_d^{*BS}(x)) \quad (2.41)$$

A principal diferença de $f_{RF}(x)$ em relação ao Bagging reside no processo de construção das D árvores internas. Além da utilização do método *bootstrap*, o RF introduz um critério de aleatoriedade na seleção das variáveis em cada nó de cada árvore, representadas por $f_d^{*BS}(x)$ (Suzuki; Suzuki, 2021; James *et al.*, 2023). Essa abordagem reduz a correlação entre as árvores, tornando o modelo mais robusto e mitigando o risco de sobreajuste (*overfitting*).

Por outro lado, no GB e em suas variações, como XGB, LGBM e AdaBoost, as árvores são ajustadas de forma iterativa em um processo sequencial dependente, no qual cada nova iteração busca minimizar e corrigir os erros residuais do modelo anterior (Galicia *et al.*, 2019; Faceli *et al.*, 2021).

No treinamento do GB, a função preditora $f_{GB}(x)$, representada pela Equação 2.42, é construída como uma soma ponderada das árvores internas, otimizando uma função de perda por meio do método de descida do gradiente (Bishop; Nasrabadi, 2006; James *et al.*, 2023). Esse processo ajusta iterativamente os modelos para minimizar os resíduos do modelo anterior, minimizando a função de perda em cada iteração.

$$f_{GB}(x) = \sum_{d=1}^D \gamma \times f_d(x) \quad (2.42)$$

sendo D o número total de árvores simples treinadas ao longo do processo, $f_d(x)$ representa a função base na iteração da árvore d , que é treinada para minimizar o erro residual do modelo, e γ é a taxa de aprendizado que regula o impacto de cada nova árvore no modelo final, garantindo um

ajuste controlado (James *et al.*, 2023). Para problemas de regressão, a função de perda otimizada pelo GB é geralmente o RSS (ou o Mean Absolute Error (MSE), que é uma versão normalizada do RSS), enquanto para problemas de classificação, utiliza-se a log-verossimilhança negativa (*Log Loss*) (Bishop; Nasrabadi, 2006; Marsland, 2015).

Do mesmo modo, o XGB é uma extensão otimizada do GB, que usa regularização e técnicas avançadas de otimização para melhorar sua eficiência computacional. A função preditora de $f_{XGB}(x)$ é semelhante à de $f_{GB}(x)$ da Equação 2.42, el cual es construido como una suma de arboles simple pero con un enfoque más eficiente y robusto. A função preditora $f_{XGB}(x)$ é apresentada na Equação 2.43:

$$f_{XGB}(x) = \sum_{d=1}^D f_d(x) \quad (2.43)$$

Do mesmo modo, o XGB é uma extensão otimizada do GB, que incorpora regularização e técnicas avançadas de otimização para melhorar sua eficiência computacional. A função preditora $f_{XGB}(x)$ é semelhante à $f_{GB}(x)$ da Equação 2.42, sendo construída como uma soma de árvores simples, mas com um enfoque mais eficiente e robusto, graças à inclusão de termos de regularização e à otimização de segunda ordem (Chen; Guestrin, 2016).

Por outro lado, o LGBM é outra variante do GB que adota estratégias avançadas para aprimorar a eficiência computacional, utilizando a mesma estrutura preditiva geral do XGB, na qual a predição final é obtida por meio da soma das árvores individuais. Dessa forma, $f_{LGBM}(x)$ segue a mesma formulação básica de $f_{XGB}(x)$, mas com diferenças na construção das árvores e no processo de otimização.

Assim, para aumentar a eficiência e reduzir o custo computacional, o LGBM incorpora técnicas como o crescimento de árvores baseado em folhas (*leaf-wise*), que seleciona para divisão o nó que proporciona a maior redução na função de perda, e a técnica de agrupamento de características (*feature bundling*), que reduz a dimensionalidade efetiva ao combinar atributos altamente correlacionados (Ke *et al.*, 2017). Essas abordagens minimizam redundâncias e aceleram significativamente o treinamento do modelo, tornando-o mais escalável para grandes volumes de dados.

Em relação ao AdaBoost, trata-se de um método de *boosting* que, diferentemente do XGB e do LGBM, ajusta iterativamente os pesos das instâncias do conjunto de dados, atribuindo maior importância às observações classificadas incorretamente a cada iteração (Bishop; Nasrabadi, 2006; Alpaydin, 2021). A função preditora $f_{AdaBoost}(x)$ é uma combinação ponderada de modelos fracos, sendo definida de maneira distinta para problemas de regressão e classificação.

Para modelos de regressão, a predição é obtida pela média ponderada das estimativas individuais, conforme a Equação 2.44:

$$f_{AdaBoost}(x) = \sum_{d=1}^D \alpha_d f_d(x) \quad (2.44)$$

onde α_d representa o peso atribuído a cada modelo base $f_d(x)$, proporcional à sua precisão na iteração correspondente (Marsland, 2015).

Já para modelos de classificação, a predição é feita com base na soma ponderada dos classificadores fracos, conforme a Equação 2.45:

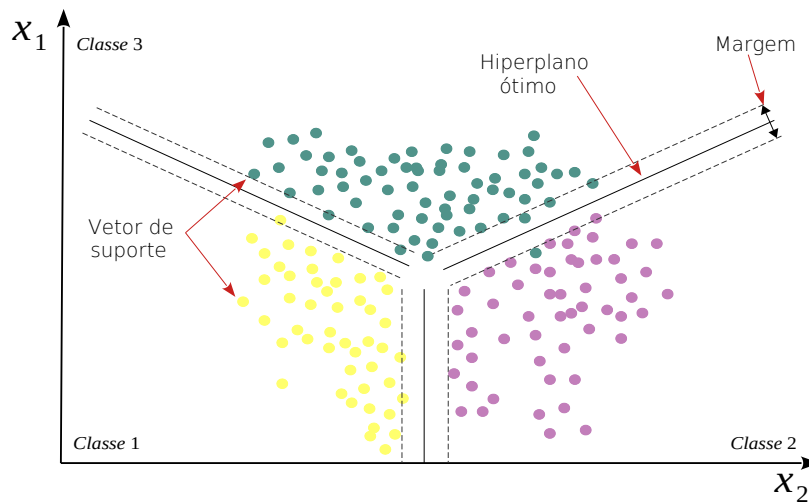
$$f_{AdaBoost}(x) = \text{sign} \left(\sum_{d=1}^D \alpha_d f_d(x) \right) \quad (2.45)$$

sendo que a função $\text{sign}(\cdot)$ define a classe predita com base no sinal da soma ponderada das previsões individuais (Bishop; Nasrabadi, 2006; Marsland, 2015).

2.6.3 Máquinas de Vetores de Suporte

O SVM mapeia os dados de entrada em um espaço de características de maior dimensão, conforme ilustrado na Figura 28, buscando encontrar um hiperplano ótimo que maximize a margem entre as classes na classificação ou minimize o erro de predição na regressão (Parmezan; Souza; Batista, 2019; James *et al.*, 2023).

Figura 28 – Estrutura de máquina de vetores de suporte



Fonte: Elaborado pelo autor (2025).

A técnica do SVM é amplamente utilizada para problemas de classificação binária e multi-classe, bem como para problemas de regressão, devido à sua capacidade de modelar relações complexas entre variáveis.

No contexto da classificação binária, o SVM busca encontrar um hiperplano que maximize a margem entre as amostras das diferentes classes. Assim, a formulação do problema de otimização é expressa pela Equação 2.46:

$$\min_{f \in H} \frac{1}{2} \|f\|^2 + C \sum_{i=1}^l \xi_i \quad (2.46)$$

sujeito a:

$$y_i f(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, l$$

sendo $f(x)$ a função de decisão do classificador, y_i são os rótulos das classes $\{0, 1\}$, $\|f\|^2$ corresponde ao termo de regularização que controla a complexidade do modelo, C é o parâmetro de regularização que balanceia a maximização da margem e a minimização do erro, e ξ_i são variáveis de folga que permitem a classificação de dados não linearmente separáveis (Alpaydin, 2021).

A formulação dual do problema de otimização é dada pela Equação 2.47:

$$\max_{\alpha \in R^l} \sum_{i=1}^l \alpha_i - \frac{1}{2} \alpha^T K \alpha \quad (2.47)$$

sujeito a:

$$\sum_{i=1}^l y_i \alpha_i = 0 \quad \wedge \quad 0 \leq \alpha_i \leq C \quad \wedge \quad i = 1, 2, \dots, l$$

sendo α_i os coeficientes de *Lagrange*, e K corresponde à matriz de *kernel*, definida como $K_{ij} = k(x_i, x_j)$.

A função *kernel* é responsável por mapear os dados para um espaço de características de maior dimensão. Os tipos de *kernel* mais comuns são apresentados nas Equações 2.48 - 2.51:

- Kernel Linear:

$$k(x_i, x_j) = x_i x_j \quad (2.48)$$

- Kernel Polinomial:

$$k(x_i, x_j) = (c_1 x_i x_j + c_2)^{c_3} \quad (2.49)$$

- Kernel de Base Radial (RBF):

$$k(x_i, x_j) = e^{-\beta \|x_i - x_j\|^2} \quad (2.50)$$

- Kernel Sigmoidal:

$$k(x_i, x_j) = \tanh(c_1 x_i x_j + c_2) \quad (2.51)$$

A função de decisão final é dada pela função preditora $f_{SVM}(x)$, conforme apresentado na Equação 2.52:

$$f_{SVM}(x) = \sum_{i=1}^l y_i \alpha_i k(x, x_i) + b \quad (2.52)$$

onde $\alpha_i > 0$, e b representa o termo de bias (intercepto) (Marsland, 2015; James *et al.*, 2023).

Para problemas de classificação multiclasse, o SVM pode ser estendido usando abordagens como One-vs-All (OvA) ou All-vs-All (AvA). No OvA, um classificador binário é treinado

para cada classe, tratando-a como a classe positiva e as demais como negativas. Já no AvA, um classificador binário é treinado para cada par de classes. Em ambas as abordagens, a decisão final é tomada com base na função de decisão de cada classificador binário.

Além disso, independentemente da abordagem utilizada, OvA ou AvA, a função preditora $f_{SVM}(x)$ para modelos multiclasse pode ser expressa conforme a Equação 2.53:

$$f_{SVM}(x) = \arg \max_k \left(\sum_{i=1}^l y_i^{(k)} \alpha_i^{(k)} k(x, x_i) + b^{(k)} \right) \quad (2.53)$$

sendo $y_i^{(k)}$ os rótulos ajustados para a classe k , $\alpha_i^{(k)}$ são os coeficientes de Lagrange associados a cada classificador, e $b^{(k)}$ representa o termo de viés (*bias*).

No contexto de regressão, o SVM busca encontrar uma função que minimize o erro de predição, garantindo que os valores estimados permaneçam dentro de uma margem de tolerância ϵ (Bishop; Nasrabadi, 2006; Marsland, 2015; James *et al.*, 2023).

A formulação do problema de otimização para o SVM para regressão é dada pela Equação 2.54:

$$\min_{f \in H} \frac{1}{2} \|f\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (2.54)$$

sujeito a:

$$y_i - f(x_i) \leq \epsilon + \xi_i, \quad f(x_i) - y_i \leq \epsilon + \xi_i^*, \quad \xi_i, \xi_i^* \geq 0, \quad i = 1, 2, \dots, l$$

sendo $f(x_i)$ a função de regressão, ϵ representa a margem de tolerância para erros de predição, e $\{\xi_i, \xi_i^*\}$ são variáveis de folga que possibilitam a flexibilidade da margem ϵ (James *et al.*, 2023).

A formulação dual do problema de otimização do modelo de regressão é dada pela Equação 2.55:

$$\max_{\alpha, \alpha^* \in R^l} \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) - \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) - \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) k(x_i, x_j) \quad (2.55)$$

sujeito a:

$$\sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, 2, \dots, l$$

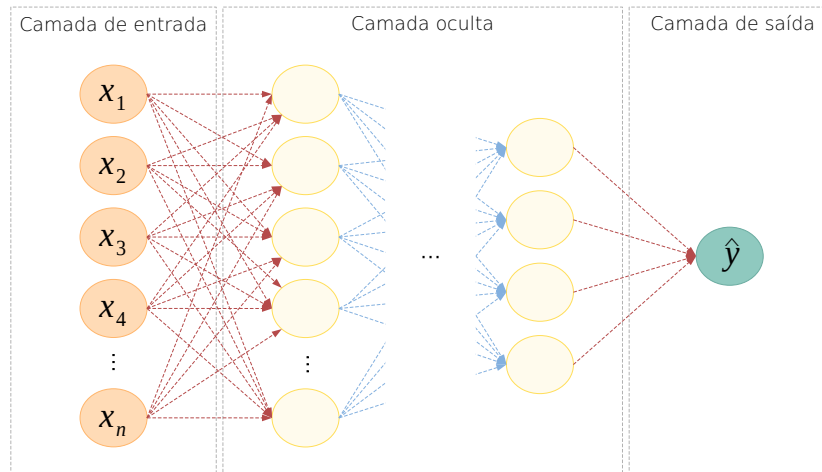
A função de regressão final é dada pela função preditora $f_{SVM}(x)$, conforme apresentado na Equação 2.56:

$$f_{SVM}(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) k(x, x_i) + b \quad (2.56)$$

2.6.4 Rede Neural Perceptron Multicamada

O MLP, também conhecido como rede neural perceptron multicamada, é inspirado no funcionamento do cérebro humano e modela a relação entre um conjunto de variáveis de entrada e a variável de saída (objetivo). Conforme ilustrado na Figura 29, a estrutura básica do MLP é composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, sendo seus neurônios internos conhecidos como *perceptrons* (Alpaydin, 2021).

Figura 29 – Estrutura de rede neural perceptron multicamada



Fonte: Elaborado pelo autor (2025).

A camada de entrada contém os neurônios responsáveis por receber os valores a serem processados. A camada oculta é responsável pelo processamento intermediário, onde ocorre a maior parte da transformação dos dados. Por fim, a camada de saída realiza a predição ou classificação com base nas informações extraídas nas camadas ocultas (Bishop; Nasrabadi, 2006; Marsland, 2015).

As camadas interconectadas do MLP utilizam a técnica de retropropagação (em inglês, *back-propagation*) para ajustar os pesos das conexões e melhorar a predição. Durante o treinamento, o gradiente do erro é calculado em relação aos pesos da rede, e esses pesos são atualizados para minimizar a função de erro. Além disso, cada neurônio do modelo utiliza uma função de ativação não linear para introduzir não linearidades ao modelo.

Para problemas de regressão, a função preditora $f_{MLP}(x)$ é definida pela Equação 2.57:

$$f_{MLP}(x) = W^{(L)} \phi^{(L-1)} \left(W^{(L-1)} \dots \phi^{(1)} \left(W^{(1)} x + b^{(1)} \right) + b^{(L-1)} \right) + b^{(L)} \quad (2.57)$$

onde $W^{(l)}$ é a matriz de pesos, $b^{(l)}$ representa o vetor de vies, e $\phi^{(l)}$ é a função de ativação da camada l , além disso, L indica o número total de camadas.

No processo de treinamento, o modelo minimiza uma função de perda, geralmente o MSE,

definido como a Equação 2.58:

$$\mathcal{L}(y, f_{MLP}(x)) = \frac{1}{n} \sum_{i=1}^n (y_i - f_{MLP}(x_i))^2 \quad (2.58)$$

sendo y_i os valores reais e $f_{MLP}(x_i)$ as predições do modelo.

Por outro lado, para problemas de classificação, a função preditora $f_{MLP}(x)$ é dada na Equação 2.59:

$$f_{MLP}(x) = \text{softmax} \left(W^{(L)} \phi^{(L-1)} \left(W^{(L-1)} \dots \phi^{(1)} \left(W^{(1)} x + b^{(1)} \right) + b^{(L-1)} \right) + b^{(L)} \right) \quad (2.59)$$

A camada de saída do modelo de classificação com MLP é utilizada a função softmax para transformar as saídas em probabilidades. Na Equação 2.60 é apresentada a função softmax:

$$\text{softmax}(z)_k = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}} \quad (2.60)$$

onde z_k é a saída da camada oculta final para a classe k .

Além disso, na função de perda geralmente usado no treinamento é a entropia cruzada, conforme apresentado na Equação 2.61:

$$\mathcal{L}(y, f_{MLP}(x)) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \log(f_{MLP}(x_i)_k) \quad (2.61)$$

onde y_{ik} assume o valor de 1 se a instância i pertence à classe k ; caso contrário, o valor é 0 (Bishop; Nasrabadi, 2006; Marsland, 2015).

Por outro lado, no processo de treinamento do MLP, além da utilização do algoritmo de retropropagação, as funções de ativação não lineares desempenham um papel fundamental, pois introduzem não linearidades no modelo, permitindo que a rede neural aprenda relações complexas (Bishop; Nasrabadi, 2006; Alpaydin, 2021). A seguir, apresentam-se as funções de ativação mais comuns:

- **Rectified Linear Unit (ReLU):** A função ReLU é amplamente empregada devido à sua simplicidade e à sua eficácia na mitigação do problema do desaparecimento do gradiente (*vanishing gradient*).

$$\text{ReLU}(z) = \max(0, z) \quad (2.62)$$

- **Sigmoid:** A função sigmoide é utilizada principalmente em problemas de classificação binária, pois sua saída está restrita ao intervalo $[0, 1]$, facilitando a interpretação dos valores como probabilidades.

$$\sigma = \frac{1}{1 + e^{-z}} \quad (2.63)$$

- Tangente Hiperbólica (Tanh): A função Tanh é semelhante à signóide, mas mapeia os valores para o intervalo $[-1, 1]$, proporcionando uma melhor centralização dos dados em torno de zero, o que pode facilitar o aprendizado do modelo.

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (2.64)$$

2.7 OTIMIZAÇÃO DE HIPERPARÂMETROS E AVALIAÇÃO DOS MODELOS

2.7.1 Otimização de Hiperparâmetros

Desenvolver modelos de ML pode ser um processo complexo devido ao tempo computacional envolvido e à necessidade de identificar a técnica adequada para ser utilizada no problema em questão. Na literatura, há uma variedade de técnicas de ML que podem ser usadas, e o desempenho dos modelos pode variar dependendo da configuração de seus hiperparâmetros (Alpaydin, 2021).

A configuração de hiperparâmetros é uma etapa importante para melhorar a adaptação e a precisão do modelo, permitindo que os parâmetros aprendam seus valores a partir dos dados. Especificamente, os hiperparâmetros são parâmetros externos que controlam o processo de treinamento dos modelos, os quais podem ser ajustados manualmente (*babysitting*) ou por meio de um processo de otimização (Probst; Boulesteix; Bischl, 2019).

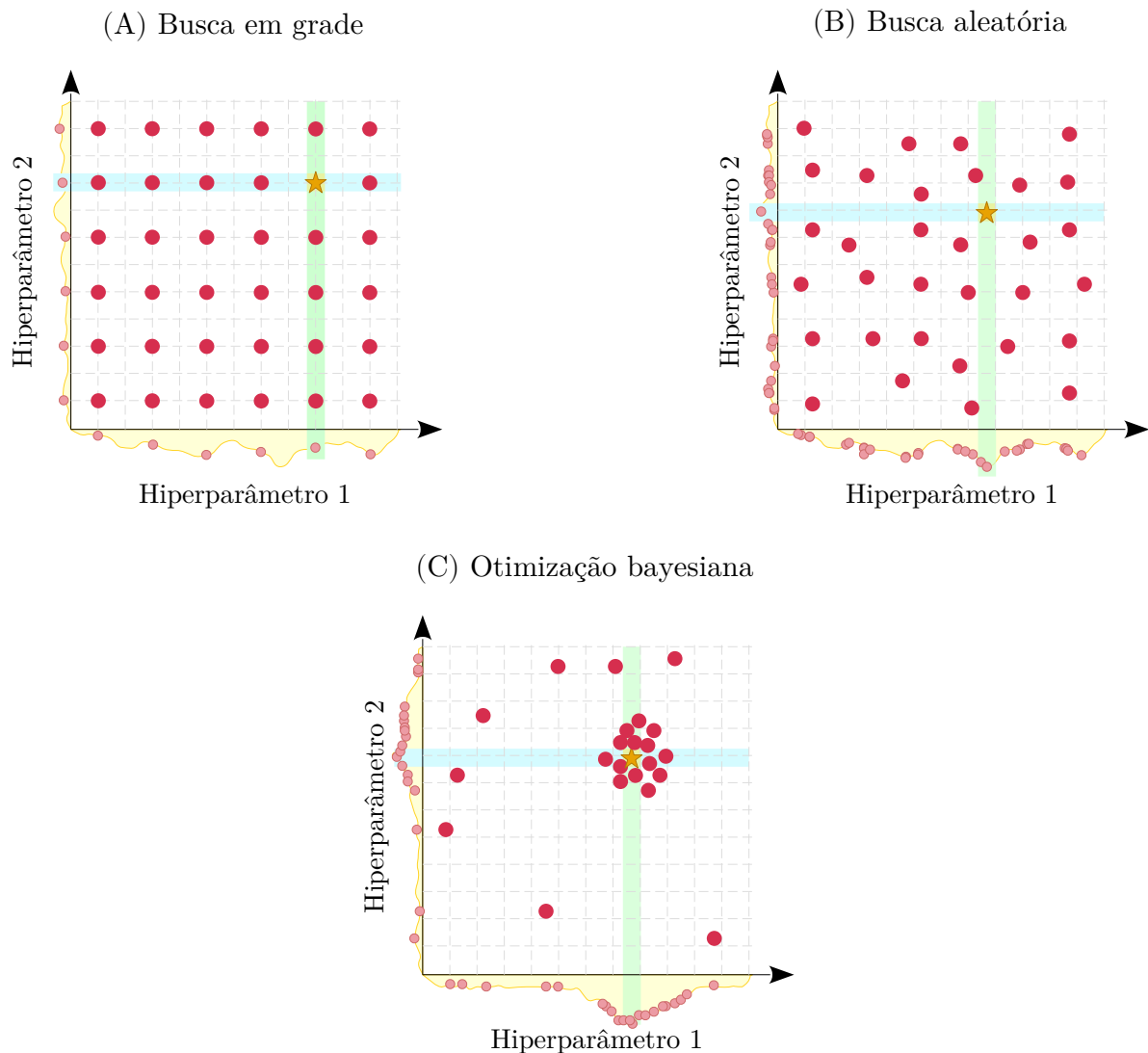
Assim, a busca pelos melhores hiperparâmetros permite minimizar a taxa de erro dos modelos, sendo que o domínio de busca pode envolver valores discretos, contínuos, categóricos ou binários. Em conformidade com Yang & Shami (2020), nos modelos supervisionados o valor ótimo do estimador pode ser obtido por meio da Equação 2.65:

$$f^* = \arg \min_{f \in F} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(\mathbf{x}_i), y_i) \quad (2.65)$$

sendo n o número de instâncias utilizadas no treinamento, $\mathbf{x}_i$ e y_i são as variáveis de entrada e saída, f^* o estimador com os hiperparâmetros ótimos, que varia de acordo com a estrutura do modelo, na qual a função f é restrito a um conjunto de funções F . Além disso, na Equação 2.65, o objetivo é minimizar a função de custo $\mathcal{L}(f(\mathbf{x}_i), y_i)$, que mede a diferença entre os valores reais e estimados.

De modo geral, os métodos tradicionais de otimização não são adequados para este processo de busca em modelos de ML; portanto, outros métodos de otimização apropriadas podem ser aplicadas (Probst; Boulesteix; Bischl, 2019). De acordo com Probst, Boulesteix & Bischl (2019) e Yang & Shami (2020), entre os métodos de otimização mais populares estão os métodos de *babysitting*, busca em grade (*grid search*), busca aleatória (*random search*) e otimização bayesiana, que são descritos a seguir:

Figura 30 – Métodos de otimização de hiperparâmetros



Fonte: Elaborado com informações de Probst, Boulesteix & Bischl (2019) e Yang & Shami (2020).

- ***Babysitting***

O método *babysitting* consiste em um procedimento simples de ajuste manual, também denominado como método de "tentativa e erro". Esse processo envolve a verificação de n possíveis valores de hiperparâmetros, selecionados com base na experiência do pesquisador. No entanto, como desvantagem significativa, o *babysitting* não é viável para algumas técnicas de ML que requerem um número maior de hiperparâmetros, o que, por sua vez, exigiria um tempo maior para as iterações.

- ***Busca em grade***

O método de busca em grade (Figura 30 (A)), também conhecido como busca exaustiva, é um processo de busca e seleção da melhor configuração obtida a partir do produto cartesiano de um conjunto finito de valores possíveis. A desvantagem deste método é

que o produto cartesiano das configurações gera um espaço de busca de alta dimensão, resultando em tempo de processamento e custo computacional bastante elevados.

- **Busca aleatória**

O método de busca aleatória (Figura 30 (B)), ajuda a superar as limitações da busca em grade, por meio de uma seleção aleatória de possíveis configurações a partir de um conjunto inicial de valores definidos pelo pesquisador. Este método é útil para identificar configurações ótimas globais ou quase globais. No entanto, a principal limitação é que, na busca, cada iteração é independente da avaliação de iterações anteriores.

- **Otimização bayesiana**

O método de otimização bayesiana (Figura 30 (C)) diferentemente dos outros métodos, constrói um modelo probabilístico que permite explorar o espaço de busca de valores possíveis de forma eficiente. Este processo parte de um conjunto inicial de configurações, que ao longo das iterações, ajusta suas previsões sobre o espaço de busca, quantificando a incerteza em regiões inexploradas, o que permite que iterações subsequentes sejam realizadas em áreas mais promissoras. Como principal vantagem, pode-se destacar a sua capacidade de explorar grandes espaços de busca com menor consumo computacional. No entanto, as limitações incluem complexidade computacional ao trabalhar com grandes volumes de dados e espaços de busca de alta dimensão.

2.7.2 Métricas de Desempenho

De acordo com a literatura, não há uma única métrica de avaliação padrão para quantificar o erro das previsões (ou predições) dos modelos (Rodríguez *et al.*, 2022). Portanto, nesta pesquisa, a avaliação e a seleção dos modelos foi realizada por meio da comparação de um conjunto de métricas de avaliação bem conhecidas na literatura tanto para problemas de regressão como de classificação.

- **Métricas para problemas de regressão**

As métricas de erro usadas foram: Mean Absolute Error (MAE), MSE, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE) e Mean Poisson Deviance (MPD) (Faceli *et al.*, 2021; Rodriguez *et al.*, 2022). Os cálculos dessas métricas podem ser realizadas por meio das Equações 2.66 – 2.71

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.66)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.67)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.68)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \times 100 \right| \quad (2.69)$$

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \left(\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \right) \quad (2.70)$$

$$\text{MPD} = 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{\hat{\mu}} \right) - (y_i - \hat{\mu}) \right] \quad (2.71)$$

sendo, em cada uma das equações, y_i o valor real, $\hat{\mu}$ é a média de $\hat{y}_i$ e $\hat{y}_i$ é o valor previsto pelo modelo (Faceli *et al.*, 2021; Rodriguez *et al.*, 2022).

Adicionalmente, o melhor modelo também pode ser selecionado por meio do teste-t pareado, para verificar as diferenças significativas entre os valores reais e previstos, adotando um nível de significância (Rodriguez *et al.*, 2022).

- **Métricas para problemas de classificação**

As métricas de erro utilizadas para modelos de classificação foram a matriz de confusão e suas métricas associadas de acurácia, precisão, revocação, F1-Score, Área sob a Curva, em inglês *Area Under the Curve* (AUC) e a curva Característica Operativa do Receptor, em inglês *Receiver Operating Characteristic* (ROC), além do teste estatístico de McNemar para comparar os modelos por pares (Alpaydin, 2021).

A Figura 31 apresenta a matriz de confusão:

Figura 31 – Matriz de confusão

		Valores estimados	
		Positivos	Negativos
Valores reais	Positivos	VP Verdadeiro Positivo	FP Falso Positivo
	Negativos	FN Falso Negativo	VN Verdadeiro Negativo

Fonte: Adaptação de Alpaydin (2021).

A matriz de confusão ilustrada na Figura 31 é uma representação matricial que permite visualizar as previsões e o desempenho do modelo de classificação (Marsland, 2015).

Conforme Marsland (2015), Chicco (2020) e Alpaydin (2021), são descritas as métricas de acurácia, precisão, revocação, F1-Score, AUC e a curva ROC, apresentadas a seguir:

A acurácia (A_{cur}) representa a proporção de previsões corretas em relação ao total de instâncias avaliadas, englobando tanto as classificações verdadeiramente positivas quanto as verdadeiramente negativas. Sua fórmula é apresentada na Equação 2.72:

$$A_{cur} = \frac{VP + VN}{VP + FP + FN + VN} \quad (2.72)$$

sendo VP verdadeiro positivo, VN verdadeiro negativo, FP falso positivo e FN falso negativo.

Uma limitação importante dessa métrica é que, na presença de classes desbalanceadas, sua confiabilidade como medida de desempenho pode ser comprometida. Portanto, é fundamental considerar a avaliação de outras métricas para obter uma análise mais abrangente do desempenho do modelo.

A métrica de precisão ($Prec$) quantifica a proporção de previsões positivas corretas em relação ao total de instâncias classificadas como positivas pelo modelo. Sua fórmula é apresentada na Equação 2.73:

$$Prec = \frac{VP}{VP + FP} \quad (2.73)$$

Por outro lado, a métrica de revocação (Rev), também conhecido como sensibilidade, também conhecida como sensibilidade, mede a capacidade do modelo de identificar corretamente as instâncias positivas em relação ao total de instâncias verdadeiramente positivas. A fórmula para calcular essa métrica de desempenho é apresentada na Equação 2.74:

$$Rev = \frac{VP}{VP + FN} \quad (2.74)$$

A medida F1-Score ($F1$) é amplamente utilizada e considerada muito importante, pois é a única medida que combina as métricas de precisão e revocação. A fórmula para calcular o $F1$ está apresentada na Equação 2.75, que corresponde à média harmônica dos valores de $Prec$ e Rev .

$$F1 = 2 \times \frac{Prec \times Rev}{Prec + Rev} \quad (2.75)$$

Os valores da métrica $F1$ variam também de 0 e 1. Quando o valor do $F1$ está próximo de 1, indica uma classificação perfeita, com FN e FP próximos de zero.

A curva ROC e a AUC são ferramentas fundamentais para avaliação de modelos de classificação. Especificamente, a curva ROC é uma representação gráfica do desempenho do classificador, que relaciona a Taxa de Verdadeiros Positivos (TVP) no eixo y com a Taxa de Falsos Positivos (TFP) no eixo x (Marsland, 2015). As métricas de TVP e TFP são definidas conforme as Equações 2.76 e 2.77:

$$TVP = \frac{VP}{VP + FN} \quad (2.76)$$

$$TFP = \frac{FP}{FP + VP} \quad (2.77)$$

A AUC quantifica a área sob a curva ROC, representando a probabilidade do modelo de classificar corretamente um par aleatório de amostras positivas e negativas. Seu cálculo pode ser aproximado pela regra dos trapézios da curva ROC, conforme apresentado na Equação 2.78:

$$AUC = \sum_{i=1}^{n-1} (TFP_{i+1} - TFP_i) \times \frac{TVP_{i+1} + TVP_i}{2} \quad (2.78)$$

Por fim, o teste estatístico de McNemar é utilizado para comparar pares de modelos, analisando a frequência de acertos e erros de cada um. A Figura 32 apresenta a matriz de contingência, onde são mostrados os contagens de acertos e erros dos modelos comparados.

Figura 32 – Matriz de McNemar

erro_{M1,M2} Número de dados classificados incorretamente por ambos	erro_{M1} Número de dados mal classificados apenas por $M1$
erro_{M2} Número de dados mal classificados apenas por $M2$	sucesso_{M1,M2} Número de dados classificados corretamente por ambos

Fonte: Adaptação de Alpaydin (2021).

A comparação entre os modelos baseia-se na hipótese nula de que ambos os modelos apresentam a mesma taxa de erro ($H_0 : error_{M1} = error_{M2}$), e que a taxa de erro de cada

um é igual à indicada pela Equação 2.79.

$$\frac{erro_{M1} + erro_{M2}}{2} \quad (2.79)$$

A estatística de McNemar (X^2) é calculada conforme a Equação 2.80:

$$X^2 = \frac{(|erro_{M1} - erro_{M2}| - 1)^2}{erro_{M1} - erro_{M2}} \quad (2.80)$$

O teste de McNemar rejeita a hipótese nula com um nível de significância α se o valor de X^2 for maior que $X_{\alpha,1}$.

2.8 PROCESSOS METODOLÓGICOS EM APRENDIZADO DE MÁQUINA

A implementação de modelos de ML requer a execução de uma série de etapas, sendo possível encontrar na literatura alguns procedimentos que podem ser aplicados por serem flexíveis e apresentarem bons resultados de seu uso em diversas áreas de estudo.

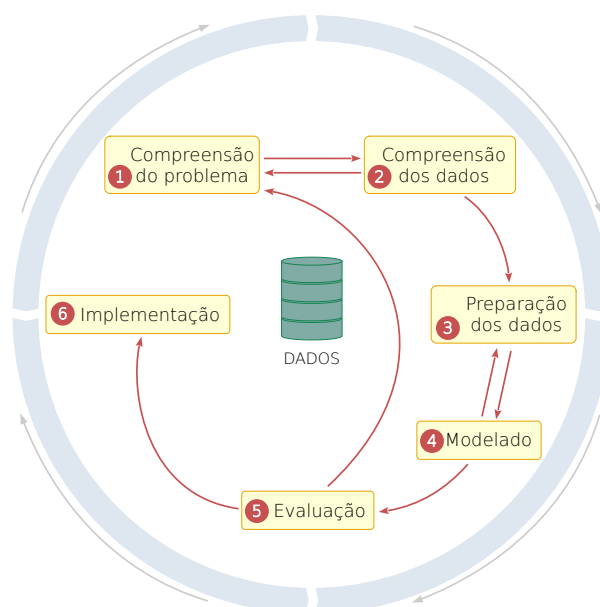
Dentre os processos metodológicos mais populares estão: Processo Padrão Inter-Indústrias para Mineração de Dados (CRISP-DM), Descoberta de Conhecimento em Bases de Dados (KDD) e Amostra, Explorar, Modificar, Modelar e Avaliar (SEMMA). Esses processos são descritos a seguir, e caracterizam-se por ter uma abordagem sólida e estruturada no desenvolvimento de pesquisas com mineração de dados e ML (Swamynathan, 2019).

2.8.1 Processo Padrão Inter-Indústrias para Mineração de Dados

O CRISP-DM, acrônimo do inglês *Knowledge Discovery Databases*, é um processo iterativo, confiável e capaz de se adaptar a diferentes problemas de estudo, que foi proposta inicialmente em 1996 (Swamynathan, 2019; Ghavami, 2020). Conforme mostrado na Figura 33, a aplicação dessa metodologia é dividida em seis etapas principais, que são brevemente descritos abaixo.

- *Etapa 1 – Compreensão do problema:* Etapa orientada para a compreensão e definição do problema, seguido pela definição dos objetivos e a identificação dos requisitos da pesquisa.
- *Etapa 2 – Compreensão dos dados:* Etapa que inclui a coleta do conjunto de dados alinhada com os objetivos definidos. Além disso, esta etapa centra-se sobretudo na seleção e compreensão das características, seguida da verificação da sua qualidade.
- *Etapa 3 – Preparação dos dados:* Após a coleta de dados, eles são preparados por meio de técnicas de transformação, correção de dados faltantes, limpeza do banco de dados e/ou seleção de variáveis, a fim de dar-lhes um formato adequado para seu processamento.

Figura 33 – Processo metodológico de CRISP-DM



CRISP-DM: *Cross-Industry Standard Process for Data Mining*

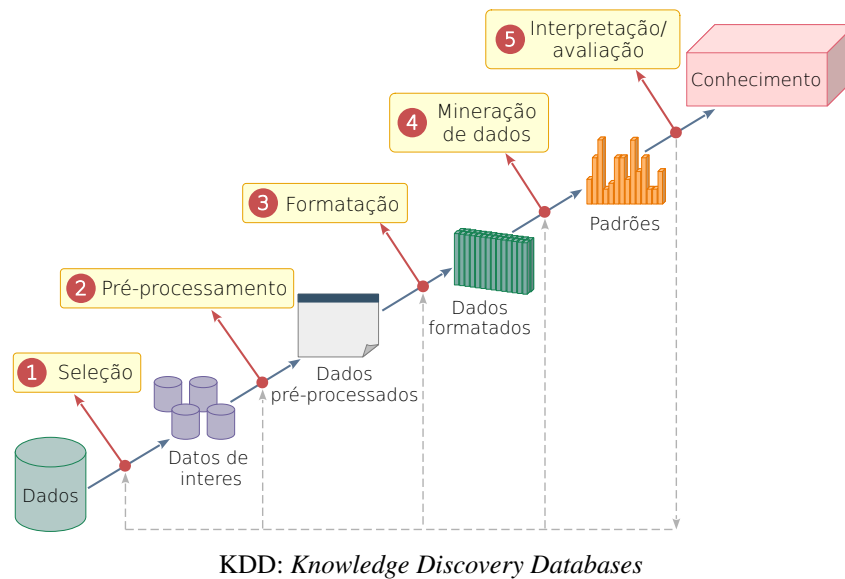
Fonte: Adaptação de Swamynathan (2019) e Ghavami (2020).

- *Etapa 4 – Modelado*: Etapa em que as técnicas a serem utilizadas são selecionadas com base no objetivo principal do modelo. As técnicas selecionadas são aplicadas aos dados e seus parâmetros são calibrados para obter um modelo otimizado.
- *Etapa 5 – Avaliação*: Os modelos desenvolvidos na etapa anterior são avaliados por meio de diferentes métricas estatísticas, a fim de garantir que os resultados atendam ao objetivo da investigação. Além disso, para tomar uma decisão sobre o modelo final obtido, as etapas anteriores e seus resultados são cuidadosamente revisados.
- *Etapa 6 – Implementação*: Na etapa final é desenvolvido um plano de implementação para o modelo final selecionado, seguido de um relatório final.

2.8.2 Descoberta de Conhecimento em Bases de Dados

O KDD, acrônimo do inglês *Knowledge Discovery Databases*, foi utilizado e implementada inicialmente entre 1991 e 1996, e refere-se particularmente ao processo de descoberta de conhecimento baseado nos dados (Frawley; Piatetsky-Shapiro; Matheus, 1992; Fayyad; Piatetsky-Shapiro; Smyth, 1996). A aplicação do processo iterativo do KDD consta de cinco etapas (Swamynathan, 2019; Ghavami, 2020), conforme ilustrado na Figura 34:

Figura 34 – Processo metodológico de KDD



Fonte: Adaptação de Fayyad, Piatetsky-Shapiro & Smyth (1996), Swamynathan (2019) e Ghavami (2020).

Cada etapa da metodologia de KDD é descrita a seguir:

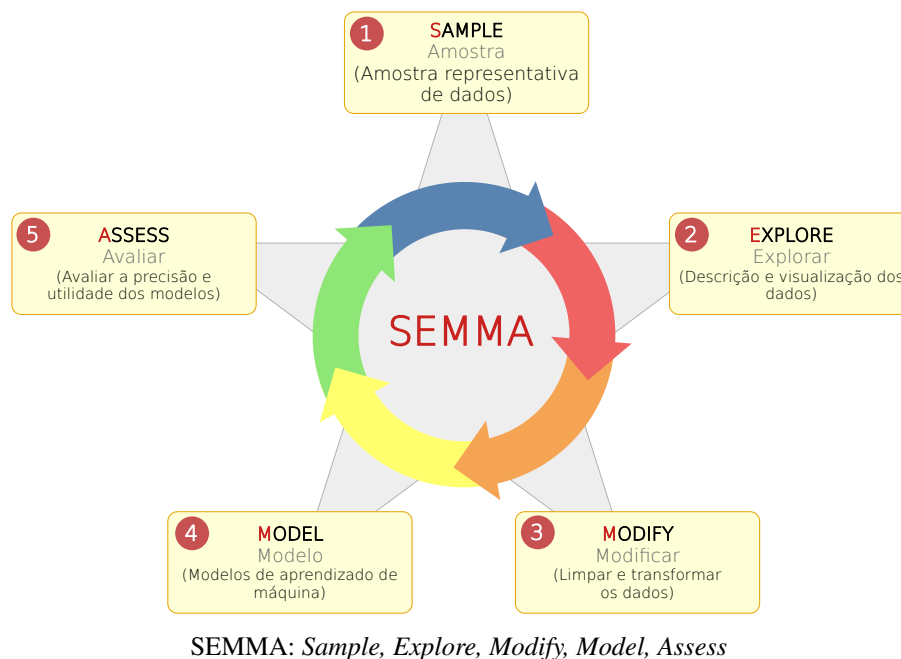
- *Etapa 1 – Seleção*: Etapa que consiste na identificação, seleção e integração dos dados que serão utilizados para atender o problema estudado. Os dados coletados podem vir de diferentes fontes e ter diferentes formatos.
- *Etapa 2 – Pré-processamento*: Processo orientado para a limpeza e pré-processamento com o intuito de corrigir valores faltantes, ruído ou inconsistências nos dados, a fim de melhorar sua qualidade.
- *Etapa 3 – Formatação*: Focado na preparação de dados, com consolidação do banco de dados no formato adequado para aplicação de análises estatísticas, mineração de dados ou ML.
- *Etapa 4 – Mineração de dados*: Compõe-se pelo processo de extração de padrões através do uso de técnicas de exploração estatística, técnicas de mineração de dados e/ou técnicas de ML, aplicadas para a criação de modelos de classificação, agrupamento, regressão, etc.
- *Etapa 5 – Interpretação/avaliação*: Processo focado na validação e interpretação dos resultados obtidos na etapa anterior.

2.8.3 Amostra, Explorar, Modificar, Modelar e Avaliar

O SEMMA, acrônimo do inglês *Sample, Explore, Modify, Model, Assess*, é um processo metodológico semelhante do CRISP-DM e KDD. O SEMMA foi proposto pelo Instituto SAS,

e caracteriza-se por permitir a organização e exploração dos dados, além do desenvolvimento e manutenção de projetos de mineração de dados e/ou ML (Statistical Analysis System, 2017; Ghavami, 2020). Conforme apresentado na Figura 35 este processo compreende a execução de cinco etapas.

Figura 35 – Processo metodológico de SEMMA



Fonte: Adaptação de Statistical Analysis System (2017), Swamynathan (2019) e Ghavami (2020).

A seguir são descritos cada uma das etapas da metodologia SEMMA.

- *Etapa 1 – Amostra:* São coletados os dados relacionados ao problema de estudo. A extração de dados é um processo cauteloso, que deve conter uma amostra representativa da realidade com informações relevantes para a identificação de padrões.
- *Etapa 2 – Exploração:* Etapa em que os dados são analisados por meio de técnicas univariadas, bivariadas ou multivariadas, a fim de identificar as relações existentes entre as variáveis. Neste processo é comum o uso de gráficos de visualização para explorar os dados.
- *Etapa 3 – Modificação:* Esta etapa está relacionada à anterior e, se necessário, os dados são limpos, transformados ou selecionados para serem utilizados na próxima etapa. Ou seja, o conjunto de dados final é preparado neste processo.
- *Etapa 4 – Modelo:* As técnicas que serão utilizadas são selecionadas, e logo são aplicadas aos dados para a criação dos modelos. Esse processo é iterativo em um esforço para obter um modelo confiável que atenda aos requisitos do problema.

- *Etapa 5 – Avaliação:* É o processo final que consiste em avaliar o modelo final por meio de métricas estatísticas. Com os resultados deste processo, decide-se se o modelo é útil para o problema abordado.

3 MATERIAIS E MÉTODOS

Neste Capítulo aborda-se a descrição da área geográfica de estudo, do conjunto de dados coletados e a descrição da malha municipal. Além disso, é feito o detalhamento do procedimento metodológico seguido para o desenvolvimento desta pesquisa, além de apresentar as considerações éticas.

3.1 LOCAL DE ESTUDO

Os dados deste estudo foram da RMVPLN, que é uma das nove regiões metropolitanas do estado de São Paulo (Assembleia Legislativa do Estado de São Paulo, 2022). Em conformidade com a Lei no. 13.089 de 12 de janeiro de 2015, uma metrópole é um espaço territorial urbano, constituído pelo agrupamento de municípios limítrofes, que têm influência regional ou nacional, seja por sua população, relevância socioeconômica ou relevância política (Brasil, 2015).

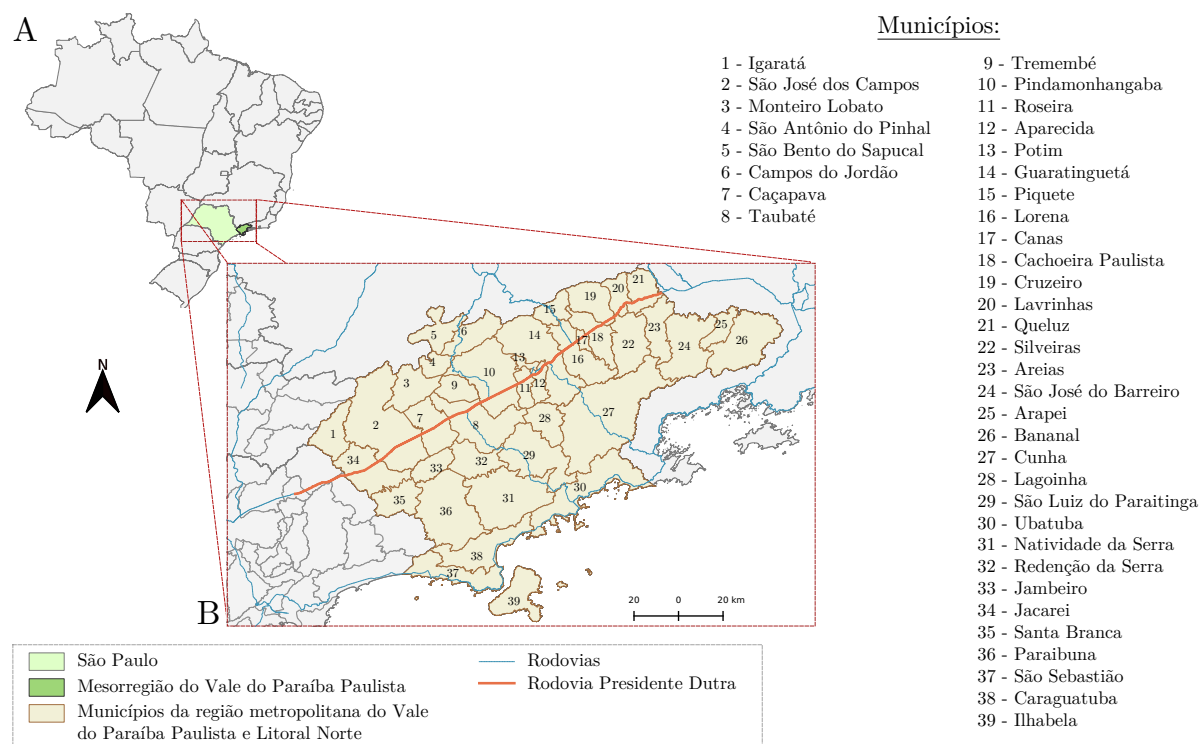
A RMVPLN tem uma área de 16.178 km^2 , equivalente a 6,5% do estado de São Paulo (EMTU/SP, 2022). Outrossim, a região possui aproximadamente 2,5 milhões de habitantes, correspondendo a aproximadamente 5,5% da população do estado, o que a torna a quarta região com maior concentração de habitantes após as regiões metropolitanas de Sorocaba, Campinas e São Paulo (Mello; Teixeira; Mello, 2016).

Além disso, a RMVPLN contém os mesmos municípios da mesorregião do Vale do Paraíba Paulista do estado de São Paulo (Figura 36 (A)), constituída por 39 municípios, conforme ilustrado na Figura 36 (B), agrupados em cinco sub-regiões.

As cinco sub-regiões desta região metropolitana estão organizadas pelos municípios de:

- Sub-região 1 de São José dos Campos: Igaratá, Jacareí, Caçapava, Jambeiro, Paraibuna, Santa Branca, Monteiro Lobato e São José dos Campos.
- Sub-região 2 de Taubaté: Lagoinha, Natividade da Serra, Campos do Jordão, Pindamonhangaba, Redenção da Serra, São Bento do Sapucaí, São Luiz do Paraitinga, Santo Antônio do Pinhal, Taubaté e Tremembé.
- Sub-região 3 de Guaratinguetá: Aparecida, Roseira, Cunha, Potim, Piquete, Guaratinguetá, Lorena, Canas e Cachoeira Paulista.
- Sub-região 4 de Cruzeiro: Arapeí, Bananal, Cruzeiro, Areias, São José do Barreiro, Lavrinhas, Queluz e Silveiras.
- Sub-região 5 do Litoral Norte: Ubatuba, Caraguatatuba, Ilhabela e São Sebastião.

Figura 36 – Região metropolitana do Vale do Paraíba e Litoral Norte

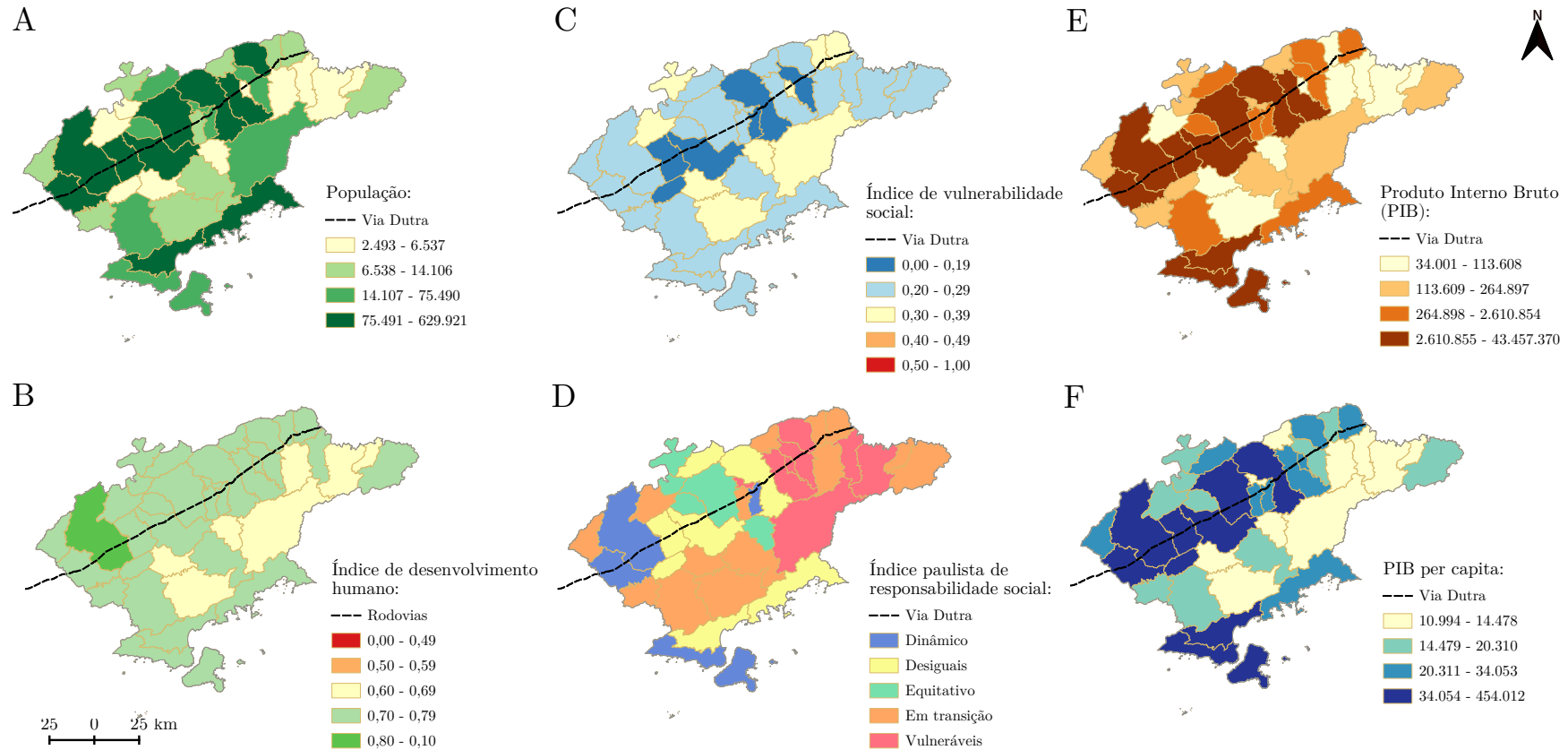


Fonte: Elaborado pela autora (2025).

Conforme ilustrado na Figura 36, a região é cortada pela Rodovia Presidente Dutra, que desde sua construção trouxe desenvolvimento industrial, principalmente crescimento nos municípios localizados ao longo da Via Dutra (Mello; Teixeira; Mello, 2016). Esse desenvolvimento municipal, foi observado no Índice de Desenvolvimento Humano (IDH) da região, do último censo de 2010, onde os municípios com maiores valores foram aqueles cortados pela rodovia (Nascimento *et al.*, 2007), além dos municípios do litoral norte (Figura 37 (B)).

Da mesma forma, de acordo com as informações fornecidas pela Fundação Sistema Estadual de Análise de Dados (SEADE) e pelo Instituto de Pesquisa Econômica Aplicada (IPEA), os municípios com menor Índice de Vulnerabilidade Social (IVS) e maior Produto Interno Bruto (PIB) e PIB *per capita* foram os localizados ao longo da Via Dutra e dos municípios do litoral norte, como ilustra a Figura 37 (C-E-F) (Instituto de Pesquisa Econômica Aplicada, 2010; Fundação SEADE, 2018b; Fundação SEADE, 2018a). Os valores apresentados na Figura 37 encontram-se na Tabela 34 do Anexo - A.

Figura 37 – Indicadores sociais da região metropolitana do Vale do Paraíba e Litoral Norte

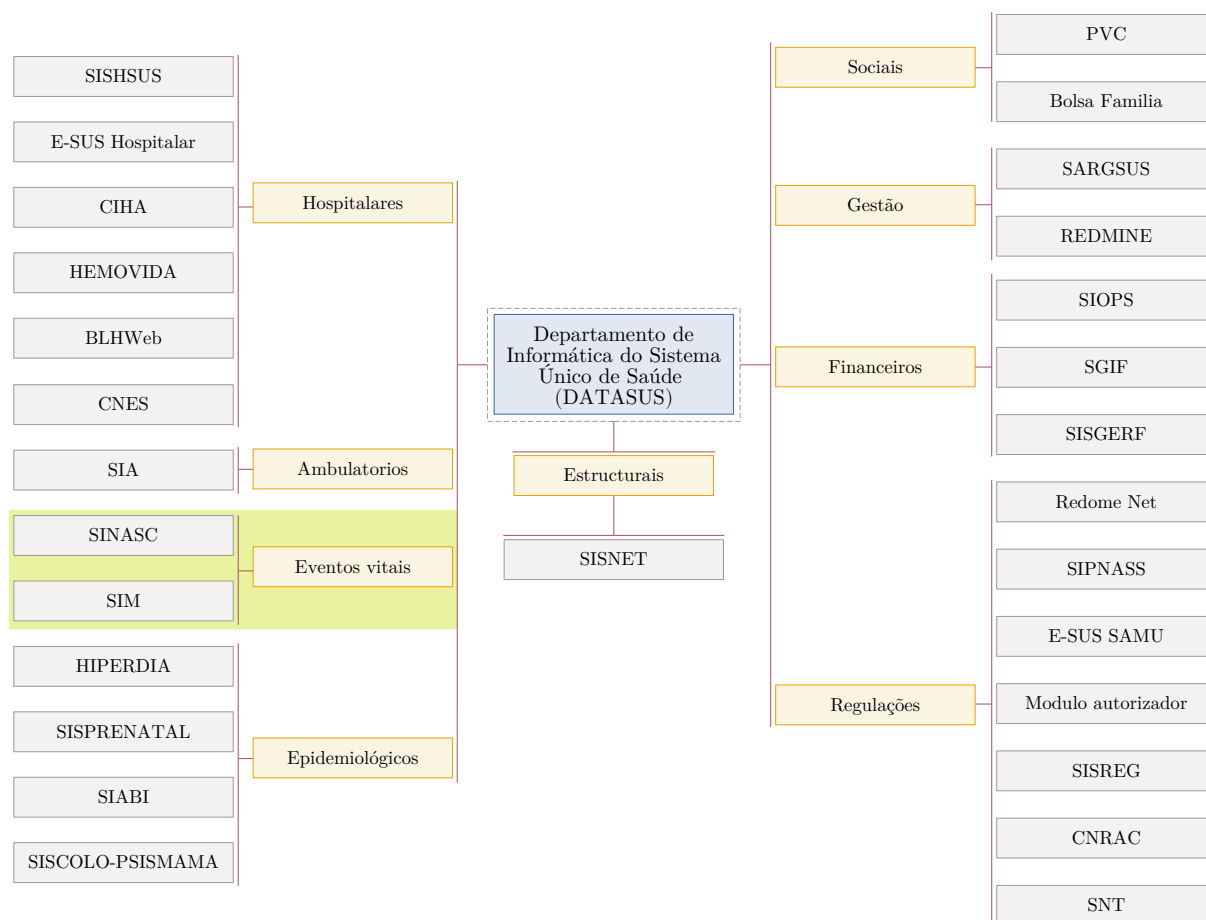


Fonte: Elaborado pela autora (2025), com dados de DATASUS (2010), Instituto de Pesquisa Econômica Aplicada (2010), Fundação SEADE (2018b), Fundação SEADE (2018a).

3.2 BANCO DE DADOS E MALHA CARTOGRÁFICA

O Departamento de Informática do Sistema Único de Saúde (DATASUS) foi a principal fonte utilizada para a coleta de dados deste trabalho. O DATASUS caracteriza-se por disponibilizar uma grande quantidade de dados referente a diferentes sistemas, além de proporcionar amplo acesso a informações que permitem analisar e avaliar os dados para o planejamento de novas políticas públicas para melhorar a qualidade do serviço ou reduzir riscos de mortalidade e/ou morbidade, como apresentado na Figura 38.

Figura 38 – Sistemas que integram o repositório do DATASUS



Fonte: Adaptação de Marys & Autran (2019).

Os dados referentes aos nascidos vivos de mães residentes dos municípios da RMVPLN e os óbitos ocorridos durante o primeiro ano de vida foram coletados do sistema relacionado a eventos vitais (Figura 38), especificamente do Sistema de Informações Sobre Nascidos Vivos (SINASC) e Sistema de Informações sobre Mortalidade (SIM), disponível na seção de transferência de arquivos do DATASUS.

O SINASC baseia-se na declaração de nascido vivo (Anexo - B), e tem por finalidade caracterizar as condições da gestação, do parto e do nascimento (Leal *et al.*, 2018). O SIM

contém informações relacionadas à declaração de óbito (Anexo - C), permitindo a análise e avaliação da magnitude da mortalidade (Correia; Padilha; Vasconcelos, 2014).

Dessa forma, foram coletados dados do sistema SINASC correspondentes ao período de 2010 a 2020. Quanto às variáveis do banco de dados do SINASC, elas tiveram mudança na quantidade de variáveis disponibilizadas durante o período de estudo, com 57 colunas em 2010 para 62 colunas no 2020. Entre as principais variáveis de interesse para o desenvolvimento desta investigação estão aquelas apresentadas no Quadro 5, categorizadas conforme disponibilizados pelo DATASUS (maior detalhe das variáveis do SINASC no Anexo – D).

Quadro 5 – Variáveis do conjunto de dados do SINASC

(continua)

Variável	Tipo	Descrição
LOCNASC	Categórico	Local de nascimento.
IDADEMAE	Numérico	Idade materna
ESTCIVMAE	Categórico	Estado civil da mãe.
ESCMAE	Categórico	Escolaridade materna.
CODOCUPMAE	Categórico	Código da ocupação materna, conforme o Código Brasileiro de Ocupações
QTDFILVIVO	Numérico	Número de filhos vivos.
QTDFILMORT	Numérico	Número de filhos mortos (perdas fetais ou abortos).
CODMUNRES	Categórico	Código IBGE ^a do município de residência.
GESTACAO	Categórico	Semanas de gestação da mãe.
GRAVIDEZ	Categórico	Tipo de gravidez.
PARTO	Categórico	Tipo de parto.
CONSULTAS	Categórico	Número de consultas de pré-natal.
DTNASC	Numérico	Data de nascimento do recém-nascido: <i>dd mm aaaa.</i>
HORANASC	Numérico	Hora do nascimento: <i>hh mm.</i>
SEXO	Categórico	Sexo do recém-nascido.
APGAR1	Numérico	<i>Apgar</i> do recém-nascido no 1º minuto, de 0 – 10.
APGAR5	Numérico	<i>Apgar</i> do recém-nascido no 5º minuto, de 0 – 10.
RACACOR	Categórico	Raça/cor do recém-nascido.
PESO	Numérico	Peso do recém-nascido em gramas.
PARIDADE	Categórico	Define se é a primeira gravidez ou não.
IDANOMAL	Categórico	Anomalia identificada no recém-nascido.
CODANOMAL	Categórico	Código CID-10 ^b da anomalia.
RACACORMAE	Categórico	Raça/cor da mãe.
QTDGESTANT	Numérico	Número de gestações passadas.
QTDPARTNOR	Numérico	Número de partos vaginais passados.
QTDPARTCES	Numérico	Número de partos cesáreos passados.
IDADEPAI	Numérico	Idade do pai do recém-nascido.
DTULTMENST	Numérico	Data da última menstruação: <i>dd mm aaaa.</i>
SEMAGESTAC	Numérico	Número de semanas de gestação.
CONSPRENAT	Numérico	Número de consultas de pré-natal.

(a) IBGE: Instituto Brasileiro de Geografia e Estatística, (b) CID-10: Classificação Internacional de Doenças.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Quadro 5 – Variáveis do conjunto de dados do SINASC

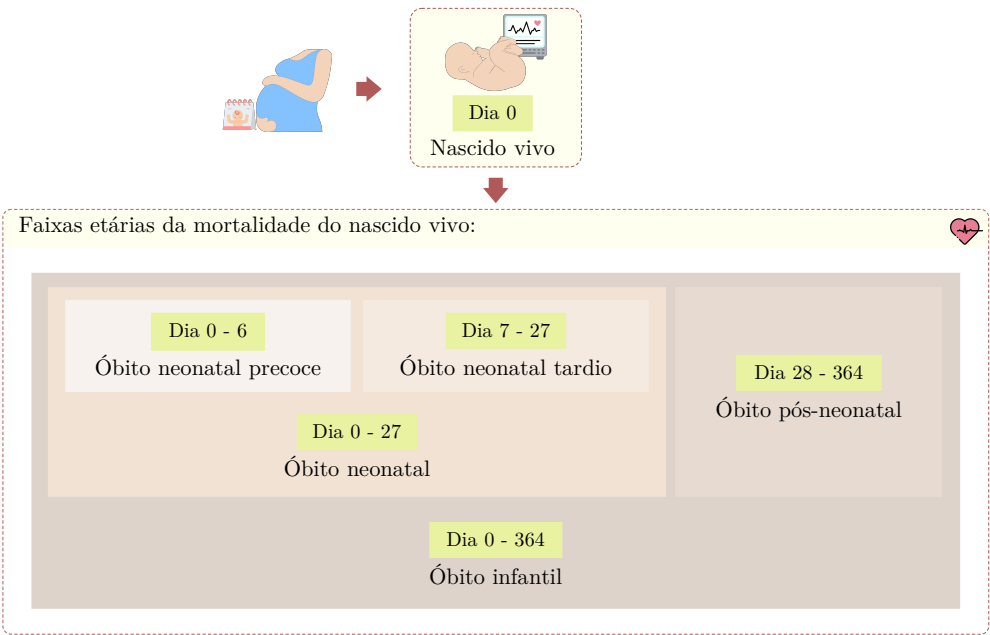
(conclusão)

Variável	Tipo	Descrição
MESPRENAT	Categórico	Mês de gestação da mãe em que iniciou o pré-natal.
TPAPRESENT	Categórico	Posição ou apresentação do recém-nascido.
STTRABPART	Categórico	Trabalho de parto induzido.
STCESPARTO	Categórico	Cesárea ocorreu antes de iniciar o trabalho de parto.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Assim, diferentemente do período selecionado para coleta de dados dos nascidos vivos do SINASC (até 2020), para a coleta dos registros de óbito do SIM foi considerado um ano adicional (até 2021), pois, independentemente da duração da gestação, a faixa etária da mortalidade do nascido vivo varia de 0 a 364 dias, conforme mostra a Figura 39. Ou seja, por exemplo, dos nascidos vivos no mês de dezembro de 2020, caso o óbito do recém-nascido tenha ocorrido, este pode ter acontecido na primeira semana de vida, no primeiro mês ou no primeiro ano de vida (2021).

Figura 39 – Divisão da faixa etária da mortalidade do nascido vivo no primeiro ano de vida



Fonte: Elaborado pela autora (2025).

Em relação às variáveis do SIM, também apresentaram variação quanto à quantidade disponível no período considerado neste estudo. O banco de dados de 2010 apresentou 60 colunas, e já para 2021 o banco tinha 88 colunas disponíveis, porém nem todas as variáveis do banco de dados foram usadas neste trabalho. O Quadro 6 apresenta as principais variáveis de interesse do SIM para este trabalho, de acordo com as categorias designadas pelo DATASUS (maior detalhe das variáveis do SIM no Anexo – E).

Quadro 6 – Variáveis do conjunto de dados do SIM

Variável	Tipo	Descrição
DTOBITO	Numérico	Data do óbito: <i>dd mm aaaa.</i>
HORAOBITO	Numérico	Hora do óbito: <i>hh : mm</i> (padrão 24 horas).
DTNASC	Numérico	Data de nascimento do óbito: <i>dd mm aaaa.</i>
SEXO	Categórico	Sexo do óbito.
CODMUNRES	Categórico	Código IBGE ^a do município de residência.
LOCOCOR	Categórico	Local de ocorrência do óbito.
IDADE	Numérico	Idade do óbito, refere-se a se a morte ocorreu em minutos, horas, meses ou anos.
CODMUNOCOR	Categórico	Código do município em que ocorreu o óbito.
IDADEMAE	Numérico	Idade materna.
ESMAE	Categórico	Escolaridade materna.
OCUPMAE	Categórico	Código da ocupação materna, conforme o Código Brasileiro de Ocupações.
QTDFILVIVO	Numérico	Número de filhos vivos.
QTDFILMORT	Numérico	Número de filhos mortos (perdas fetais ou abortos).
GRAVIDEZ	Categórico	Tipo de gravidez.
GESTACAO	Categórico	Semanas de gestação da mãe.
PARTO	Categórico	Tipo de parto.
PESO	Numérico	Peso do óbito em gramas.
OBITOPUERP	Categórico	Óbito no puerpério:
CAUSABAS	Categórico	Causa básica do óbito, de acordo com os códigos CID-10 ^b .
CIRCOBITO	Categórico	Tipo de circunstâncias em que se deu o óbito.
TPMORTEOCO	Categórico	Quando a morte ocorreu.
TPOBITOCOR	Numérico	Sem descrição.
MORTEPARTO	Numérico	Sem descrição.
CIRURGIA	Categórico	Indica se houve cirurgia.
LINHAA	Strings	Linha A do atestado, conforme a CID-10 ^b .
LINHAB	Strings	Linha B do atestado, conforme a CID-10 ^b .
LINHAC	Strings	Linha C do atestado, conforme a CID-10 ^b .
LINHAD	Strings	Linha D do atestado, conforme a CID-10 ^b .
LINHAI I	Strings	Linha II do atestado, conforme a CID-10 ^b .
ESMAE2010	Numérico	Escolaridade 2010.
CODMUNNATU	Categórico	Código do município de naturalidade.
CODESTAB	Categórico	Código do estabelecimento.
ASSISTMED	Categórico	Indica se houve assistência médica.

(a) IBGE: Instituto Brasileiro de Geografia e Estatística, (b) CID-10: Classificação Internacional de Doenças.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Por outro lado, embora as bases de dados do SINASC e SIM do DATASUS compartilhem uma grande quantidade de dados e tenham melhorado substancialmente ao longo do tempo, sua qualidade ainda é afetada por campos incompletos (Correia; Padilha; Vasconcelos, 2014), além de ter uma defasagem na disponibilidade dos dados, aproximadamente entre dois ou três anos (Silva *et al.*, 2014).

Em adição, é relevante destacar que os dados do SIM e do SINASC coletados do DATASUS

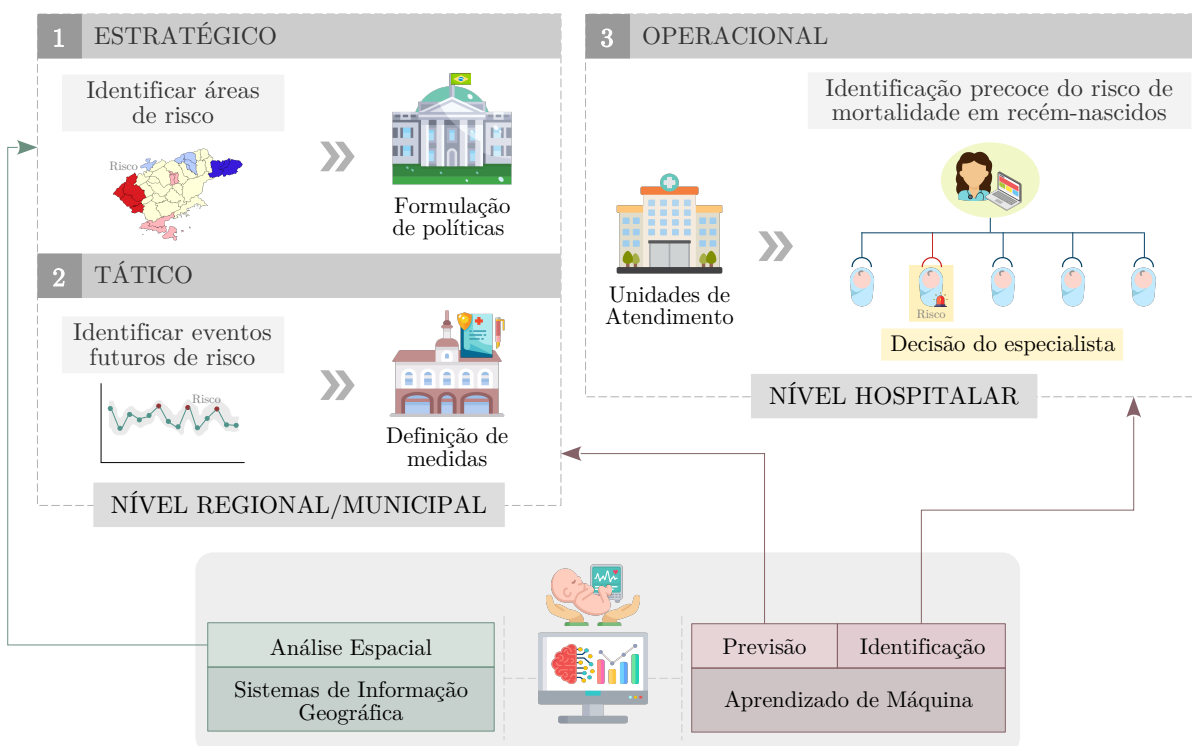
foram extraídos no formato DBC¹, e para facilitar sua utilização foram convertidos no formato CSV², sendo este formato o utilizado ao longo deste trabalho. Por outro lado, cabe destacar que a conversão do formato DBC para CSV foi realizada utilizando código em R, através do RStudio. Mais informações sobre o processo podem ser encontradas no Apêndice D).

Outras variáveis extraídas do DATASUS foram o número de médicos do município (pediatras, médicos da estratégia saúde da família e médicos de família e comunidade) e o número de leitos pediátricos, coletados especificamente do Cadastro Nacional de Estabelecimentos de Saúde (CNES) (DATASUS, 2023). Além disso, as informações sobre o IDH e IVS de cada município foram obtidos do repositório do IPEA (Instituto de Pesquisa Econômica Aplicada, 2010), e os dados relacionados ao PIB e PIB *per capita* municipal foram extraídos do repositório da SEADE (Fundação SEADE, 2018a).

Por outro lado, a malha cartográfica digital dos 39 municípios da RMVPLN foi obtida do Instituto Brasileiro de Geografia e Estatística (IBGE). As malhas digitais do IBGE são uma representação vetorial dos limites municipais do local de estudo, e são referenciadas ao Sistema Geodésico Brasileiro (SIRGAS 2000) (IBGE, 2023).

3.3 PROCEDIMENTO METODOLÓGICO

Figura 40 – Sequência de soluções propostas com análise especial e aprendizado de máquina



Fonte: Elaborado pela autora (2025).

¹ Abreviação do termo inglês DataBase Container.

² Abreviação do termo inglês Comma Separated Values, que é um arquivo simples delimitando por vírgulas.

Para atingir os objetivos definidos, esta tese envolveu uma sequência de diferentes etapas e processos realizados, explorando os métodos de análise espacial e técnicas de ML.

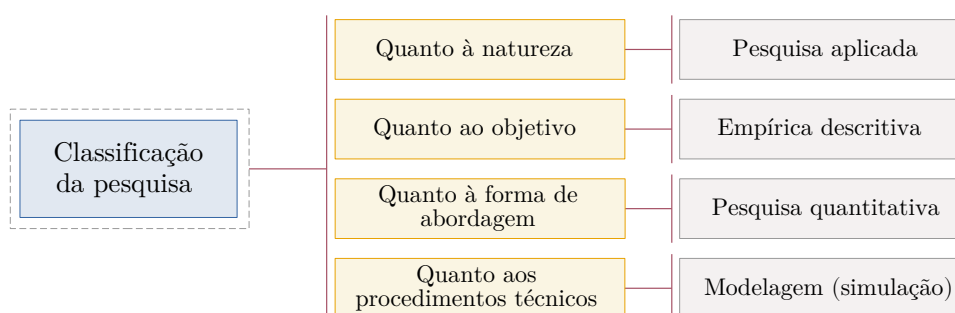
Especificamente, nesta pesquisa é proposta uma sequência de soluções que podem ser utilizadas como suporte para a tomada de decisões tanto em nível estratégico, tático e operacional. A Figura 40 mostra a sequência de soluções propostas nesta pesquisa, sobre a aplicação do SIG com análise espacial e o ML para a mitigação da mortalidade do nascido vivo na RMVPLN.

Portanto, conforme a Figura 40, este estudo apresenta três propostas de ferramentas de apoio à decisão, sendo a primeira a nível estratégico com a identificação de áreas de risco com os SIG e métodos de análise espacial. A segunda proposta a nível tático com a aplicação de ML com séries temporais para previsão do número de óbitos na região analisada. Já a terceira proposta é a nível operacional com a aplicação de ML para a identificação de recém nascido com risco de óbito.

Essas propostas foram levantadas com o objetivo de desenvolver uma sequência de ferramentas que possam servir de apoio aos médicos especialistas, bem como subsidiar a geração de políticas de prevenção.

Deste jeito, de acordo com os tipos de pesquisa (Gerhardt *et al.*, 2009; Prodanov; Freitas, 2013), este trabalho pode ser classificado quanto à sua natureza como pesquisa aplicada, em relação aos seus objetivos é uma pesquisa empírica descritiva, quanto à forma de abordagem é uma pesquisa quantitativa, e no que se refere aos procedimentos técnicos é modelagem e simulação, conforme apresentado na Figura 41.

Figura 41 – Classificação da pesquisa

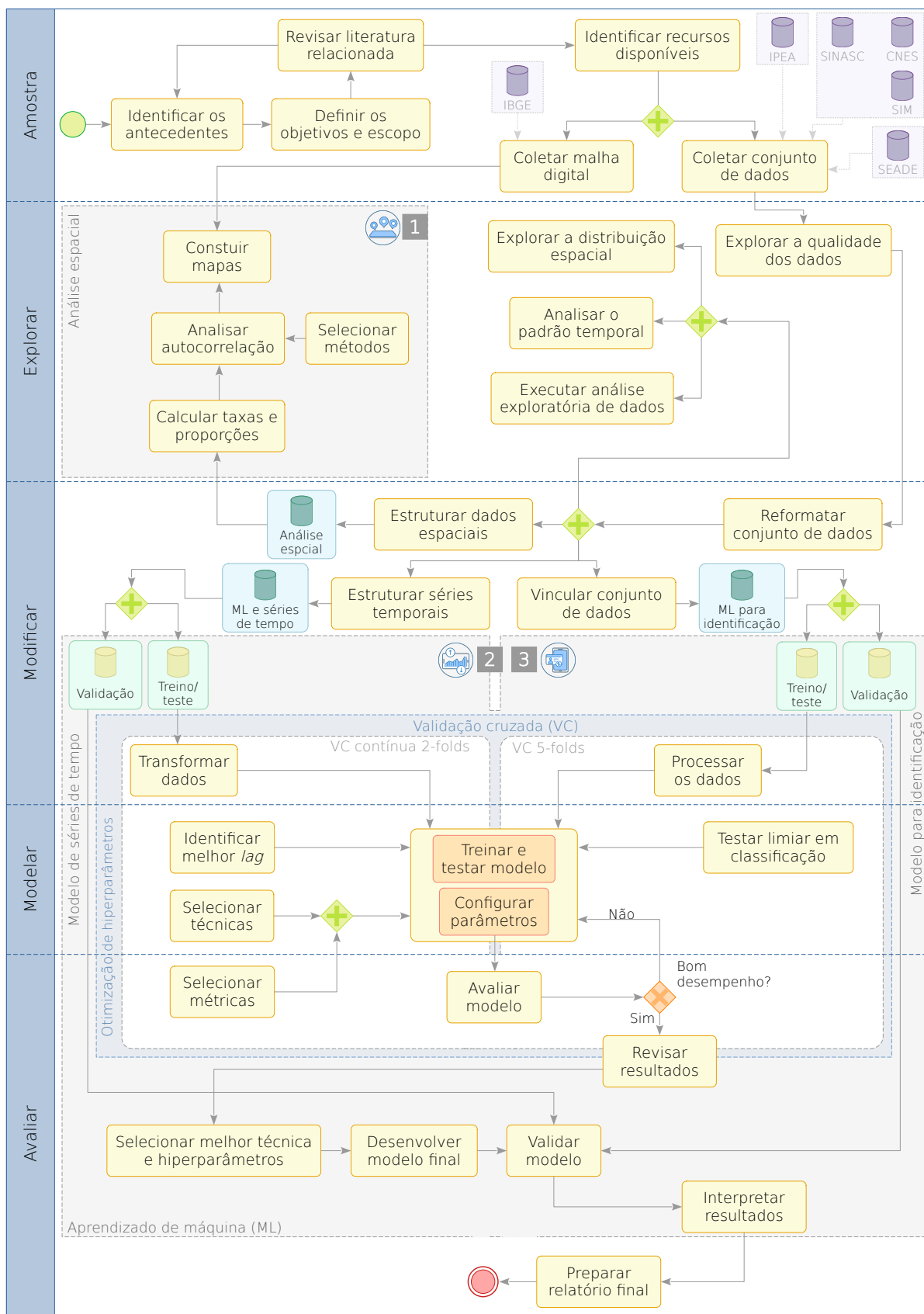


Fonte: Adaptação de Gerhardt *et al.* (2009) e Prodanov & Freitas (2013).

A sistemática metodológica utilizada nesta pesquisa foi SEMMA, por ser popular na área de mineração de dados e ML, aplicado para descoberta de conhecimento em grandes conjuntos de dados (Obenshain, 2004; Ghavami, 2020). Além disso, a SEMMA foi selecionada por ser flexível, iterativa, e fácil de customizar (Swamynathan, 2019), pois, nesta investigação, além de técnicas de ML foram utilizados métodos de análise espacial.

A Figura 42 ilustra de forma geral as etapas da pesquisa, que foi uma adaptação do processo SEMMA com as cinco etapas principais de: amostra, exploração, modificação, modelo e avaliação.

Figura 42 – Fluxograma das etapas da pesquisa adaptada da metodologia SEMMA

SEMMA: Amostra (*Sample*), Explorar (*Explore*), Modificar (*Modify*), Modelar (*Model*) e Avaliar (*Assess*)

Fonte: Elaborado pela autora (2025).

Uma breve descrição de cada etapa da Figura 42 é apresentada a seguir:

Etapa 1 – Amostra: Refere-se à parte inicial do procedimento metodológico, que incluiu preliminarmente a identificação e definição dos antecedentes, objetivos e escopo, com base em uma revisão da literatura relevante. Em seguida, realizou-se a identificação dos recursos disponíveis para o desenvolvimento da pesquisa, como a coleta da malha digital do local de estudo (IBGE), para a elaboração de mapas temáticos, e a coleta do conjunto de dados de diferentes origens (SINASC, SIM, CNES, IPEA e SEADE). A descrição mais detalhada do local de estudo, banco de dados e malha digital está apresentada em 3.1 e 3.2.

Etapa 2 – Explorar: Concerne à exploração geral dos conjuntos de dados do SINASC e SIM, iniciando com uma acentuada verificação da qualidade dos dados. A seguir foi realizada uma análise estatística, descritiva e gráfica sobre a natureza dos dados e seu comportamento. O intuito foi identificar padrões, eventos atípicos ou valores ausentes (*missing*) para garantir uma reformatação adequada dos dados e uma integração precisa das informações sobre os nascidos vivos e os óbitos.

A verificação da qualidade dos bancos de dados possibilitou a identificação de valores ausentes e de informações preenchidas de forma incorreta, tanto nas variáveis provenientes do SINASC quanto do SIM. Este processo ajudou a certificar a qualidade dos dados.

Em seguida, dado que a metodologia SEMMA é caracterizada por seu caráter iterativo, a execução desta etapa está relacionada à Etapa 3 de modificação, na qual o banco de dados foi reformatado para permitir a análise exploratória das informações, além de ter sido adaptado conforme as especificações da aplicação de análise espacial, previsão e predição (Figura 40).

Dessa forma, além da verificação da qualidade, nesta etapa foram realizadas principalmente duas atividades: a análise exploratória dos dados e a análise espacial.

- **Análise exploratória de dados:**

A análise exploratória dos dados foi realizada para cada abordagem, permitindo uma avaliação e descrição mais detalhada das informações coletadas e a geração de gráficos de visualização. Entre essas análises, destacam-se a análise exploratória da qualidade dos conjuntos de dados (Apêndice H), a análise exploratória dos nascidos vivos e óbitos entre os municípios da região de estudo (Apêndice I), a análise do padrão temporal dos óbitos ocorridos durante o primeiro ano de vida por faixa etária (Apêndice J), e a análise exploratória das características maternas, paternas e físicas dos nascidos vivos e óbitos na região (Apêndice K). Detalhes sobre o código em Python desta etapa da exploração estão no Apêndice G.

- **Análise espacial:**

Com as informações do novo subconjunto para análise espacial, foram geradas as taxas de mortalidade, avaliadas por faixas etárias. A TMI, TMN e TMPN por mil

nascidos vivos foram calculadas para cada município, usando-se como denominador o número de nascidos vivos, conforme descrito nas Equações 3.1 – 3.3:

$$TMI = \frac{\text{Número de óbitos menores de 1 ano de vida}}{\text{Número total de nascidos vivos no período}} \times 1.000 \quad (3.1)$$

$$TMN = \frac{\text{Número de óbitos de 0 até 27 dias de vida}}{\text{Número total de nascidos vivos no período}} \times 1.000 \quad (3.2)$$

$$TMPN = \frac{\text{Número de óbitos de 28 até 364 dias de vida}}{\text{Número total de nascidos vivos no período}} \times 1.000 \quad (3.3)$$

As variáveis relacionadas às características da mãe e do recém-nascido foram avaliadas por prevalência de casos, que foram calculadas pela razão de seu número em relação ao número total de nascidos vivos em cada município, conforme apresentado na Equação 3.4:

$$\text{Prevalência}_x = \frac{\text{Número de casos de } x}{\text{Número total de nascidos vivos no período}} \times 100 \quad (3.4)$$

sendo x a variável de estudo utilizada no cálculo da prevalência. As variáveis maternas consideradas foram: mães com menos de 20 anos, mães com mais de 34 anos, gravidez dupla, gravidez tripla ou mais, mães com filhos morto, mães com parto normal, parto cesáreo, mães solteiras, viúvas ou separadas, mães com escolaridade menor de 8 anos, pré-natal inadequado (menos de 7 consultas), além de avaliar a idade paterna com menos de 20 anos e mais de 40 anos.

As variáveis analisadas relacionadas ao recém-nascido foram: nascidos vivos prematuros (<37 semanas), baixo peso ao nascer (<2.500 g), anomalia congênita e nascidos vivos com *apgar* menor a 7 no 1º minuto e no 5º minuto de vida.

Além disso, as variáveis relacionadas ao atendimento de saúde foram: taxa de leitos pediátricos por mil nascidos vivos e taxa de médicos por mil habitantes, sendo considerados os pediatras, médicos da estratégia saúde da família e médicos de família e comunidade. A taxa de leitos e médicos foram calculados conforme as Equações 3.5 e 3.6:

$$\text{Taxa}_{\text{leitos}} = \frac{\text{Número de leitos pediátricos}}{\text{Número total de nascidos vivos no período}} \times 1.000 \quad (3.5)$$

$$\text{Taxa}_{\text{médicos}} = \frac{\text{Número de médicos}}{\text{População no período}} \times 1.000 \quad (3.6)$$

Por outro lado, também foram analisadas variáveis relacionadas ao município como o IDH, IVS, Índice Paulista de Responsabilidade Social (IPRS), PIB e PIB *per capita*. Subsequentemente, na análise espacial (Figura 42), analisou-se a distribuição espacial

dos dados coletados, do período 2010 – 2020, sobre mortalidade do nascido vivo, a fim de identificar a existência de dependência espacial, avaliar sua relação com as outras variáveis de análise e identificar municípios com maior risco.

Na análise, as taxas brutas de mortalidade de cada município foram plotadas em mapas temáticos utilizando-se a categoria quartil. Além disso, as taxas brutas foram suavizadas com o método bayesiana empírico (2.4.3), com o intuito de estabilizar a variabilidade dos valores.

Na análise de autocorrelação espacial foram utilizados o índice univariado e bivariado de Moran (2.4.1 e 2.4.2) num panorama global e local, para medir o grau de associação espacial e na RMVPLN e sua relação com as outras variáveis selecionadas, a fim de identificar conglomerados espaciais de risco. Além disso, foram elaborados mapas temáticos utilizando o mapa de Moran para a identificação de conglomerados espaciais, com a divisão por categorias mostrando a associação de acordo com AA, BB, AB e BA. O nível de significância estatística adotado foi de 5%, e seu valor foi calculado pelo teste de significância com Monte-Carlo com 999 permutações aleatórias.

Etapas 3 – Modificar: Trata-se da reformatação dos dados coletados com o intuito de construir novos subconjuntos de dados com as especificações necessárias para a execução de cada uma das análises e modelos propostos na Figura 40.

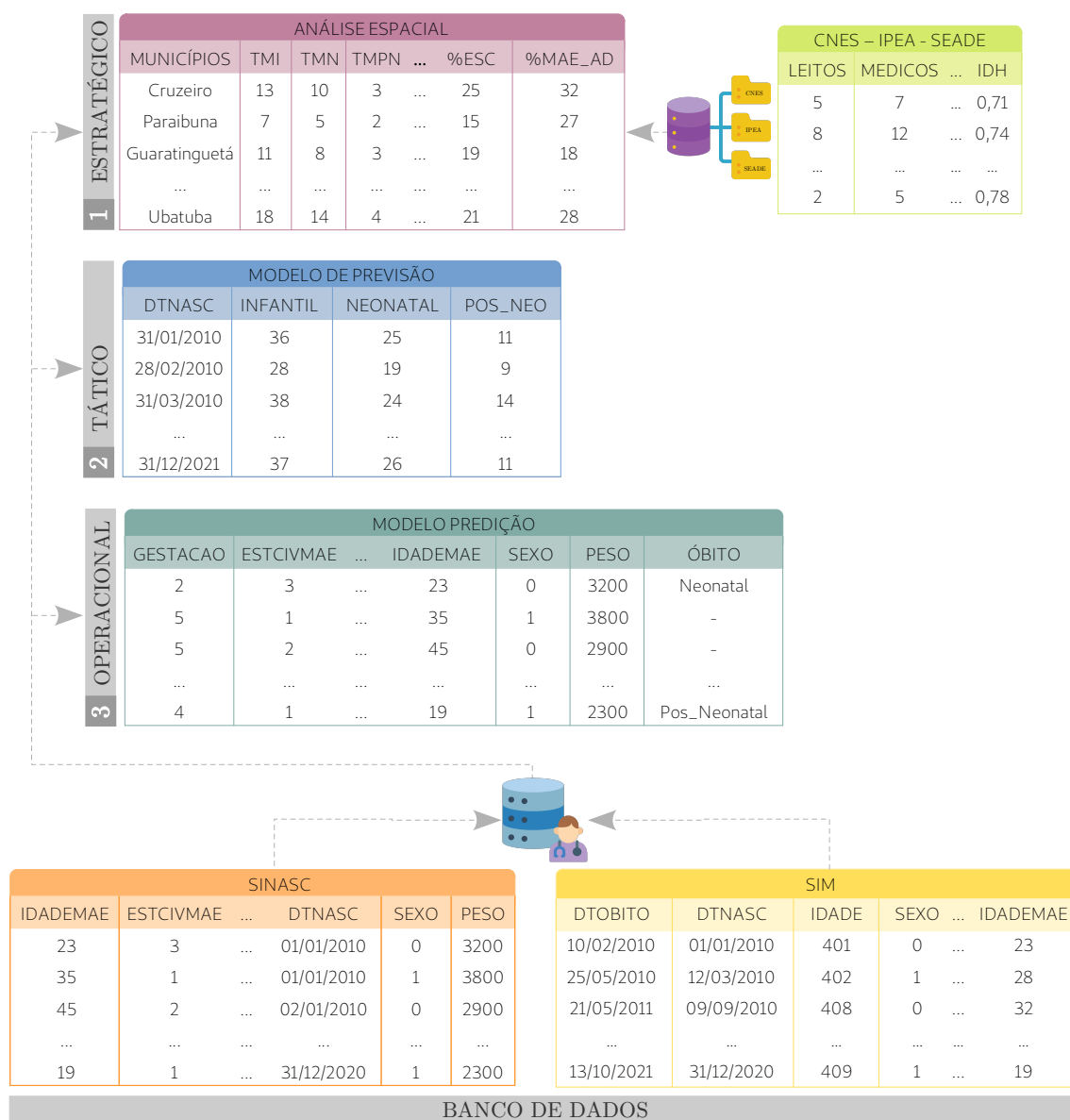
Assim, o processo de reformatação foi iniciado considerando que os dados fornecidos pelos sistemas são anuais. Esses dados, foram consolidados em um formato único e filtrados com base nas informações da RMVPLN, por município e sub-região. Além disso, para a análise exploratória das características, as variáveis CODOCUPMAE, CODANOMAL e CAUSABAS foram decodificadas.

No banco de dados do SIM, realizou-se a rotulação das faixas etária dos óbitos, classificando-os como neonatais (precoce e tardio), pós-neonatais e infantis. O processo de rotulação foi realizado com base na coluna IDADE.

Além disso, foi realizada uma fase de limpeza e padronização superficial das informações, garantindo maior consistência e confiabilidade para as análises subsequentes. A limpeza e padronização dos dados foram conduzidas conforme a descrição dos conjuntos de dados nos Anexos D e E, relativos à declaração de nascido vivo (Anexo B) e à declaração de óbito (Anexo C). Mais detalhes sobre o código em Python desta etapa estão no Apêndice E.

Em seguida, o conjunto de dados corrigido serviu para a reformatação dos novos subconjuntos de dados, de acordo com as especificações da aplicação de análise espacial, previsão e predição, conforme a Figura 43. Mais detalhes sobre o código em Python utilizado para a geração de cada um dos subconjuntos nesta etapa podem ser encontrados no Apêndice F.

Figura 43 – Estruturação dos subconjunto de dados para cada análise



Fonte: Elaborado pela autora (2025).

Assim, o conjunto de dados consolidado foi reformulado com o objetivo de adaptá-lo às características necessárias para as três subetapas ou abordagens propostas neste trabalho. Estas subetapas incluem a análise espacial, a previsão de séries temporais e a predição e identificação de recém-nascidos com risco de óbito (Figura 40).

Cada um dos subconjuntos de dados para cada subetapa é descrito abaixo.

- **Subconjunto da análise espacial:**

Os dados para a análise espacial foram estruturados com informações sobre o total de nascidos vivos, óbitos e o total das informações de seus fatores associados, referentes ao período de 2010 a 2020, para cada um dos municípios da RMVPLN.

O subconjunto de dados inclui variáveis relevantes como o número de óbitos por faixas etárias, variáveis relacionadas às características da mãe e do recém-nascido, além de variáveis relacionadas ao município.

- ***Subconjunto do modelo de previsão:***

Para a formatação dos dados deste subconjunto foram utilizadas exclusivamente as informações do SIM. Os dados para o modelo de previsão foram organizados em formato de séries temporais, considerando a ocorrência mensal de óbitos nas faixas etárias neonatal, pós-neonatal e infantil. O conjunto de dados mensais de óbitos, abrangendo o período de 2010 a 2021, foi empregado nas etapas de treinamento, teste e validação dos modelos. Diferentemente dos demais subconjuntos, nos modelos de previsão com ML, as informações de 2021 foram reservadas especificamente para a validação dos modelos treinados.

- ***Subconjunto do modelo de predição:***

Os dados para o modelo de predição foram estruturados considerando as informações individuais de cada nascido vivo na RMVPLN do período 2010 – 2020.

Neste processo, considerou-se o banco de dados rotulado por meio de uma integração dos conjuntos de dados do SIM e do SINASC, também conhecido como *linkage* de dados. Esse processo foi realizado após a análise de qualidade, pois a presença de dados incorretos pode comprometer os resultados. A partir dessa etapa, foi possível identificar os nascidos vivos que evoluíram para óbito dentro do primeiro ano de vida, como ilustrado na Figura 43.

Na vinculação de bases de dados foi utilizada a ferramenta `recordlinkage` do Python (Christen, 2012). Assim, no processo de *linkage*, foram identificados pares de registros correspondentes entre as informações do SINASC e do SIM. Esse procedimento foi realizado com base em um conjunto de regras utilizando variáveis-chave para garantir uma integração precisa dos dados. Deste modo, as características de DTNASC, PESO, IDADEMAE e GESTACAO permitiram a combinação correta dos registros e a identificação dos nascidos vivos que evoluíram para óbitos.

Etapas 4 – Modelar: Refere-se principalmente ao processo de treinamento dos modelos com técnicas de ML. Parte desta etapa foi compartilhada com a Etapa 3 e 5, conforme a Figura 42.

- ***Subetapa do modelo de previsão:***

No modelo de previsão os dados correspondentes aos 132 meses do período de 2010 a 2020 foram utilizados para treinar e testar os modelos, empregando a técnica de validação cruzada contínua com 2 *folds* (divisões). A técnica de validação cruzada contínua dividiu o conjunto de dados original em dois subconjuntos de treino e teste, com intervalos de tempo fixos e de menor tamanho, respeitando a lógica temporal.

As técnicas de ML ($\mathcal{M}$) usadas neste processo foram: RF, SVM, MLP, XGB, GB, DT, AdaBoost e LGBM, que foram comparadas com as técnicas clássica de ARIMA e SARIMA.

Em relação ao processo de treinamento dos modelos com as técnicas ARIMA e SARIMA, estes foram ajustados utilizando os dados de treinamento previamente normalizados com a técnica Min – Max. Esta técnica permitiu que todas as observações da série temporal tivessem a mesma escala.

Além disso, dado que a estacionariedade da série temporal é uma qualidade importante para as técnicas ARIMA e SARIMA, os dados de treinamento de cada divisão foram analisados a fim de identificar a possível presença de não estacionariedade. A não estacionariedade da série temporal foi verificada por meio dos testes ADF e KPSS, considerando um nível de significância de 5%. Para as séries não estacionárias, foi aplicada uma diferenciação de primeira ordem para torná-las estacionárias.

Seguidamente, após testar a estacionariedade das séries temporais de treinamento de cada faixa etária, foram treinados os modelos ARIMA e SARIMA para cada divisão de dados da validação cruzada contínua. Esses modelos clássicos foram treinados aplicando-se uma busca computacional exaustiva dos melhores parâmetros dos modelos $ARIMA(p, d, q)$ e $SARIMA(p, d, q)(P, D, Q)_s$.

Por outro lado, considerando que, para utilizar as técnicas de ML na previsão de valores futuros em uma série temporal, foi necessário criar variáveis baseadas em valores anteriores da série. Assim, foi realizado um processo de busca dos d eventos passados, conhecido como janela deslizante (lag), para prever o novo evento futuro, conforme foi apresentado na Equação (2.29).

A busca pelo melhor valor de lag foi realizada por meio de um processo iterativo, com o objetivo de encontrar o valor ideal do lag , ou seja, o número de períodos anteriores considerados para o cálculo de uma nova previsão.

Esse processo experimental foi conduzido utilizando validação cruzada contínua, identificando, para cada técnica $mod \in \mathcal{M}$, o valor de lag que minimizou o erro de previsão, conforme apresentado na Equação (3.7):

$$lag_{mod}^* = \arg \min_{lag \in \mathcal{L}} \mathcal{E}(mod, lag) \quad (3.7)$$

onde $\mathcal{E}(mod, lag)$ representa o erro de previsão associado ao modelo da técnica mod para um dado valor de lag , e $\mathcal{L}$ é o conjunto de valores possíveis de lag .

Após a seleção do lag ótimo para cada técnica, o valor mais recorrente entre os lag_{mod}^* foi escolhido como o lag ideal final, utilizando o voto majoritário (moda), conforme apresentado na Equação (3.8):

$$lag^* = \text{moda}(lag_{mod}^*) \quad (3.8)$$

sendo que $\text{moda}(\cdot)$ retorna o valor de lag mais frequente entre os lag_{mod}^* selecionados para cada técnica mod .

Dessa forma, essa abordagem permitiu selecionar o lag ideal considerando o desempenho de diversas técnicas de ML, assegurando que o valor final do lag fosse o mais representativo para a geração da matriz de dados de entrada (x) e saída (y), necessários para o treinamento dos modelos de ML.

Por outro lado, para garantir a eficiência dos modelos treinados, os dados da matriz de entrada (x) foram processados por meio da técnica de normalização Min – Max, enquanto os valores de saída (y) foram transformados utilizando a técnica de padronização.

Além disso, em cada processo de treinamento, os hiperparâmetros de cada modelo foram otimizados utilizando o método de busca aleatória em cada rodada, com o intuito de identificar a melhor configuração de hiperparâmetros para cada técnica testadas.

Por fim, os 12 meses restantes de 2021 foram utilizados para a validação dos modelos finais.

- ***Subetapa do modelos de predição:***

Nesta subetapa, foi realizado um processo de experimentação com o objetivo de desenvolver modelos preditivos, tanto binários quanto multiclasse, para identificar o risco de óbito em diferentes faixas etárias. Especificamente, foram criados dois conjuntos de dados:

- *Modelo binário:* para a predição do risco de óbito infantil (rótulos: Vivo – 0 e Óbito – 1).
- *Modelo multiclasse:* para a predição do risco de óbito neonatal ou pós-neonatal (rótulos: Vivo – 0, Neonatal – 1 e Pós-neonatal – 2).

Dessa forma, além dos rótulos, os bancos de dados foram compostos por variáveis contínuas e categóricas, conforme lista a seguir:

- *Variáveis contínuas consideradas inicialmente:*

* APGAR1	* QTDPARTNOR	* QTDPARTCES
* APGAR5	* QTDGESTANT	* IDADEMAE
* PESO	* QTDFILVIVO	* IDADEPAI
* SEMAGESTAC	* QTDFILMORT	

- *Variáveis categóricas consideradas inicialmente:*

* RACACORMAE	* GRAVIDEZ	* CONSULTAS
* RACACOR	* SEXO	* ESTCIVMAE
* STCESPARTO	* GESTACAO	* PARIDADE
* PARTO	* IDANOMAL	* CODOCUPMAE
* TPAPRESENT	* LOCNASC	* ESCMAE
* STTRABPART	* TPNASCASSI	* SUB-REGIAO

A variável SUB-REGIAO está relacionada à dimensão espacial, fornecendo informações referentes aos municípios de residência das mães. Cada município, pertencente a uma das cinco sub-regiões da RMVPLN, é atribuído a uma categoria numérica correspondente à sua respectiva sub-região.

Em cada conjunto de dados foi realizada uma etapa preliminar de seleção de características, que além de considerar a qualidade da informação, também se baseou nos resultados de uma análise de correlação de Pearson (r) para as variáveis numéricas, e uma análise de associação com o coeficiente V de Cramer (v_c) para as variáveis categóricas. Esse processo permitiu selecionar as características relacionadas ao recém-nascido e à mãe, que foram utilizadas no desenvolvimento dos modelos.

Em seguida, após a seleção preliminar das variáveis que constituíram o Conjunto de Dados Completo (CDC), procedeu-se à divisão dos dados. Para o processo de treinamento e teste das técnicas de ML, foram considerados os dados dos nascidos vivos do período de 2010 a 2019, enquanto as informações de 2020 foram reservadas exclusivamente para o processo de validação dos modelos finais.

Antes do treinamento dos modelos de predição para a identificação de risco, foi realizada uma análise da importância das características, com o intuito de identificar as variáveis com maior influência nos modelos preditivos. Para essa análise, utilizou-se o método embutido com as técnicas RF, que permite capturar interações complexas entre variáveis, e LASSO, com seu componente de penalização $L1$, devido à sua capacidade de selecionar variáveis relevantes por meio da penalização de coeficientes. O processo de seleção de variáveis iniciou com uma primeira filtragem das características, com base em um critério de seleção tanto para LASSO (CS_{LASSO}) quanto para RF (CS_{RF}). A segunda filtragem de variáveis foi baseada na combinação dessas duas abordagens, o que permitiu a identificação de variáveis importantes tanto de forma linear quanto não linear. Este processo foi selecionado pois o uso de ambas as técnicas pode reduzir o *overfitting* durante o treinamento dos modelos.

Esse processo permitiu a remoção de variáveis pouco relevantes, possibilitando a comparação entre as métricas dos modelos treinados (binários e multiclasse) com o Conjunto de Dados Simplificado (CDS) e aqueles treinados com o CDC, sem a remoção das variáveis pouco relevantes.

Portanto, fase de treino e teste foi realizada em quatro blocos de experimentação, considerando os conjuntos de dados simplificado e completo, para os modelos binários e multiclasse. Este processo foi iterativo, utilizando a técnica de validação cruzada 5-fold (5 dobras), onde, em cada dobra, o pré-processamento de dados foi realizado separadamente para os conjuntos de treinamento e teste.

Para o conjunto de treinamento, foram aplicados ajustes como imputação de dados ausentes, remoção de informações irrelevantes e transformações de variáveis, permitindo o ajuste dos parâmetros com base nos dados de treino. Esses parâmetros, obtidos durante o treinamento, foram aplicados ao conjunto de teste, garantindo consistência no processo de transformação, além de evitar *overfitting* e vazamento de dados (*data leakage*), com modelos mais generalizáveis e robustos.

A Figura 44 ilustra os procedimentos empregados no pré-processamento para aprimorar a qualidade dos dados. Esses procedimentos englobam etapas fundamentais de verificação, imputação, exclusão e transformação, aplicadas de forma específica às variáveis categóricas e numéricas, conforme detalhado a seguir.

- *Verificação*: Os valores de cada variável categórica do banco de dados foram verificados com base nas diferentes categorias das variáveis da Declaração de Nascido Vivo, conforme os Anexos B e D.

Em relação às variáveis numéricas de QTDFILVIVO e QTDFILMORT, foi realizada uma verificação por meio de uma sequência de regras de processamento, com base na relação $QTDFILMORT = QTDFILVIVO + QTDFILMORT$.

- *Imputação*: Nas variáveis categóricas, os valores diferentes das categorias da Declaração de Nascido Vivo (0, 1 e 99), assim como os valores ausentes (NaN), foram substituídos pelo valor 9, correspondente a "Ignorado".

Nas variáveis numéricas, os valores ausentes (NaN) de IDADEMAE, QTDFILVIVO e QTDFILMORT foram imputados pela média baseada no conjunto de treinamento.

Durante o pré-processamento do conjunto de teste, os valores ausentes foram imputados utilizando os parâmetros (média ou mediana) previamente calculados no conjunto de treinamento. Essa abordagem garantiu consistência no tratamento dos dados entre os dois conjuntos, prevenindo vazamento de dados (*data leakage*) e mantendo a integridade da avaliação dos modelos.

As variáveis não mencionadas neste procedimento não foram tratadas pelo processo de imputação.

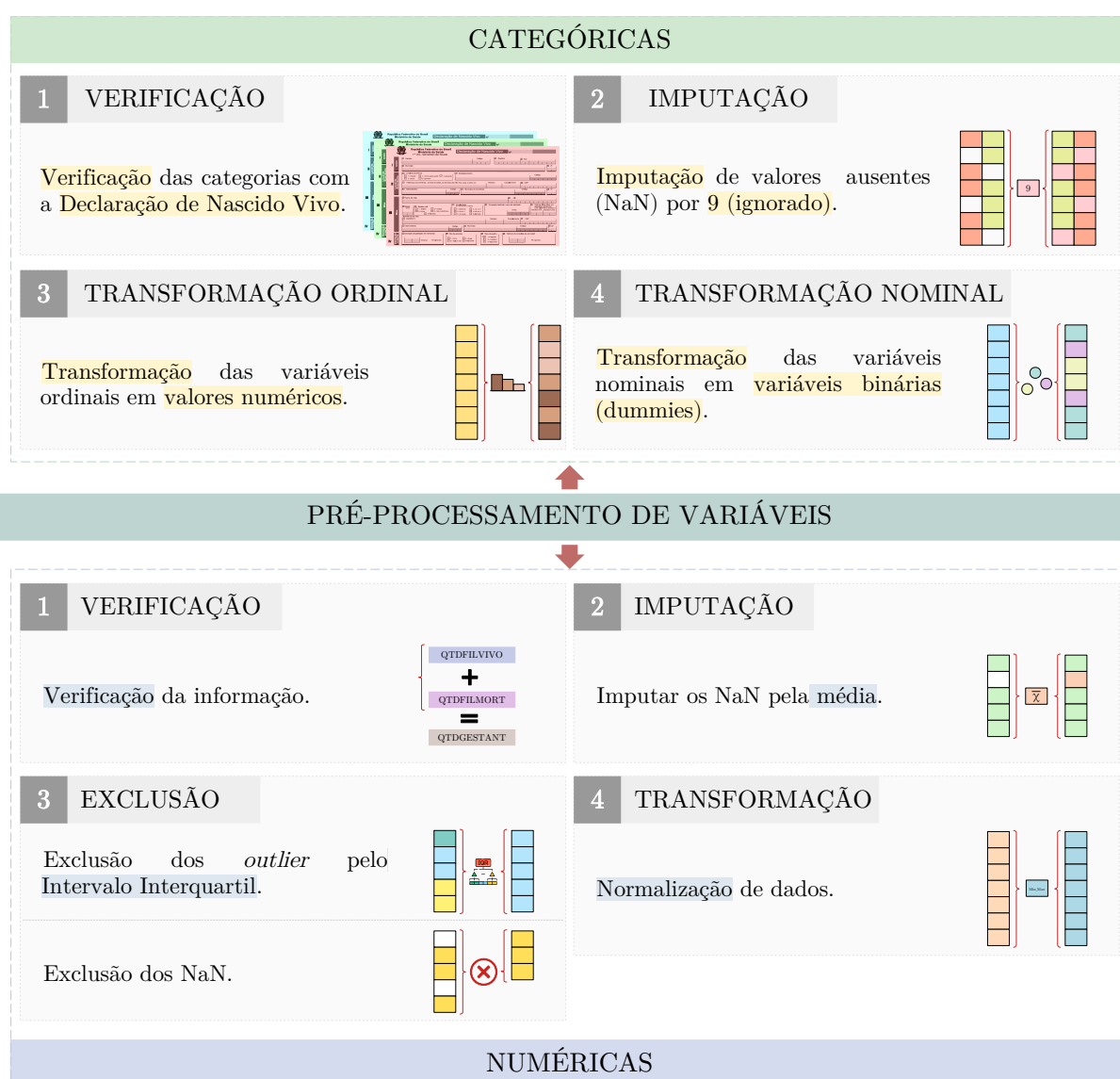
- *Exclusão*: Em relação às variáveis QTDFILVIVO e QTDFILMORT, a identificação dos *outliers* nas foi realizada com base nos percentis de cada variável. Valores que estavam fora dos percentis inferior e superior também foram considerados *outliers* e excluídos do banco de dados de treinamento.

Por outro lado, as linhas com valores ausentes (NaN) das variáveis PESO, APGAR1 e APGAR5 foram excluídos do banco de dados de treinamento.

No conjunto de teste, entretanto, os *outliers* foram mantidos, com o intuito de preservar a integridade dos dados e permitir uma avaliação do modelo em condições que reflitam cenários reais e diversos. Essa abordagem contribui para o desenvolvimento de modelos de ML mais robustos e generalizáveis, com a capacidade de lidar com valores extremos em dados futuros.

As variáveis que não foram indicadas neste procedimento não passaram por nenhum tipo de tratamento ou exclusão.

Figura 44 – Procedimentos aplicados no pré-processamento das variáveis



Fonte: Elaborado pela autora (2025).

– *Transformação*: Para assegurar uma codificação adequada das variáveis, diferentes abordagens foram aplicadas conforme o tipo de variável (nominal, ordinal

ou numérica).

As variáveis ordinais, como ESCMAE, GESTACAO e CONSULTAS, foram convertidas em valores numéricos inteiros, preservando a ordem hierárquica de suas categorias. Por outro lado, as variáveis nominais, como LOCNASC, IDANOMAL, TPAPRESENT, STTRABPART, SEXO, GRAVIDEZ, STCESPARTO, PARTO, SUB-REGIAO, ESTCIVMAE e RACACORMAE, foram transformadas por meio da criação de variáveis binárias (*dummies*) para cada categoria,

Para as variáveis numéricas, como IDADEMAE, APGAR1, APGAR5, PESO, QTDFILMORT e QTDFILVIVO, foi utilizada a técnica de normalização Min-Max, que ajustou os valores para um intervalo comum de $[0, 1]$. Essa padronização garantiu uma escala consistente entre as variáveis numéricas, facilitando o aprendizado dos modelos.

Durante este processo, os parâmetros necessários para as transformações foram aprendidos exclusivamente a partir do conjunto de treinamento e, em seguida, aplicados aos dados de teste ou validação. Essa abordagem assegurou a consistência nas transformações e manteve a integridade da avaliação por meio da validação cruzada.

Adicionalmente, em cada iteração da validação cruzada, foi implementado um processo de balanceamento das classes por meio da técnica de subamostragem, com o objetivo de reduzir a classe majoritária. Essa abordagem visou evitar que os modelos treinados fossem excessivamente influenciados pela classe dominante, além de minimizar o risco de *overfitting*. O balanceamento das classes contribuiu para a redução de vieses nos resultados, garantindo uma avaliação mais equitativa do desempenho dos modelos durante o processo de avaliação.

Em relação às técnicas de *ML* utilizadas neste processo foram: RLog, DT, SVM, MLP, RF, XGB, LGBM, GB e Bagging. Além disso, os hiperparâmetros dos modelos foram otimizados em cada rodada por meio da otimização bayesiana, visando identificar os melhores valores para maximizar o desempenho dos modelos em cada iteração de treinamento.

Etapas 5 – Avaliar: Trata-se de um processo iterativo, que foi conduzido em conjunto com a Etapa 4 para o desenvolvimento dos modelos. Esta etapa tem como objetivo a avaliação, seleção e validação dos modelos de ML desenvolvidos, utilizando diferentes métricas estatísticas de desempenho.

No caso da subetapa dedicada aos modelos de previsão, foram adotadas as seguintes métricas de erro para modelos de regressão: MAE, MSE, RMSE, MAPE, SMAPE e MPD. Além disso, foi aplicado o teste-t pareado para comparar as médias das previsões dos modelos com os valores reais, utilizando um intervalo de confiança de 95% (Montgomery;

Runger, 2020). Esse procedimento buscou identificar e selecionar o modelo com o melhor desempenho preditivo.

Por outro lado, em relação à subetapa do modelo de predição, foram utilizadas métricas de erro para modelos de classificação, como a matriz de confusão e suas métricas associadas: acurácia, precisão, revocação (sensibilidade), F1-Score, AUC e a curva ROC. Além disso, foi aplicado o teste estatístico de McNemar para comparar os modelos treinados em pares (Alpaydin, 2021), com o objetivo de analisar a frequência de seus erros e acertos e, assim, selecionar o modelo final mais robusto e adequado para a aplicação proposta.

3.4 CONSIDERAÇÕES ÉTICAS

Esta pesquisa foi desenvolvida com dados disponibilizados pelos domínios públicos do DATASUS, SEADE, IPEA e IBGE, portanto não foi necessária a submissão do projeto ao Comitê de Ética em Pesquisa, conforme estabelecido na Resolução nº 510/16 do Conselho Nacional de Saúde, de 07 de abril de 2016.

4 RESULTADOS E DISCUSSÃO

Este Capítulo apresenta a descrição e discussão dos resultados obtidos na análise espacial da mortalidade dos nascidos vivos e no desenvolvimento dos modelos de previsão e predição com ML.

Especificamente, o Capítulo está dividido em três seções principais. A primeira, 4.1, apresenta os resultados da análise espacial das taxas de mortalidade neonatal, pós-neonatal e infantil na RMVPLN, bem como sua associação com fatores determinantes. A Seção 4.2 descreve os resultados do desenvolvimento e da aplicação de técnicas de ML para prever o número mensal de óbitos na área analisada. Já a Seção 4.3 apresenta os resultados da aplicação de ML no desenvolvimento de um modelo preditivo para identificar recém-nascidos com risco de óbito.

A descrição da análise da qualidade dos dados, assim como a análise exploratória da distribuição espacial, do padrão temporal e das características dos dados, encontram-se nos Apêndices H, I, J e K, respectivamente.

4.1 ANÁLISE ESPACIAL DAS TAXAS DE MORTALIDADE

Na análise espacial foram avaliadas as taxas de mortalidade e sua relação com as outras variáveis de análise. As taxas de mortalidade neonatal, pós-neonatal e infantil foram calculadas utilizando as Equações 3.1 – 3.3. As variáveis relacionadas às características maternas e do recém-nascido foram calculadas pela Equação 3.4, e as características relacionadas ao atendimento de saúde foram calculadas com as Equações 3.5 e 3.6.

4.1.1 Mapeamento das taxas de mortalidade e determinantes

A média da TMI na RMVPLN no período 2010-2020 foi de 11,96 por mil nascidos vivos, além disso, no mesmo período, a TMN foi de 8,61 por mil nascidos vivos, e a TMPN foi de 3,35 por mil nascidos vivos.

A Tabela 3 apresenta os valores mínimos, máximos e valor médio das variáveis analisadas espacialmente dos 39 municípios da região, no período 2010-2020.

Tabela 3 – Análise descritiva das taxas de mortalidade e das outras variáveis de análise nos municípios da RMVPLN, no período 2010-2020

Variáveis	Mínimo	Máximo	(continua)
			Média
Mortalidade			
TMI ^a	6,24	20,58	13,12
TMN ^b	5,75	15,74	9,50
TMPN ^c	0,00	6,22	3,62

(a) Taxa de mortalidade infantil, (b) Taxa de mortalidade neonatal, (c) Taxa de mortalidade pós-neonatal.
Fonte: Elaborado pela autora (2025).

Tabela 3 – Análise descritiva das taxas de mortalidade e das outras variáveis de análise nos municípios da RMVPLN, no período 2010-2020

	(conclusão)		
Variáveis	Mínimo	Máximo	Média
Caraterísticas dos pais			
Mães com menos de 20 anos (%)	11,42	21,00	16,19
Mães com mais de 34 anos (%)	8,57	20,90	13,90
Gravidez dupla (%)	0,56	2,70	2,00
Gravidez tripla ou mais (%)	0,00	0,20	0,03
Mães com filhos morto (%)	11,55	19,90	15,05
Mães com parto normal (%)	15,23	59,91	36,95
Mães com parto cesáreo (%)	40,09	84,45	62,95
Mães solteiras, viúvas ou separadas (%)	32,47	63,08	48,26
Mães com baixa escolaridade ^d (%)	6,79	27,18	16,42
Pré-natal inadequado ^e (%)	11,52	37,50	25,52
Pais com menos de 20 anos (%)	0,18	5,33	2,21
Pais com mais de 40 anos (%)	29,42	94,23	61,25
Caraterísticas do recém-nascido			
Nascidos vivos prematuros ^f (%)	7,91	15,92	10,33
Baixo peso ao nascer ^g (%)	4,19	10,59	8,67
Anomalia congênita (%)	0,19	2,37	0,84
Nascidos vivos com <i>apgar</i> < 7 no 1 o minuto (%)	3,75	8,98	5,73
Nascidos vivos com <i>apgar</i> < 7 no 5 o minuto (%)	0,56	2,13	1,06
Atenção à saúde			
Taxa de leitos pediátricos	38,74	240,84	83,51
Taxa de médicos	0,0	0,57	0,16
Caraterísticas do município			
IDH ^h	0,66	0,81	0,74
IVS ⁱ	0,18	0,36	0,25
IPRS ^j	1,00	5,00	3,33
PIB ^k	27.254	33.991.526	2.539.468,56
PIB <i>per capita</i>	9.059	195.597	26.695,15

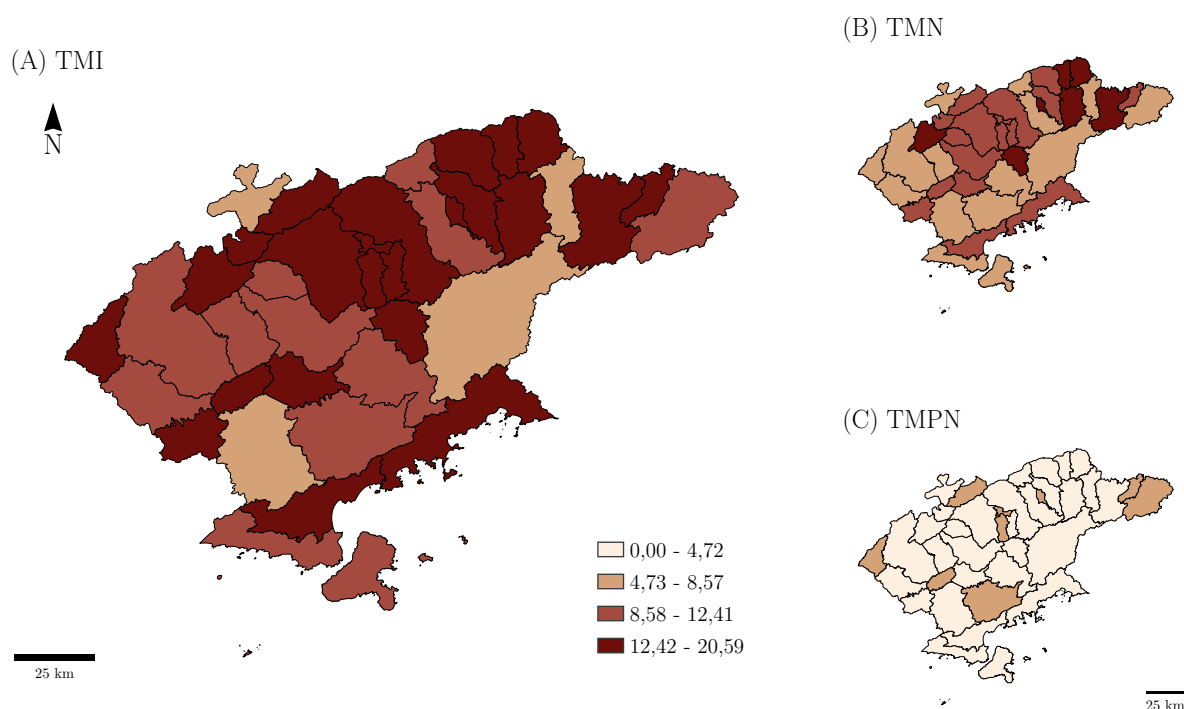
(d) Menos de 8 anos, (e) Menos de 7 consultas, (f) Idade gestacional < 37 semanas, (g) Peso < 2.500 g,

(h) Índice de desenvolvimento humano, (i) IVS: Índice de vulnerabilidade social, (j) IPRS: Índice paulista de responsabilidade social e (k) PIB: Produto interno bruto.

Fonte: Elaborado pela autora (2025).

Na Tabela 3 observa-se que as taxas de mortalidade apresentaram uma variação diferente quando avaliados por faixa etária. Assim, a distribuição espacial das taxas de mortalidade municipal foram apresentadas na Figura 45, exibida no critério de quartis. Os valores consolidados da TMI (Figura 45A), revelaram que nenhum dos municípios apresentaram taxas abaixo de 4,73, e 10,26% (4) estiveram no intervalo 4,73 e 8,57. No intervalo 8,58 – 12,41 estiveram 30,77% (12) dos municípios, e os demais 58,97% (23) dos municípios apresentaram taxas acima de 12,41.

Figura 45 – Mapa temático da distribuição da (A) Taxa de Mortalidade Infantil (TMI), (B) Taxa de Mortalidade Neonatal (TMN), (C) Taxa de Mortalidade Pós-neonatal (TMPN), período 2010 – 2020



Fonte: Elaborado pela autora (2025).

Da mesma forma que na TMI, a análise da distribuição da TMN (Figura 45B) revelou que nenhum município apresentou taxa de mortalidade abaixo de 4,73. Já 38,46% (15) e 43,59% (17) dos municípios apresentaram taxas de mortalidade no intervalo de 4,73 – 8,57 e 8,58 – 12,41, respectivamente, bem como 17,95% (7) dos municípios restantes apresentaram taxas acima de 12,42.

A distribuição da TMPN (Figura 45C) apresentou um comportamento diferente, sendo que 76,92% (30) dos municípios apresentaram taxas abaixo de 4,73 e os demais 23,08% (9) dos municípios apresentaram taxas no intervalo de 4,73 - 8,57. Nenhum município apresentou taxa de mortalidade acima de 8,57.

Os achados nesta análise indicaram que a redução das taxas de mortalidade na RMVPLN tem se refletido principalmente em diminuição da mortalidade pós-neonatal, dado que o número de óbitos neonatais é maior que o número de óbitos pós-neonatais. Além disso, tais resultados não são muito diferentes daqueles observados em outros estudos no Brasil (Paixao *et al.*, 2021; Bernardino *et al.*, 2022).

4.1.2 Autocorrelação espacial univariado e bivariado

A identificação de aglomerados espaciais das condições de saúde dos nascidos vivos tem sido objeto de estudo nos últimos anos, e a mortalidade do recém-nascido não tem sido exceção

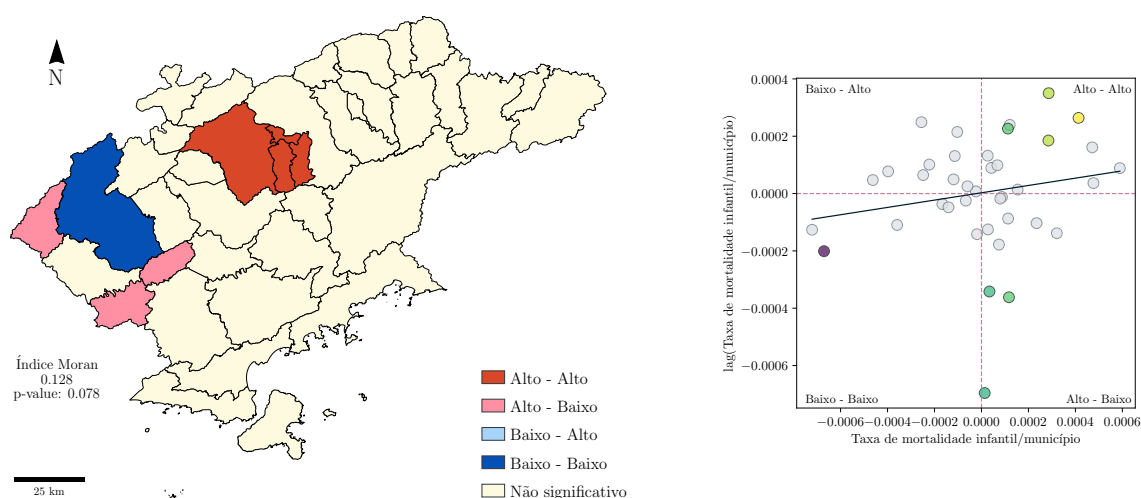
(Nascimento *et al.*, 2007; Gonçalves; Costa; Braga, 2011; Almeida; Gomes; Nascimento, 2014; Nascimento; Almeida; Gomes, 2014).

Deste modo, na análise de autocorrelação espacial foi utilizado o método bayesiana empírico (Equação 2.17), para suavizar as taxas de mortalidade, a fim de reduzir a instabilidade nos dados analisados.

Na análise de autocorrelação espacial com o índice de Moran global, identificou-se a presença de autocorrelação espacial estatisticamente significativa na taxa suavizada da mortalidade neonatal ($I_U = 0,238$; p-valor = 0,012). Por outro lado, na mortalidade infantil ($I_U = 0,128$; p-valor = 0,078) e pós-neonatal ($I_U = -0,063$; p-valor = 0,381) não houve significância estatística.

Na avaliação de conglomerados espaciais, com o índice de Moran local, identificaram-se agrupamentos de municípios estatisticamente significativos. Na TMI (Figura 46) 8 (20,51%) municípios apresentaram valores I_{U_i} significativos, observando-se aglomerados de risco de taxas altas (AA) na região Norte Central da RMVPLN, de modo que dentro deste agrupamento estão os municípios de Pindamonhangaba, Roseira, Aparecida e Potim.

Figura 46 – Mapa Moran e gráfico de dispersão da suavização bayesiana da Taxa de Mortalidade Infantil (TMI), período 2010 – 2020



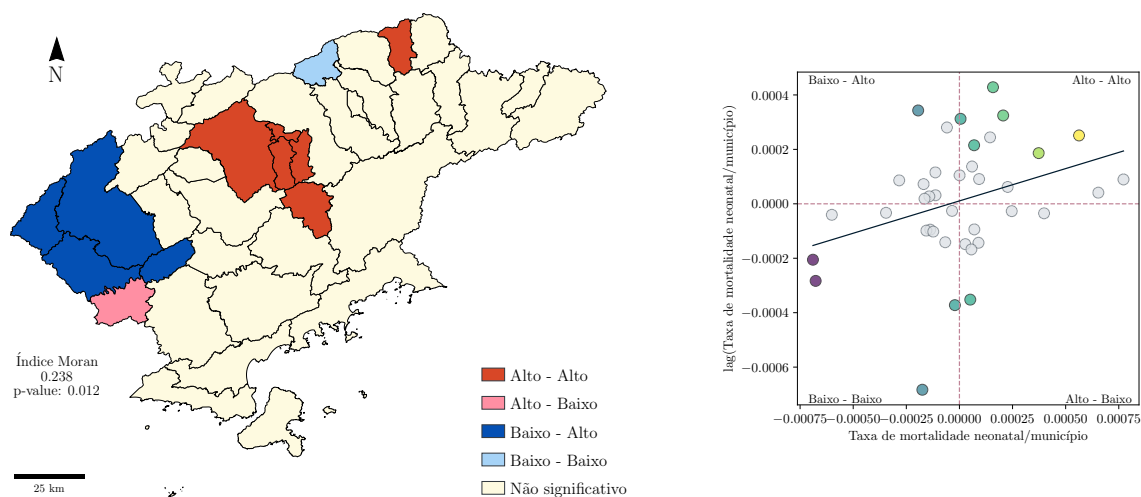
Fonte: Elaborado pela autora (2025).

Além disso, ainda na Figura 46 identificou-se um pequeno aglomerado de municípios (Santa Branca e Paraibuna) com alta taxa de mortalidade cercados por municípios vizinhos com baixos valores (AB) na região Oeste. Já na região Noroeste os municípios de Igaratá e São José dos Campos foram apresentados como aglomerados AB e BA por serem circundados por municípios com baixas e altas taxas de mortalidade, respectivamente. No entanto, esses municípios observados como AB e BA não podem ser considerados aglomerados.

Na TMN (Figura 47) 13 (33,33%) municípios foram significativos, formando-se aglomerados espaciais de municípios de risco com altos valores de mortalidade cercados por municípios com também altos valores de mortalidade infantil (AA). Entre eles, identificaram-se os municípios

de Campos do Jordão, Pindamonhangaba, Roseira, Aparecida, Potim e Lagoinha, localizado na região Norte Central.

Figura 47 – Mapa Moran e gráfico de dispersão da suavização da Taxa de Mortalidade Neonatal (TMN), período 2010 – 2020

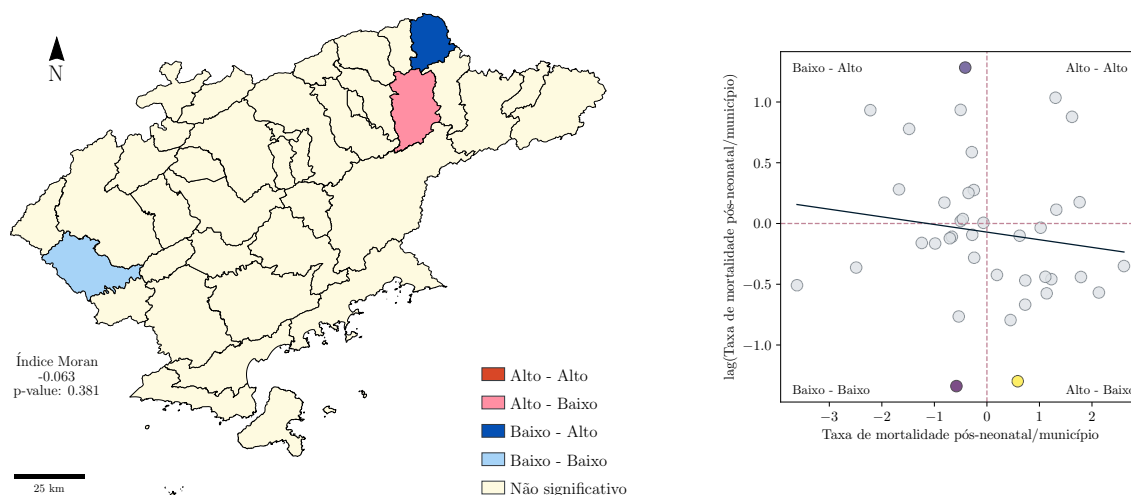


Fonte: Elaborado pela autora (2025).

Também, foi identificado um grupo de municípios com baixos valores de mortalidade vizinhos de municípios com baixos valores na TMN (BB), na região Noroeste (Igaratá, São José dos Campos, Jambuí e Jacareí). Por outro lado, Lavrinhas, Santa Branca e Piquete, apesar de serem municípios significativos, não podem ser considerados aglomerações por serem áreas isoladas.

Em relação à TMPN (Figura 48), do mesmo modo que na TMI e TMN, os municípios de Silveiras, Queluz e Jacareí não podem ser considerados aglomerações.

Figura 48 – Mapa Moran e gráfico de dispersão da suavização bayesiana da Taxa de Mortalidade Pós-neonatal (TMPN), período 2010 – 2020



Fonte: Elaborado pela autora (2025).

Os aglomerados identificados na análise permitiram a identificação de municípios com risco agravado nas taxas de mortalidade (Filho *et al.*, 2007), possibilitando a identificação das áreas com necessidade de intervenção, de acordo com a faixa etária do óbito. Além disso, ao se comparar estes resultados com os encontrados na literatura (Nascimento *et al.*, 2007), observou-se que embora a TMN tenha diminuído na região, há municípios que ainda continuam apresentando alta mortalidade neonatal, como os municípios de Silveira e Lagoinha.

Por outro lado, na análise bivariada global identificaram-se diferentes graus de associação espacial estatisticamente significativa entre as diferentes variáveis analisadas com as taxas de mortalidade suavizadas, conforme apresentado na Tabela 4.

Tabela 4 – Índice de Moran bivariado entre a mortalidade infantil, neonatal e pós-neonatal com as variáveis selecionadas da RMVPLN, no período 2010-2020

Variáveis	TMI ^a	(p-valor)	TMN ^b	(p-valor)	TMPN ^c	(p-valor)
Caraterísticas dos pais						
Mães com menos de 20 anos (%)	-0,037	(0,344*)	0,017	(0,432*)	-0,125	(0,078*)
Mães com mais de 34 anos (%)	-0,192	(0,017)	-0,218	(0,012)	0,122	(0,085*)
Gravidez dupla (%)	-0,034	(0,357*)	-0,089	(0,194*)	0,245	(0,003)
Gravidez tripla ou mais (%)	-0,044	(0,312*)	-0,04	(0,336*)	0,058	(0,255*)
Mães com filhos morto (%)	-0,120	(0,097*)	-0,141	(0,083*)	0,03	(0,362*)
Mães com parto normal (%)	-0,283	(0,000)	-0,337	(0,000)	-0,044	(0,308*)
Mães com parto cesáreo (%)	0,283	(0,001)	0,337	(0,000)	0,043	(0,319*)
Mães solteiras (%)	0,134	(0,076*)	0,221	(0,011)	-0,116	(0,096*)
Mães com baixa escolaridade (%)	-0,136	(0,069*)	-0,14	(0,082*)	-0,125	(0,079*)
Pré-natal inadequado (%)	-0,003	(0,494*)	0,046	(0,324*)	-0,098	(0,137*)
Pais com menos de 20 anos (%)	0,174	(0,029)	0,173	(0,039)	-0,016	(0,425*)
Pais com mais de 40 anos (%)	-0,237	(0,002)	-0,231	(0,008)	0,039	(0,335*)
Caraterísticas do recém-nascido						
Nascidos vivos prematuros (%)	-0,194	(0,019)	-0,220	(0,016)	0,067	(0,228*)
Baixo peso ao nascer (%)	-0,104	(0,138*)	-0,112	(0,135*)	0,042	(0,314*)
Anomalia congênita (%)	0,009	(0,475*)	0,016	(0,454*)	-0,032	(0,360*)
Apgar < 7 no 1 o minuto (%)	-0,027	(0,381*)	-0,036	(0,354*)	-0,025	(0,390*)
Apgar < 7 no 5 o minuto (%)	-0,128	(0,086*)	-0,111	(0,134*)	-0,052	(0,277*)
Atenção à saúde						
Taxa de leitos pediátricos	-0,071	(0,210*)	-0,084	(0,205*)	0,107	(0,107*)
Taxa de médicos	0,041	(0,341*)	0,038	(0,356*)	0,018	(0,415*)
Caraterísticas do município						
IDH ^d	0,012	(0,443*)	-0,039	(0,348*)	0,165	(0,034)
IVS ^e	-0,038	(0,340*)	-0,002	(0,493*)	-0,134	(0,064*)
IPRS ^f	0,119	(0,096*)	0,214	(0,016)	-0,104	(0,124*)
PIB ^g	-0,117	(0,116*)	-0,212	(0,024)	0,213	(0,011)
PIB <i>per capita</i>	-0,05	(0,271*)	-0,126	(0,110*)	0,072	(0,194*)

* Não significativo (p-valor > 0,05), (a) Taxa de mortalidade infantil, (b) Taxa de mortalidade neonatal, (c) Taxa de mortalidade pós-neonatal, (d) Índice de desenvolvimento humano, (e) IVS: Índice de vulnerabilidade social, (f) IPRS: Índice paulista de responsabilidade social e (g) PIB: Produto interno bruto.

Fonte: Elaborado pela autora (2025).

Em relação às variáveis analisada na Tabela 4, a TMI associou-se de forma positiva estatisticamente significativa com o parto cesáreo ($I_{U_i} = 0,283$) e a idade paterna menor a 20 anos

($I_{U_i} = 0, 174$). Além disso, constatou-se uma associação significativa negativa com mães com mais de 34 anos ($I_{U_i} = -0, 192$), parto normal ($I_{U_i} = -0, 283$), pais com mais de 40 anos ($I_{U_i} = -0, 237$) e prematuridade ($I_{U_i} = -0, 194$).

A TMN associou-se significativamente de forma positiva com o parto cesáreo ($I_{U_i} = 0, 337$), mães solteiras ($I_{U_i} = 0, 221$), pais com menos de 20 anos ($I_{U_i} = 0, 173$) e IPRS ($I_{U_i} = 0, 214$), além de associar-se negativamente com mães com mais de 34 anos ($I_{U_i} = -0, 218$), parto normal ($I_{U_i} = -0, 337$), pais com mais de 40 anos ($I_{U_i} = -0, 231$), prematuridade ($I_{U_i} = -0, 220$) e PIB ($I_{U_i} = -0, 212$).

Com relação à TMPN, identificou-se uma associação positiva estatisticamente significativa com a gravidez dupla ($I_{U_i} = 0, 245$), IDH ($I_{U_i} = 0, 165$) e PIB ($I_{U_i} = -0, 213$). Por outro lado, as outras variáveis não indicadas da Tabela 4 não mostraram associação estatisticamente significativa com a TMI, TMN e TMPN.

Deste modo, a análise de associação global ajudou a identificar as possíveis relações entre as taxas de mortalidade com as características analisadas. Os resultados da análise local foram apresentados no Apêndice M, dado que, foram encontradas aglomerações espaciais estatisticamente significativas na avaliação de autocorrelação bivariada mesmo com ausência de significância no índice global bivariado (Anselin, 1995; Almeida, 2012).

Neste cenário, em muitos casos, os falecimentos dos nascidos vivos podem ser eventos evitáveis por estar associado a diversas características (Bernardino *et al.*, 2022), como os analisados na Tabela 4.

Assim, a maternidade em idade jovem e idade mais avançada estão associadas com a vulnerabilidade do recém-nascido e risco de óbito infantil, pois tem uma ligação com fatores socioeconômicos, situação conjugal, ocupação, etc (Ferraz; Neves, 2011; Gaiva; Fujimori; Sato, 2016; Costa-Nobre *et al.*, 2021; Rodríguez *et al.*, 2023). Por outro lado, a idade paterna é uma característica frequentemente avaliada com a mortalidade infantil, e estudos indicam que está associada com a prematuridade, baixo peso ao nascer e/ou problemas de morbidade (Khandwala *et al.*, 2018).

As gestações múltiplas apresentam um alto risco de mortalidade infantil, e o risco é maior nas primeiras semanas de vida (Santos *et al.*, 2012; Mathews; Macdorman; Thoma, 2015; MS-SVE, 2019; Tesema *et al.*, 2022). Além disso, a gestações múltiplas também estão relacionadas com a idade materna, pois está associada ao envelhecimento do sistema reprodutivo nas mulheres acima dos 34 anos e à imaturidade emocional e reprodutiva nas mães adolescentes (Santos *et al.*, 2012; Gaiva; Fujimori; Sato, 2016).

Igualmente, o nível de escolaridade da mãe pode ser considerado um reflexo da estabilidade financeira e do nível socioeconômico da família, além do fato de que a baixa renda familiar também está associada com o desenvolvimento do feto durante a gestação, como a prematuridade e/ou baixo peso ao nascer (Almeida *et al.*, 2008; Capelli *et al.*, 2014; Fonseca *et al.*, 2017; Vilanova *et al.*, 2019; Rodríguez *et al.*, 2023; Rodríguez *et al.*, 2024). Assim, a prematuridade e o baixo peso ao nascer estão fortemente relacionados com a mortalidade, dado que, quanto

menor o tamanho e o peso do recém-nascido e quanto menor o tempo de gestação, maior o risco de óbito (Almeida *et al.*, 2008; Ferraz; Neves, 2011; Vilanova *et al.*, 2019; Rodríguez *et al.*, 2023).

No entanto, complicações como baixo peso ao nascer ou anomalias congênitas, podem ser eventos preveníveis ou tratados por meio de cuidados pré-natais adequados (Vilanova *et al.*, 2019). O acompanhamento adequado do pré-natal pode ajudar positivamente nos casos de gestação de risco, e na detecção precoce dos agravos durante a gestação (Ferraz; Neves, 2011; Vilanova *et al.*, 2019; Tesema *et al.*, 2022). Portanto, o aumento do número de consultas de pré-natal durante a gestação resultaria na possível diminuição da mortalidade neonatal e infantil.

Por sua vez, a anomalia congênita é considerada um forte fator de risco para mortalidade neonatal, pós-neonatal e infantil (Kurinczuk *et al.*, 2010; Kropiwiec; Franco; Amaral, 2017; MS-SVE, 2019), não obstante, como inconveniente, muitas das anomalias são detectadas após o nascimento (MS-SVE, 2019). Além disso, os recém-nascidos de mães com histórico de perdas fetais ou abortos em gestações anteriores apresentam um maior risco óbito, em comparação dos nascidos-vivos de mães que nunca tiveram perdas (MS-SVE, 2019).

Do mesmo modo, recém-nascidos com *apgar* inferior a 7 no 1º e 5º minuto também apresentam alto risco de óbito, dado que o *apgar* é um indicador da saúde do nascido vivo (Santos *et al.*, 2012; Vilanova *et al.*, 2019), e o risco é maior para recém-nascidos vulneráveis com *apgar* baixo (MS-SVE, 2019).

Em relação ao tipo de parto, existem estudos onde a cesárea está diretamente relacionada com a mortalidade infantil (Tesema *et al.*, 2022), porém, em outros estudos encontra-se associada como fator de proteção para a mortalidade (Santos *et al.*, 2012; Vilanova *et al.*, 2019), ainda mais quando se trata da mortalidade neonatal em recém-nascidos prematuros ou com baixo peso ao nascer (Vilanova *et al.*, 2019). Assim, fetos com crescimento restrito podem apresentar maior risco de sofrer complicações após o parto normal (vaginal), pois passam por um processo de estresse hipóxico crônico (Vilanova *et al.*, 2019).

No entanto, embora o parto cesáreo seja um procedimento cirúrgico utilizado em situações de emergência para resguardar a vida do recém-nascido, ele pode sofrer algumas complicações devido à falta de contato mãe-filho (Tesema *et al.*, 2022).

Com relação aos indicadores socioeconômicos de IDH, IVS, IPRS, PIB e PIB *per capita*, sua análise ajuda a compreender a realidade social da região analisada (Costa; Figueiredo, 2021). As condições socioeconômicas podem afetar as condições de saúde da população, portanto, indivíduos vulneráveis apresentam maior risco de morte (Costa; Figueiredo, 2021), porém há estudos que mostram um paradoxo social onde regiões que, mesmo com desigualdade social, possuem assistência adequada para reduzir o risco de óbito e morbidade do nascido vivo (Silva *et al.*, 2010; Rodríguez *et al.*, 2023).

Além disso, essas desigualdades também estão relacionadas à qualidade da assistência recebida pela mãe e pelo recém-nascido. A análise das taxas de leitos pediátricos e médicos podem permitir a avaliação da qualidade da assistência prestada, mostrando a necessidade de

melhoria no planejamento de recursos físicos e humanos para aprimoramento da assistência ao recém-nascido (Almeida *et al.*, 2008; Rodríguez *et al.*, 2023).

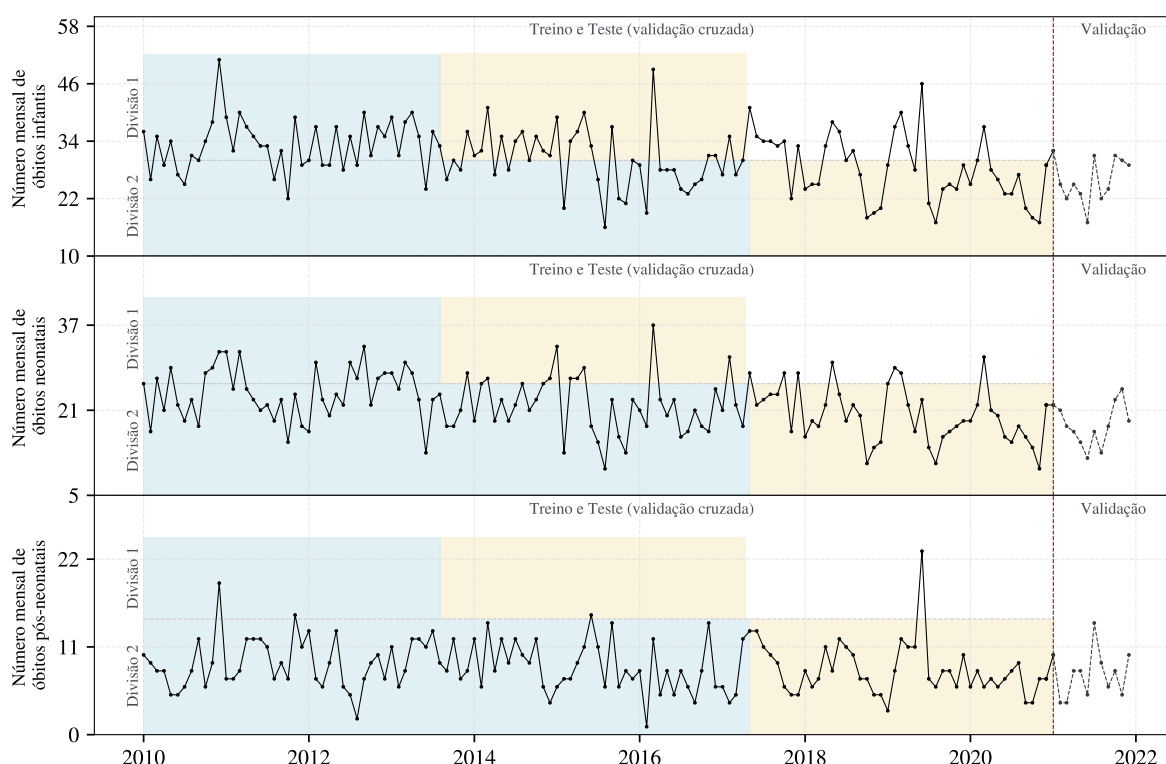
Assim, recém-nascidos com alto risco de óbito têm maior chance de sobrevivência em ambientes hospitalares que oferecem melhores condições em relação aos leitos de UTI pediátrica, melhores equipamentos para exames para diagnósticos e/ou profissionais capacitados para prevenir recém-nascido vulneráveis ou com risco de morte (Kropiweiec; Franco; Amaral, 2017).

4.2 MODELO PARA PREVISÃO DA MORTALIDADE COM SÉRIE DE TEMPO

Esta Seção está dividida em duas subseções principais, sendo 4.2.1 a primeira, apresentando o treinamento e avaliação dos modelos clássicos e com ML do número de óbitos neonatais, pós-neonatais e infantis na RMVPLN do período 2010 – 2020. Em seguida, a subseção 4.2.2 apresenta a validação dos modelos finais com os dados dos óbitos mensais ocorridos no período 2021.

4.2.1 Avaliação dos modelos clássicos e de aprendizado de máquina

Figura 49 – Intervalo de meses da validação cruzada contínua de 2 divisões



Fonte: Elaborado pela autora (2025).

O número de óbitos dos 132 meses do intervalo 2010 – 2020 das faixas etárias neonatal, pós-neonatal e infantil foram particionados em dois subconjuntos de treino-teste usando a técnica de validação cruzada contínua de 2 divisões, conforme ilustrado na Figura 49.

Conforme apresentado na Figura 49 a primeira divisão do conjunto de dados consistiu em 44:44 meses para treinamento e teste, considerando o intervalo de meses de janeiro de 2010 a agosto de 2013 para treinamento e de setembro de 2013 a abril de 2017 para teste dos modelos.

Já a segunda divisão consistiu em 88:44 meses para treinamento e teste, com o intervalo de meses de janeiro de 2010 a abril de 2017 para treinamento e de maio de 2017 até dezembro de 2020 para teste. Essas subdivisões do conjunto de dados foram utilizadas para treinar todos os modelos, seja com técnicas clássicas ou com ML. As estatísticas descritiva de cada subconjunto de dados da validação cruzada contínua é apresentado na Tabela 5.

Tabela 5 – Estatísticas descritivas da divisão do banco de dados com a validação cruzada contínua de 2 divisões

Dados	Média	Var ^a	Desv. pad. ^b	Mín ^c	Q ^d 1	Q2	Q3	Máx ^e
Infantil								
Treino (1º Divisão)	33,32	29,43	5,43	22,00	29,00	33,50	37,00	51,00
Teste (1º Divisão)	39,89	30,14	6,32	16,00	26,75	30,00	34,25	49,00
Treino (2º Divisão)	31,73	36,82	6,07	16,00	28,00	31,00	36,00	51,00
Teste (2º Divisão)	28,39	49,22	7,02	17,00	23,75	28,00	33,25	46,00
Neonatal								
Treino (1º Divisão)	24,25	23,35	4,83	13,00	21,75	24,00	28,00	33,00
Teste (1º Divisão)	29,81	21,77	5,46	10,00	18,00	21,00	26,00	37,00
Treino (2º Divisão)	23,01	27,83	5,28	10,00	19,00	23,00	27,00	37,00
Teste (2º Divisão)	20,18	28,15	5,31	10,00	16,00	19,50	23,25	31,00
Pós-neonatal								
Treino (1º Divisão)	9,07	10,02	3,17	2,00	7,00	9,00	11,25	19,0
Teste (1º Divisão)	10,42	8,36	3,23	1,00	6,00	8,00	11,25	15,00
Treino (2º Divisão)	8,72	10,23	3,20	1,00	6,00	8,00	11,25	19,00
Teste (2º Divisão)	8,20	11,47	3,39	3,00	6,00	7,50	10,00	23,00

(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.
Fonte: Elaborado pela autora (2025).

Os resultados, por cada faixa etária, obtidos com cada técnicas testada são descritos a seguir.

4.2.1.1 Modelos clássicos

A Tabela 6 apresenta os resultados obtidos dos testes ADF e KPSS para cada divisão do conjunto de dados na validação cruzada, sendo importante ressaltar que cada conjunto de dados relacionado com as faixas etárias apresentaram um comportamento diferente.

Tabela 6 – Verificação da estacionariedade dos dados de treinamento utilizando o teste de Dickey-Fuller aumentado e o teste de Kwiatkowski-Phillips-Schmidt-Shin

Dados		ADF ^a	p-valor	KPSS ^b	p-valor	Observação
Infantil						
Treino (1º divisão)	Teste Estatístico	-6,112	0,000	0,082	0,100	Estacionaria
	Valor crítico (1%)	-3,593		0,739		
	Valor crítico (5%)	-2,932		0,463		
	Valor crítico (10%)	-2,604		0,347		
Treino (2º divisão)	Teste Estatístico	-3,691	0,004	0,573	0,025	Não Estacionaria (diferenciar)
	Valor crítico (1%)	-3,510		0,739		
	Valor crítico (5%)	-2,896		0,463		
	Valor crítico (10%)	-2,585		0,347		
1º diferença Treino (2º divisão)	Teste Estatístico	-4,855	0,000	0,322	0,100	Estacionaria
	Valor crítico (1%)	-3,518		0,739		
	Valor crítico (5%)	-2,900		0,463		
	Valor crítico (10%)	-2,587		0,739		
Neonatal						
Treino (1º divisão)	Teste Estatístico	-5,281	0,000	0,073	0,100	Estacionaria
	Valor crítico (1%)	-3,593		0,739		
	Valor crítico (5%)	-2,932		0,463		
	Valor crítico (10%)	-2,604		0,347		
Treino (2º divisão)	Teste Estatístico	-7,856	0,000	0,342	0,100	Estacionaria
	Valor crítico (1%)	-3,508		0,739		
	Valor crítico (5%)	-2,895		0,463		
	Valor crítico (10%)	-2,585		0,347		
Pós-neonatal						
Treino (1º divisão)	Teste Estatístico	-5,237	0,000	0,099	0,100	Estacionaria
	Valor crítico (1%)	-3,593		0,739		
	Valor crítico (5%)	-2,932		0,463		
	Valor crítico (10%)	-2,604		0,347		
Treino (2º divisão)	Teste Estatístico	-8,336	0,000	0,252	0,100	Estacionaria
	Valor crítico (1%)	-3,508		0,739		
	Valor crítico (5%)	-2,895		0,463		
	Valor crítico (10%)	-2,585		0,347		

(a) ADF: Augmented Dickey-Fuller, (b) KPSS: Kwiatkowski-Phillips-Schmidt-Shin.

Fonte: Elaborado pela autora (2025).

Os resultados dos testes ADF ($-6,112 < -2,932_{v_c}$; p-valor $< 0,001$) e KPSS ($0,082 < 0,463_{v_c}$; p-valor = 0,100) da série temporal do conjunto de dados do número mensal de óbitos infantis mostraram que a série temporal do conjunto de treino da primeira divisão é estacionária (Tabela 6). Portanto, com um nível de significância de 5%, a hipótese nula do teste ADF ($ADF < v_c$ e p-valor $< 0,05$) é rejeitado, e a hipótese nula do teste KPSS ($KPSS < v_c$ e p-valor $> 0,05$) é aceito, sugerindo a existência de raiz unitária na série temporal da primeira divisão dos dados da mortalidade infantil.

Pelo contrário, os testes ADF ($-3,691 < -2,896_{v_c}$; p-valor = 0,004) e KPSS ($0,573 > 0,463_{v_c}$; p-valor = 0,025) da série temporal da segunda divisão revelou a presença de não estacionariedade nos dados (Tabela 6), portanto foi aplicada uma diferença de primeira ordem à série temporal para

torná-la estacionária. Seguidamente, a estacionariedade da série diferenciada foi confirmada pelos testes ADF ($-4,855 < -2,900_{v_c}$; p-valor $< 0,001$) e KPSS ($0,322 < 0,463_{v_c}$; p-valor = 0,100), indicando que a diferenciação de primeira ordem foi suficiente para atingir a estacionariedade da série.

Com relação à série temporal do conjunto de dados do número mensal de óbitos neonatais, as duas divisões de treino do conjunto de dados mostraram ser estacionários, com um nível de significância estatística de 5%, conforme apresentado na Tabela 6.

No primeiro subconjunto de dados o valor do teste ADF foi $-5,281 < -2,932_{v_c}$ com p-valor $< 0,001$, e o valor do teste KPSS foi $0,073 < 0,463_{v_c}$ com p-valor = 0,100, já no segundo subconjunto de treino o valor de teste ADF foi $-7,856 < -2,895_{v_c}$ com p-valor $< 0,001$ e do KPSS foi $0,322 < 0,463_{v_c}$ com p-valor = 0,100, resultando na rejeição e aceitação da hipótese nula dos testes ADF e KPSS (Tabela 6).

Do mesmo modo, na série temporal do conjunto de dados do número mensal de óbitos pós-neonatais foi observada a presença de estacionariedade nas duas divisões de treino do conjunto de dados. O resultado do teste ADF foi $-5,237 < -2,932_{v_c}$ (p-valor $< 0,001$) e do teste KPSS foi $0,099 < 0,463_{v_c}$ (p-valor = 0,100) para o primeiro subconjunto de treino da validação cruzada.

No segundo subconjunto de dados os teste ADF ($-8,336 < -2,895_{v_c}$; p-valor $< 0,001$) e KPSS ($0,252 < 0,463_{v_c}$; p-valor = 0,100) foram rejeitado e aceito, respetivamente, indicando que a série não tem raiz unitária.

Além disso, na busca computacional foi utilizado o critério Akaike Information Criterion (AIC) no processo de seleção dos parâmetros dos modelos com melhor ajuste. As Tabelas 7 e 8 apresentam as melhores configurações de parâmetros dos modelos ARIMA e SARIMA para cada modelo treinado na validação cruzada.

Tabela 7 – Modelos ARIMA e SARIMA ajustados para prever o número mensal de óbitos infantis, neonatais e pós-neonatais, 1^o divisão - validação cruzada contínua

Modelo	Parâmetros	Erro padrão	p-valor	IC ^a 95%	AIC ^b
Infantil					
ARIMA(0, 0, 0)	$c = 0,390$	0,029	0,000	[0,334; 0,447]	-19,656
SARIMA(0, 0, 0)(0, 0, 0) ₃	$c = 0,390$	0,029	0,000	[0,334; 0,447]	-19,656
Neonatal					
ARIMA(0, 0, 0)	$c = 0,563$	0,036	0,000	[0,491; 0,634]	2,866
SARIMA(0, 0, 0)(0, 0, 0) ₃	$c = 0,563$	0,036	0,000	[0,491; 0,634]	2,866
Pós-neonatal					
ARIMA(0, 0, 0)	$c = 0,416$	0,030	0,000	[0,358; 0,474]	-20,073
SARIMA(0, 0, 0)(0, 0, 0) ₃	$c = 0,416$	0,030	0,000	[0,358; 0,474]	-20,073

(a) IC: Intervalo de confiança, (b) AIC: Akaike Information Criterion, *Não significativo (p-valor $> 0,05$).
Fonte: Elaborado pela autora (2025).

Conforme contemplado na Tabela 7, no primeiro subconjunto de treinamento e testagem dos modelos com os dados mensais de mortalidade infantil, neonatal e pós-neonatal, todos os modelos apresentaram o processo integrado de ordem zero I(0) dado que as séries temporais de treinamento

foram estacionárias (Tabela 6). Também, observou-se que os melhores parâmetros dos modelos ARIMA e SARIMA foram zero. Estes modelos ARIMA(0, 0, 0) e SARIMA(0, 0, 0)(0, 0, 0)₃ são de ruído branco, com erro constante e distribuídos de forma idêntica, cujas previsões seriam a média dos dados utilizados no treinamento.

Tabela 8 – Modelos ARIMA e SARIMA ajustados para prever o número mensal de óbitos infantis, neonatais e pós-neonatais, 2^o divisão - validação cruzada contínua

Modelo	Parâmetros	Erro padrão	p-valor	IC ^a 95%	AIC ^b
Infantil					
ARIMA(0, 1, 1)	$\theta = -0,927$	0,042	0,000	[-1,009; -0,845]	-53,465
SARIMA(2, 1, 0)(0, 0, 1) ₃	$\phi_1 = -1,009$	0,055	0,000	[-1,116; -0,900]	-52,479
	$\phi_2 = -0,868$	0,096	0,000	[-1,057; -0,679]	
	$\Theta_1 = 0,029$	0,004	0,000	[0,021; 0,036]	
Neonatal					
ARIMA(2, 0, 0)	$c = 0,342$	0,069	0,000	[0,207; 0,476]	35,062
	$\phi_1 = 0,128$	0,112	0,255*	[-0,092; 0,348]	
	$\phi_2 = 0,162$	0,122	0,184*	[-0,077; 0,400]	
SARIMA(0, 0, 0)(0, 0, 1) ₃	$c = 0,482$	0,024	0,000	[0,435; 0,529]	-35,172
	$\Theta_1 = 0,180$	0,107	0,093*	[-0,030; 0,390]	
Pós-neonatal					
ARIMA(0, 0, 0)	$c = 0,429$	0,020	0,000	[0,390; 0,467]	-51,361
SARIMA(0, 0, 0)(0, 0, 0) ₃	$c = 0,429$	0,020	0,000	[0,390; 0,467]	-51,361

(a) IC: Intervalo de confiança, (b) AIC: Akaike Information Criterion, *Não significativo (p-valor >0,05).

Fonte: Elaborado pela autora (2025).

Do mesmo modo, os parâmetros dos modelos treinados com o segundo conjunto de dados dos dados mensais da mortalidade pós-neonatal também foram zero, com erro constante e distribuídos de forma idêntica (Tabela 8).

Por outro lado, os modelos treinados com o segundo conjunto de dados das faixas etárias infantil e neonatal mostraram um comportamento diferente em relação à configuração de seus parâmetros, conforme apresentado na Tabela 8. A segunda série temporal da validação cruzada da mortalidade infantil foi não estacionária, resultando em modelos ARIMA/SARIMA com o processo integrado de primeira ordem I(1), que é o número de vezes de diferenciação necessárias para torná-lo estacionário (Tabela 6).

Adicionalmente, o modelo ARIMA(0, 1, 1) apresentou o processo de médias móveis de primeira ordem, e o modelo SARIMA(2, 1, 0)(0, 0, 1)₃ com os parâmetros de segunda ordem no processo autoregressivo da parte não sazonal e primeira ordem no processo de médias móveis da parte sazonal (Tabela 8).

Os modelos ARIMA e SARIMA obtidos com o segundo subconjunto de dados da mortalidade mensal neonatal apresentaram o processo integrado de ordem zero I(0) por serem séries estacionárias (Tabela 6). O modelo ARIMA(2, 0, 0) apresentou o processo autoregressivo de segunda ordem, e o modelo SARIMA(0, 0, 0)(0, 0, 1)₃ teve os parâmetros da parte não sazonal com ordem zero e primeira ordem no processo de médias móveis da parte sazonal (Tabela 8).

Na sequência, as Tabelas 9 e 10 apresentam os resultados das métricas de avaliação de cada um dos modelos testados na validação cruzada.

Tabela 9 – Métricas de desempenho dos modelos ARIMA e SARIMA treinados na 1^o divisão da validação cruzada contínua

Modelos	MAE	MSE	RMSE	MAPE	SMAPE	MPD
Infantil						
ARIMA(0, 0, 0)	5,630	49,105	7,008	0,218	0,187	1,611
SARIMA(0, 0, 0)(0, 0, 0) ₃	5,630	49,105	7,008	0,218	0,187	1,611
Neonatal						
ARIMA(0, 0, 0)	4,920	35,267	5,939	0,269	0,225	1,596
SARIMA(0, 0, 0)(0, 0, 0) ₃	4,920	5,939	35,267	0,269	0,225	1,596
Pós-neonatal						
ARIMA(0, 0, 0)	2,798	10,682	3,268	0,552	0,345	1,332
SARIMA(0, 0, 0)(0, 0, 0) ₃	2,798	10,682	3,268	0,552	0,345	1,332

Fonte: Elaborado pela autora (2025).

Tabela 10 – Métricas de desempenho dos modelos ARIMA e SARIMA treinados na 2^o divisão da validação cruzada contínua

Modelos	MAE	MSE	RMSE	MAPE	SMAPE	MPD
Infantil						
ARIMA(0, 1, 1)	5,843	48,862	6,990	0,230	0,208	1,719
SARIMA(2, 1, 0)(0, 0, 1) ₃	5,861	49,044	7,003	0,231	0,209	1,726
Neonatal						
ARIMA(2, 0, 0)	4,957	35,557	5,963	0,302	0,245	1,742
SARIMA(0, 0, 0)(0, 0, 1) ₃	4,940	35,251	5,937	0,302	0,245	1,729
Pós-neonatal						
ARIMA(0, 0, 0)	2,569	11,470	3,387	0,371	0,307	1,231
SARIMA(0, 0, 0)(0, 0, 0) ₃	2,569	11,470	3,387	0,371	0,307	1,231

Fonte: Elaborado pela autora (2025).

Os modelos ARIMA e SARIMA das faixas etárias neonatal, pós-neonatal e infantil do primeiro subconjunto de dados (Tabela 9) e os modelos da mortalidade pós-neonatal do segundo subconjunto de dados apresentaram os mesmos valores nas métricas de avaliação (Tabela 10). Isto porque cada par de modelos por série temporal apresentam a mesma configuração de parâmetros encontrados na busca computacional (Tabelas 7 e 8). Além disso, pode-se observar que no caso dos modelos treinados com o segundo subconjunto de dados o modelo ARIMA foi ligeiramente melhor para mortalidade infantil e o modelo SARIMA para mortalidade neonatal.

4.2.1.2 Modelos de aprendizado de máquina

Na aplicação das técnicas de ML, também foi utilizada a técnica de validação cruzada contínua para treinar e testar os modelos, utilizando as mesmas divisões de dados de 44:44 para treino e teste na primeira iteração e 88:44 na segunda iteração (Figura 49).

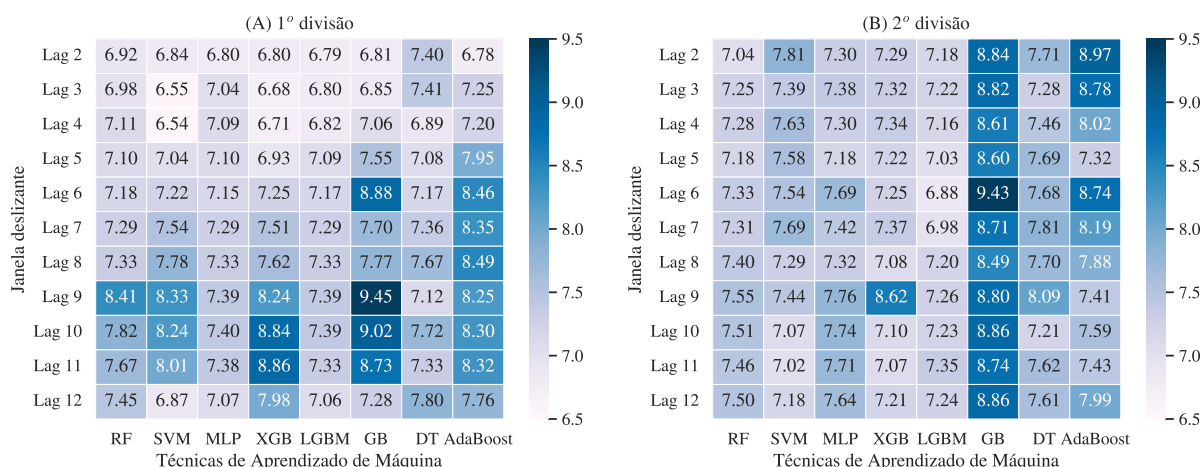
Assim, inicialmente foi analisado o número de observações (meses) ou períodos que podem ser considerados para treinar os modelos para fazer uma nova previsão. Neste sentido, foi

realizado um processo iterativo considerando 11 configurações diferentes, também conhecida como janela deslizante (*lag*), para treinar cada um dos modelos de ML.

Em específico, foram testadas entre 2 e 12 meses atrás para prever um novo valor, com as técnicas de RF, SVM, MLP, XGB, LGBM, GB, DT e AdaBoost. Além disso, foi realizada uma busca computacional exaustiva utilizando o método de busca aleatória para encontrar a melhor configuração de hiperparâmetros para cada uma das técnicas utilizadas.

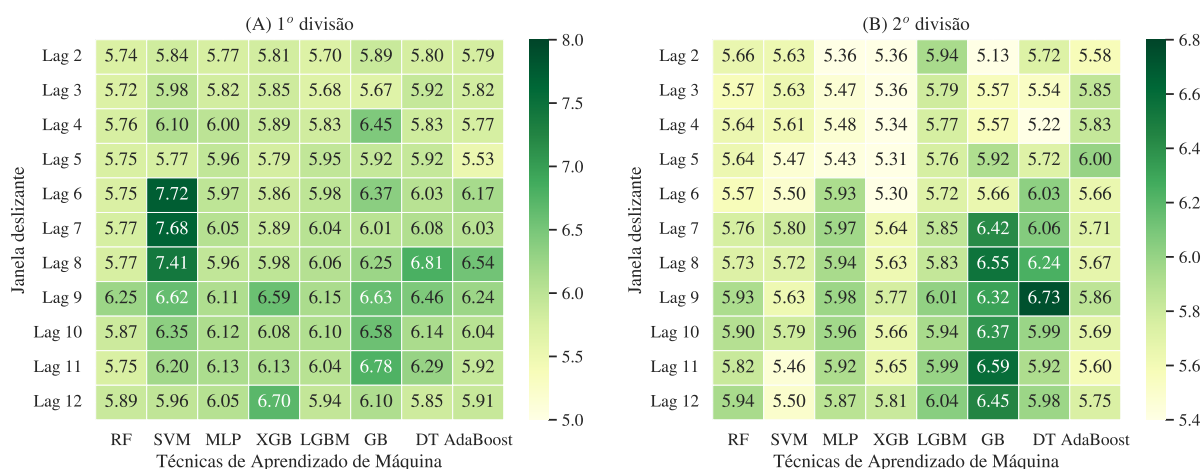
Nas Figuras 50, 51 e 52 são apresentados os valores RMSE de cada modelo treinado na validação cruzada com as 11 configurações da janela deslizante dos dados mensais da mortalidade infantil, neonatal e pós-neonatal.

Figura 50 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade infantil



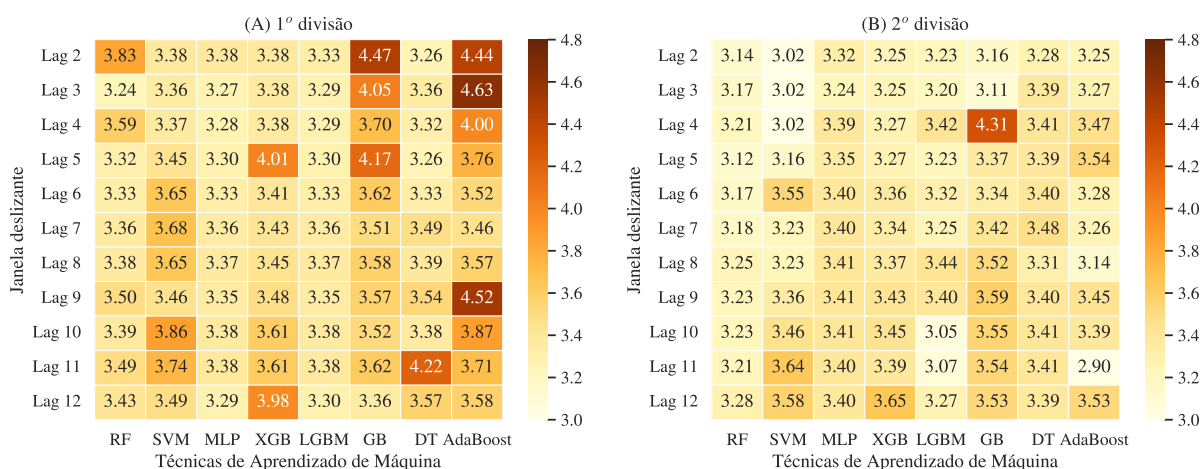
Fonte: Elaborado pela autora (2025).

Figura 51 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade neonatal



Fonte: Elaborado pela autora (2025).

Figura 52 – Mapa de calor dos valores de RMSE de cada um dos modelos treinados com as configurações de janela deslizante, com dados de mortalidade pós-neonatal

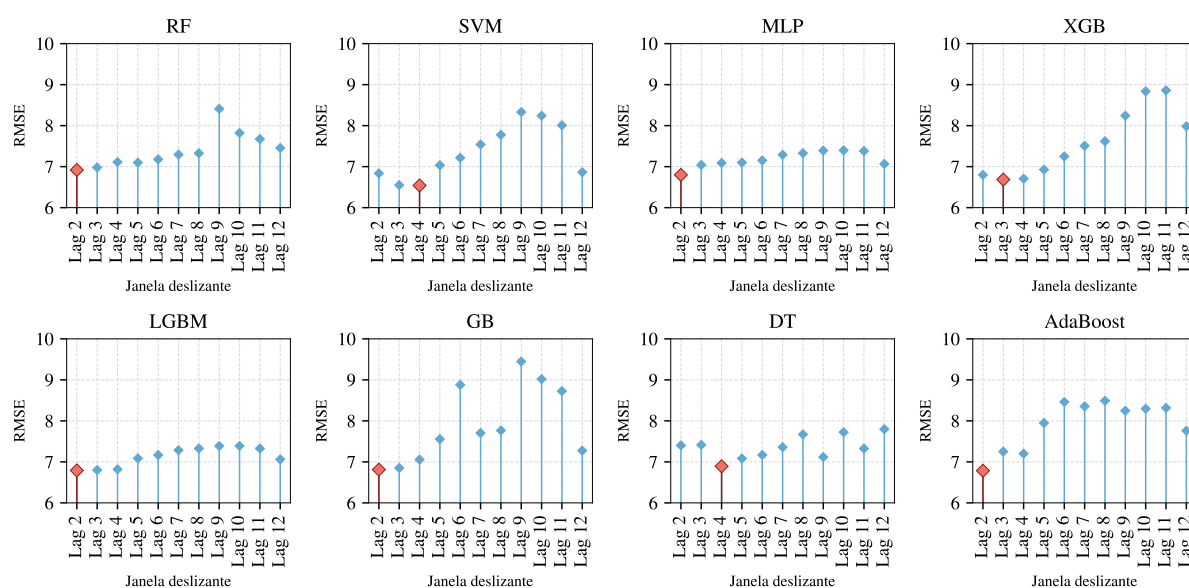


Fonte: Elaborado pela autora (2025).

Os mapas de calor apresentados nas Figuras 50, 51 e 52 permitiram observar que os menores valores de RMSE estão posicionados principalmente nos modelos com valores baixos na configuração da vantagem deslizante. Deste jeito, cerca de 50% dos modelos treinados com cada subconjunto de dados de validação cruzada apresentaram valores de RMSE abaixo de 7,5 nos modelos de mortalidade infantil, 6 nos modelos de mortalidade neonatal e 3,5 nos modelos de mortalidade pós-neonatal.

Portanto, com o primeiro subconjunto de dados dos óbitos infantis, os modelos RF, MLP, LGBM, GB e AdaBoost apresentaram o menor erro RMSE considerando dois meses anteriores para suas previsões, XGB com 3 meses e os modelos SVM e DT com 4 meses (Figura 53).

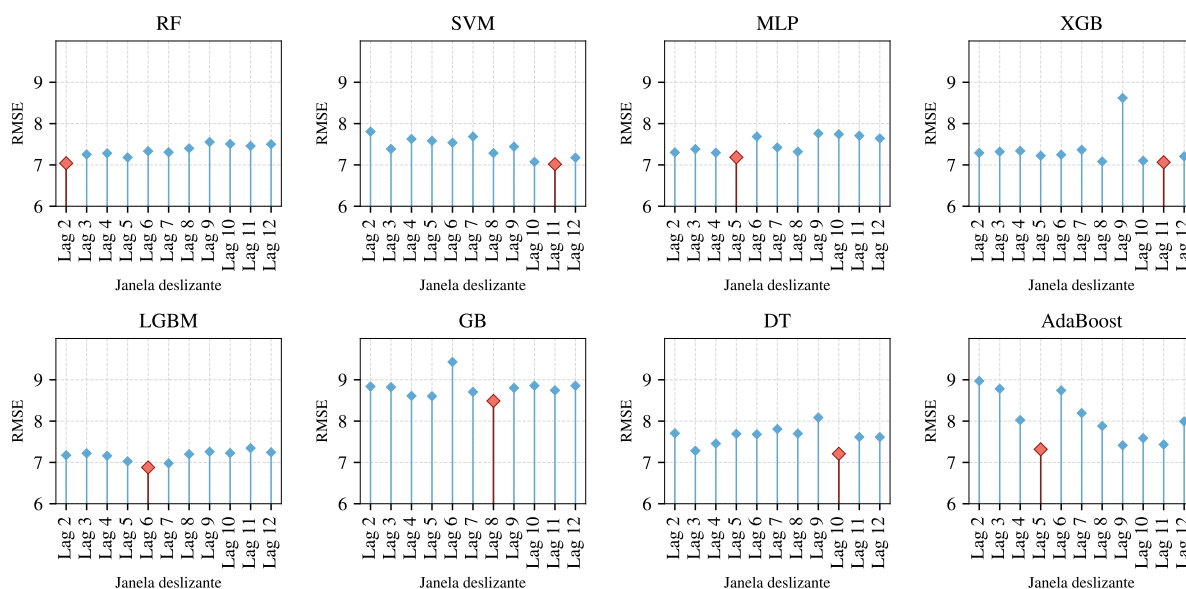
Figura 53 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1ª divisão da validação cruzada contínua da mortalidade infantil



Fonte: Elaborado pela autora (2025).

Com o segundo subconjunto de dados da mortalidade infantil houve maior variação entre as configurações da janela deslizante, observando-se que o modelo RF foi melhor com 2 meses, MLP e AdaBoost com 5, SVM e XGB com 11 meses, e LGBM, GB e DT com 6, 8 e 10 meses.

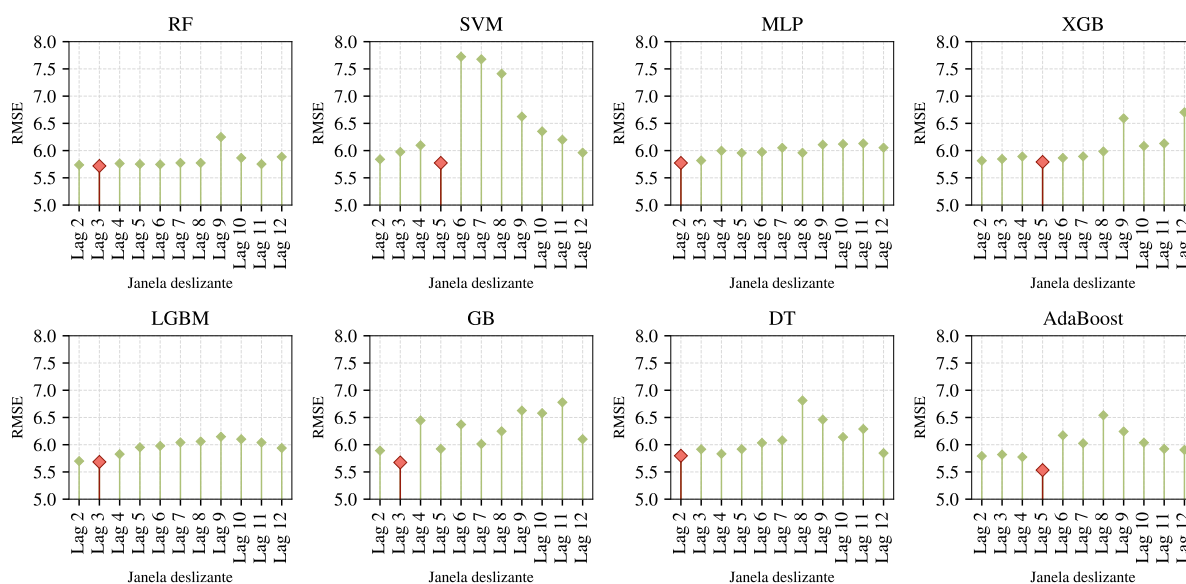
Figura 54 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade infantil



Fonte: Elaborado pela autora (2025).

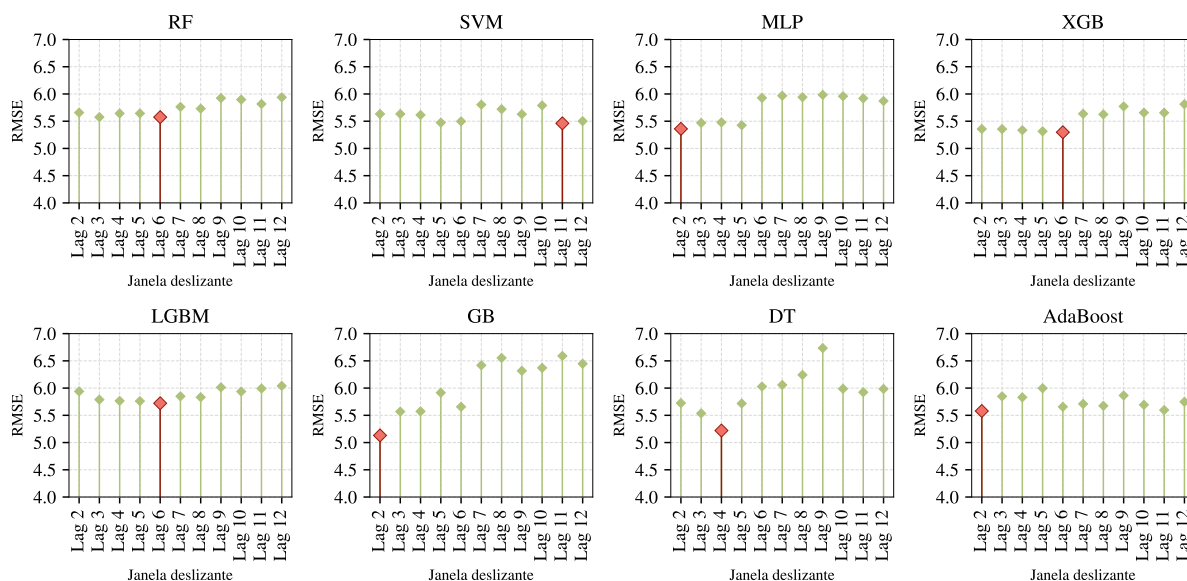
Em relação aos modelos do número mensal de óbitos neonatais, as Figuras 55 e 56 apresentam as melhores configurações da janela deslizante de cada modelo testado na validação cruzada.

Figura 55 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1ª divisão da validação cruzada contínua da mortalidade neonatal



Fonte: Elaborado pela autora (2025).

Figura 56 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade neonatal

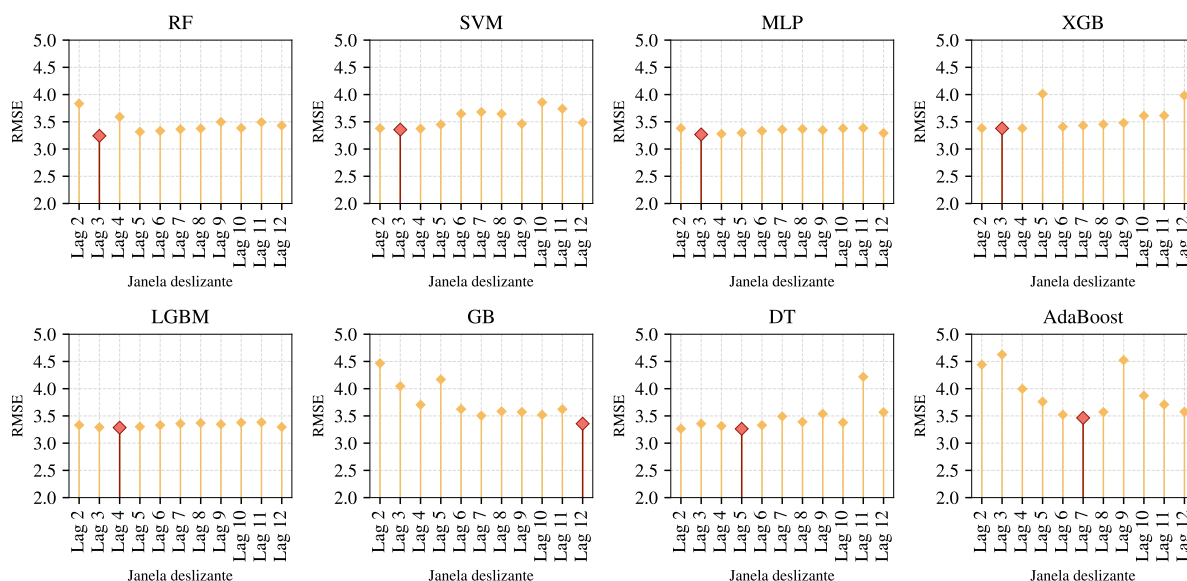


Fonte: Elaborado pela autora (2025).

Das Figuras 55 e 56, os modelos de MLP, DT, RF, LGBM e GB apresentaram melhor desempenho considerando entre 2 e 3 meses anteriores, e os modelos de SVM, XGB e AdaBoost com 5 meses, no primeiro treino. Já no segundo treino, o MLP, GB, AdaBoost e DT apresentaram menor erro com 2 e 4 meses anteriores, e os outros modelos com mais de 6 meses.

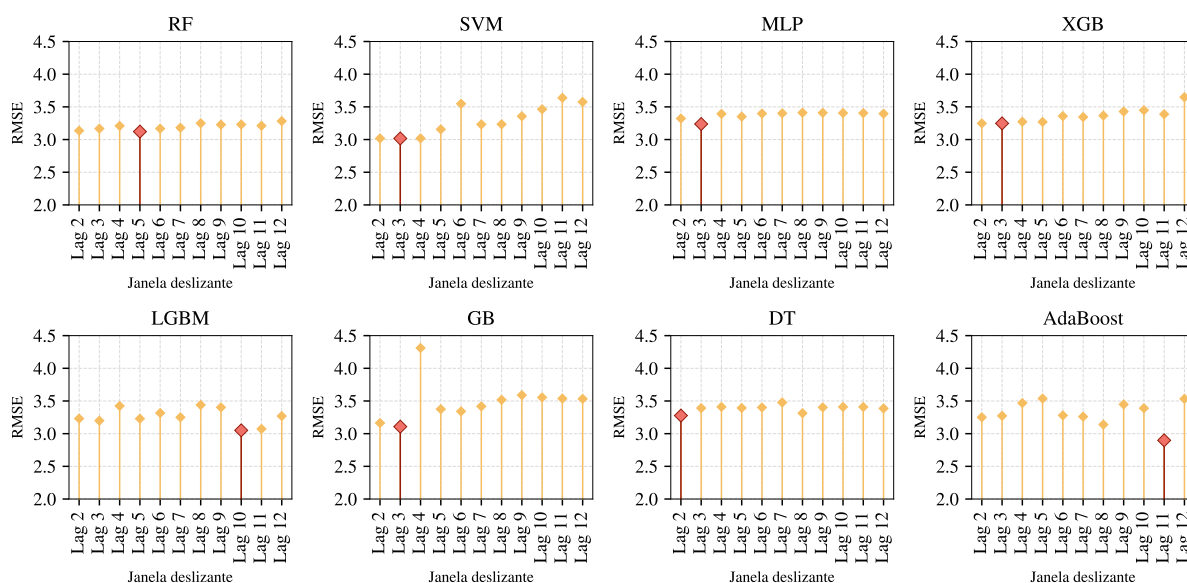
Da mesma forma, as Figuras 57 e 58 apresentam as melhores configurações para os modelos de previsão mensal de óbitos pós-neonatais.

Figura 57 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 1ª divisão da validação cruzada contínua da mortalidade pós-neonatal



Fonte: Elaborado pela autora (2025).

Figura 58 – Identificação do tamanho da janela deslizante dos modelos com menor valor de RMSE, na 2ª divisão da validação cruzada contínua da mortalidade pós-neonatal

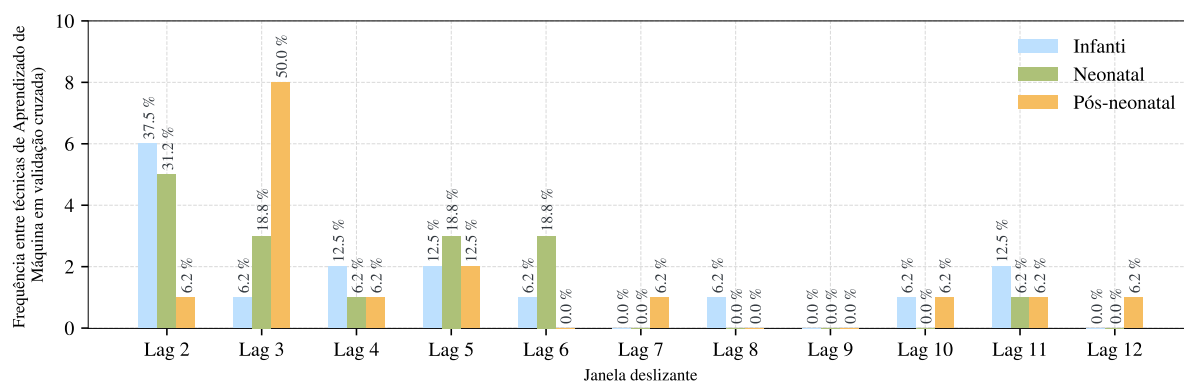


Fonte: Elaborado pela autora (2025).

No primeiro treinamento (Figura 57), os modelos com as técnicas de RF, SVM, MLP e XGB apresentaram o menor erro RMSE em suas previsões com 3 meses anteriores, LGBM com 4 meses, DT com 5, AdaBoost com 7 e GB com 12 meses. No segundo treinamento (Figura 58), o modelos com DT foi melhor considerando 2 meses meses atrás, os modelos de SVM, MLP, XGB e GB com 3 meses e, RF, LGBM e AdaBoost com 5, 10 e 11 meses respectivamente, na sua configuração da janela deslizante.

Além disso, levando em consideração a melhor configuração da janela deslizante considerando cada uma das técnicas testadas em cada iteração da validação cruzada, foi observado que dos 16 modelos treinados (8 da 1ª e 2ª divisão) para mortalidade infantil, 37,5% (6) apresentaram menor valor de RMSE considerando 2 meses anteriores para seu treinamento (Figura 59).

Figura 59 – Frequência do tamanho da janela deslizante entre os modelos treinados



Fonte: Elaborado pela autora (2025).

Dos modelos treinados na validação cruzada para a mortalidade neonatal, 31,2% (5) apresentaram melhor desempenho com 2 meses e 18,8% (3) dos modelos com 3, 5 e 6 meses na configuração da janela deslizante. Pelo contrário, nos modelos treinados para a mortalidade pós-neonatal, 50,0% (8) das técnicas apresentaram menor RMSE considerando 3 meses para seu treinamento (Figura 59).

Complementarmente, na Tabela 11 são apresentados os resultados das métricas de avaliação de cada um dos modelos testados selecionados pela melhor configuração da janela deslizante, com os dados do número mensal de óbitos infantil, neonatal e pós-neonatal da primeira divisão de dados da validação cruzada.

Tabela 11 – Métricas de desempenho dos modelos de aprendizado de máquina treinados na 1ª divisão da validação cruzada contínua

Modelos	Lag	MAE	MSE	RMSE	MAPE	SMAPE	MPD
Infantil							
RF	2	5,294	47,859	6,918	0,203	0,176	1,568
SVM	4	5,043	42,911	6,551	0,191	0,168	1,405
MLP	2	5,242	46,218	6,798	0,201	0,174	1,517
XGB	3	5,147	44,684	6,685	0,196	0,172	1,471
LGBM	2	5,375	46,115	6,791	0,204	0,179	1,517
GB	2	5,300	46,388	6,811	0,204	0,176	1,517
DT	4	5,244	47,509	6,893	0,201	0,174	1,554
AdaBoost	2	5,574	51,848	7,201	0,211	0,184	1,681
Neonatal							
RF	3	4,540	32,697	5,718	0,245	0,209	1,478
SVM	5	4,803	33,333	5,773	0,259	0,220	1,507
MLP	2	4,718	33,324	5,773	0,251	0,217	1,505
XGB	5	4,787	33,542	5,792	0,254	0,219	1,508
LGBM	3	4,539	32,308	5,684	0,243	0,209	1,461
GB	3	4,457	34,703	5,891	0,244	0,205	1,565
DT	2	4,605	34,020	5,833	0,248	0,212	1,525
AdaBoost	5	4,518	33,544	5,792	0,245	0,207	1,511
Pós-neonatal							
RF	3	2,709	10,515	3,243	0,534	0,335	1,310
SVM	3	2,768	11,262	3,356	0,529	0,341	1,394
MLP	3	2,798	10,675	3,267	0,552	0,345	1,331
XGB	3	2,819	11,417	3,379	0,547	0,347	1,410
LGBM	4	2,944	11,411	3,378	0,592	0,359	1,410
GB	12	2,763	11,270	3,357	0,513	0,344	1,431
DT	5	2,790	10,654	3,264	0,550	0,344	1,328
AdaBoost	7	3,152	13,758	3,709	0,565	0,381	1,653

Fonte: Elaborado pela autora (2025).

No primeiro treinamento das técnicas testadas (Tabela 11), foi observado que o modelo de SVM (MAE = 5,043, MSE = 42,911, RMSE = 6,551, MAPE = 0,191, SMAPE = 0,168, MPD = 1,405) apresentou as melhores métricas para previsão mensal da mortalidade infantil, em comparação das outras técnicas testadas. Para a mortalidade pós-neonatal mensal o melhor

modelos foi o desenvolvido com a técnica de RF (MAE = 2,709, MSE = 10,515, RMSE = 3,243, MAPE = 0,534, SMAPE = 0,335, MPD = 1,31).

Por outro lado, para a predição da mortalidade neonatal, observou-se que duas técnicas apresentaram melhores valores nas métricas de erro analisadas. Em relação à mortalidade mensal neonatal os melhores modelos foram os desenvolvidos com as técnicas de GB (MAE = 4,457, MSE = 34,703, RMSE = 5,891, MAPE = 0,244, SMAPE = 0,205, MPD = 1,565) e LGBM (MAE = 4,539, MSE = 32,308, RMSE = 5,684, MAPE = 0,243, SMAPE = 0,209, MPD = 1,461).

Na Tabela 12 são apresentados os resultados das métricas de avaliação dos modelos com a melhor configuração da janela deslizante, treinados com a segunda divisão de dados da validação cruzada para cada faixa etária.

Tabela 12 – Métricas de desempenho dos modelos de aprendizado de máquina treinados na 2ª divisão da validação cruzada contínua

Modelos	Lag	MAE	MSE	RMSE	MAPE	SMAPE	MPD
Infantil							
RF	2	5,639	49,573	7,041	0,232	0,201	1,738
SVM	11	6,022	54,551	7,386	0,243	0,214	1,902
MLP	5	5,890	53,325	7,302	0,244	0,209	1,859
XGB	11	5,941	53,583	7,320	0,243	0,211	1,870
LGBM	6	5,826	51,482	7,175	0,238	0,207	1,801
GB	8	6,614	78,109	8,838	0,279	0,227	2,539
DT	10	6,269	55,605	7,457	0,256	0,222	1,942
AdaBoost	5	6,275	64,376	8,023	0,249	0,221	2,231
Neonatal							
RF	6	4,674	31,078	5,575	0,277	0,233	1,540
SVM	11	4,483	29,962	5,474	0,254	0,225	1,487
MLP	2	4,474	28,723	5,359	0,265	0,224	1,430
XGB	6	4,449	28,239	5,314	0,261	0,223	1,405
LGBM	6	4,875	33,506	5,788	0,286	0,243	1,654
GB	2	4,238	26,328	5,131	0,242	0,211	1,298
DT	4	4,297	27,258	5,221	0,254	0,217	1,367
AdaBoost	2	4,650	31,127	5,579	0,272	0,229	1,505
Pós-neonatal							
RF	5	2,310	10,033	3,167	0,333	0,278	1,065
SVM	3	2,035	9,101	3,017	0,284	0,249	0,966
MLP	3	2,359	10,482	3,238	0,337	0,283	1,114
XGB	3	2,376	10,546	3,247	0,342	0,285	1,122
LGBM	10	2,270	9,307	3,051	0,333	0,274	0,986
GB	3	2,758	12,482	3,533	0,404	0,325	1,339
DT	2	2,411	10,735	3,276	0,346	0,289	1,145
AdaBoost	11	2,344	8,398	2,898	0,358	0,286	0,951

Fonte: Elaborado pela autora (2025).

Assim, no caso dos melhores modelos desenvolvidos com o segundo subconjunto de treinamento (Tabela 12), observou-se que para os dados da mortalidade infantil mensal o melhor

modelo foi RF (MAE = 5,639, MSE = 49,573, RMSE = 7,041, MAPE = 0,232, SMAPE = 0,201, MPD = 1,738) e para a mortalidade neonatal o modelo com GB (MAE = 4,238, MSE = 26,328, RMSE = 5,131, MAPE = 0,242, SMAPE = 0,211, MPD = 1,298).

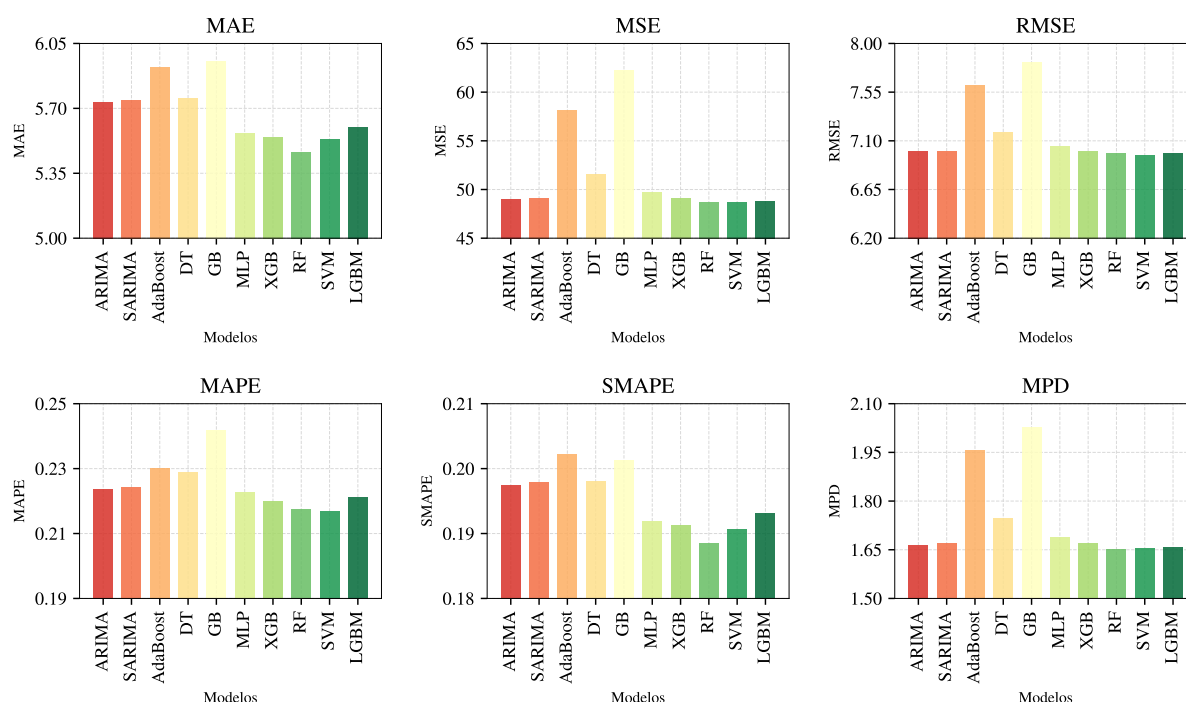
No entanto, os modelos com SVM (MAE = 2,035, MSE = 9,101, RMSE = 3,017, MAPE = 0,284, SMAPE = 0,249, MPD = 0,966) e AdaBoost (MAE = 2,344, MSE = 8,398, RMSE = 2,898, MAPE = 0,358, SMAPE = 0,286, MPD = 0,951) apresentaram um melhor desempenho para a previsão mensal da mortalidade pós-neonatal.

4.2.1.3 Comparação de modelos com métricas de avaliação

Subsequentemente, o desempenho dos modelos clássicos e com ML são comparados, considerando como valor de referência o valor médio das métricas de avaliação da validação cruzada.

A Figura 60 apresenta o valor médio das métricas de desempenho de todos os modelos testados para a previsão mensal de óbitos infantis na RMVPLN.

Figura 60 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade infantil



Fonte: Elaborado pela autora (2025).

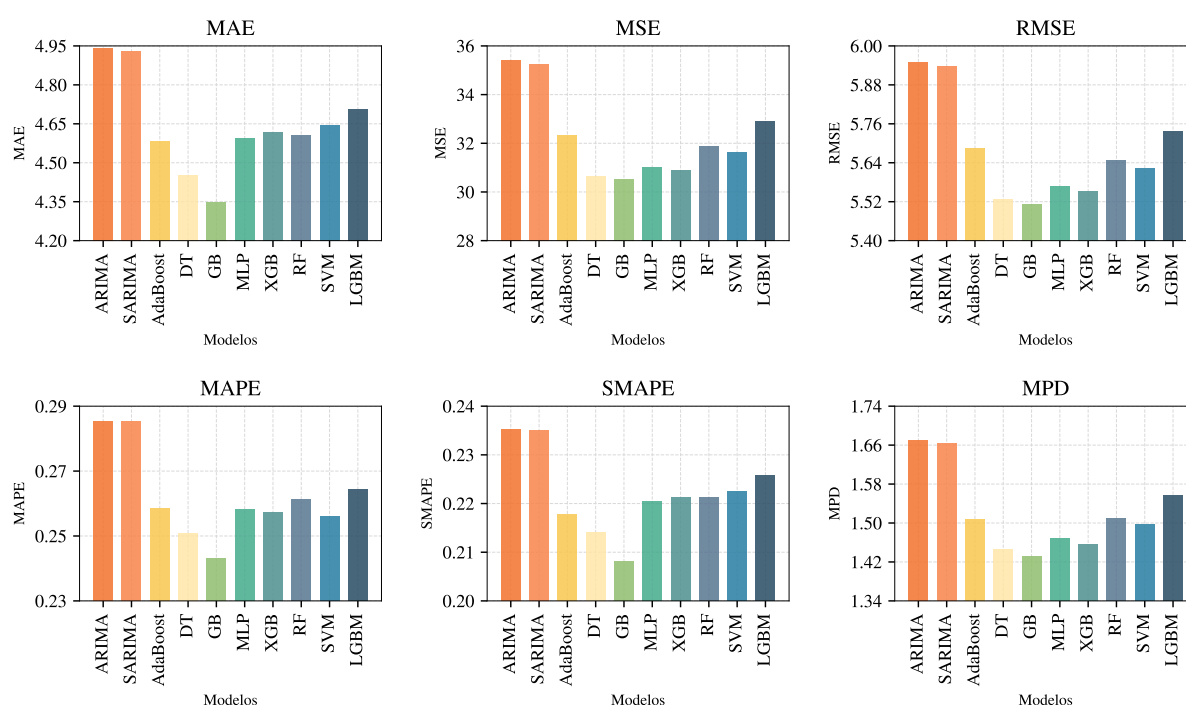
Os resultados da Figura 60 apontaram que os modelos clássicos ARIMA (MAE = 5,736, MSE = 48,984, RMSE = 6,999, MAPE = 0,224, SMAPE = 0,197, MPD = 1,665) e SARIMA (MAE = 5,746, MSE = 49,075, RMSE = 7,005, MAPE = 0,224, SMAPE = 0,198, MPD = 1,669) apresentaram menor erro médio nas métricas de erro, em comparação com os modelos de ML com GB ($Lag_1 = 2$, $Lag_2 = 8$, MAE = 5,601, MSE = 62,249, RMSE = 7,825, MAPE = 0,242, SMAPE = 0,202, MPD = 2,028), AdaBoost ($Lag_1 = 2$, $Lag_2 = 5$, MAE = 5,925, MSE = 58,112,

RMSE = 7,612, MAPE = 0,230, SMAPE = 0,203, MPD = 1,956) e DT ($Lag_1 = 4$, $Lag_2 = 10$, MAE = 5,757, MSE = 51,557, RMSE = 7,175, MAPE = 0,229, SMAPE = 0,198, MPD = 1,748).

Por outro lado, o modelo desenvolvido com RF ($Lag_1 = 2$, $Lag_2 = 2$, MAE = 5,467, MSE = 48,716, RMSE = 6,980, MAPE = 0,218, SMAPE = 0,189, MPD = 1,653) apresentou melhores métricas de avaliação, com os menores valores nas diferentes métricas de erro, em comparação com as outras técnicas de ML e modelos clássicos (Figura 60).

A comparação dos valores médios das métricas de desempenho dos modelos treinados para previsão mensal de óbitos neonatais na RMVPLN são apresentados na Figura 61.

Figura 61 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade neonatal



Fonte: Elaborado pela autora (2025).

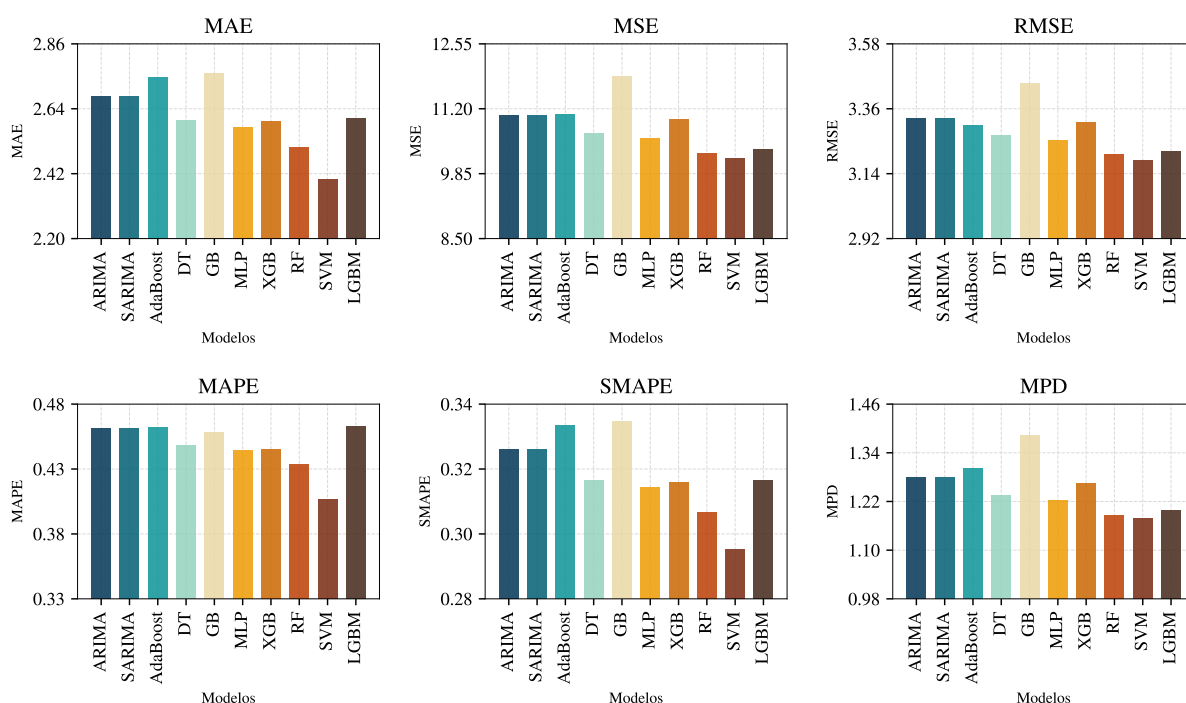
Os resultados mostraram que entre os modelos clássicos (Figura 61), o modelo SARIMA (MAE = 4,930, MSE = 35,259, RMSE = 5,938, MAPE = 0,286, SMAPE = 0,235, MPD = 1,663) apresentou melhor adequação na previsão da mortalidade neonatal mensal, diferentemente do ARIMA (MAE = 4,939, MSE = 35,412, RMSE = 5,951, MAPE = 0,285, SMAPE = 0,235, MPD = 1,669). Não obstante, em comparação com os modelos de ML, ficou evidente que os modelos desenvolvidos com técnicas de ML apresentaram uma melhor capacidade de prever o número mensal de óbitos neonatais.

Dos modelos com ML, a técnica de LGBM ($Lag_1 = 3$, $Lag_2 = 6$, MAE = 4,707, MSE = 32,907, RMSE = 5,736, MAPE = 0,265, SMAPE = 0,226, MPD = 1,558) apresentou os piores valores nas métricas avaliadas, mas, ainda assim, foram valores inferiores quando comparados aos modelos SARIMA e ARIMA (Figura 61).

Conforme a avaliação dos erros médios, o modelo de GB ($Lag_1 = 3$, $Lag_2 = 2$, MAE = 4,348, MSE = 30,516, RMSE = 5,511, MAPE = 0,243, SMAPE = 0,208, MPD = 1,432) apresentou as melhores métricas entre os modelos treinados, com menores valores nas métricas de erro (Figura 61).

Em relação à comparação dos erros médios das métricas dos modelos treinados para previsão do número mensal de óbitos pós-neonatais na RMVPLN, o comportamento das métricas dos modelos testados foi semelhante ao dos modelos de previsão de óbitos infantis (Figura 60), possivelmente porque a mortalidade pós-neonatal faz parte da mortalidade infantil, conforme ilustrado na Figura 62.

Figura 62 – Valores médios das métricas de desempenho dos modelos clássicos com os modelos de aprendizado de máquina, nos dados da mortalidade pós-neonatal



Fonte: Elaborado pela autora (2025).

De forma semelhante com os modelos da Figura 60, os modelos com GB ($Lag_1 = 12$, $Lag_2 = 3$, MAE = 2,761, MSE = 11,876, RMSE = 3,445, MAPE = 0,459, SMAPE = 0,335, MPD = 1,385) e AdaBoost ($Lag_1 = 7$, $Lag_2 = 11$, MAE = 2,748, MSE = 11,078, RMSE = 3,304, MAPE = 0,462, SMAPE = 0,334, MPD = 1,302) os maiores valores nas métricas de avaliação, ainda superiores dos modelos clássicos (MAE = 2,683, MSE = 11,076, RMSE = 3,328, MAPE = 0,461, SMAPE = 0,326, MPD = 1,281).

No entanto, a técnica de SVM ($Lag_1 = 3$, $Lag_2 = 3$, MAE = 2,402, MSE = 10,182, RMSE = 3,187, MAPE = 0,407, SMAPE = 0,295, MPD = 1,180) apresentou os melhores valores médios nas diferentes métricas de erro, e comparação com as outras técnicas de ML testadas.

Em suma, de acordo com os resultados obtidos nas métricas de erro da validação cruzada contínua, as técnicas que apresentaram melhor adequação na previsão mensal de óbitos infantis,

neonatais e pós-neonatais foram RF, GB e SVM, respectivamente. Os modelos desenvolvidos com estas técnicas demonstraram capturar melhor os padrões nos dados, tomando como referência a quantidade de óbitos dos últimos 2 ou 3 meses para fazer novas previsões, de acordo com a configuração da janela deslizante de cada modelo (Tabelas 11 e 12).

Além disso, essa configuração da janela deslizante também é semelhante dos resultados obtidos na análise de frequência da Figura 59.

4.2.2 Validação dos resultados

Com base nos resultados obtidos na validação cruzada contínua, os modelos com RF ($lag = 2$), GB ($lag = 2$) e SVM ($lag = 3$) foram treinados mais uma vez, considerando a série temporal do período 2010 – 2020 do número mensal de óbitos infantis, neonatais e pós-neonatais, respectivamente.

Além disso, do mesmo modo que no processo de experimentação, no treinamento dos modelos foi realizada uma busca computacional da melhor configuração de hiperparâmetros de cada técnica. O Quadro 7 apresenta a melhor configuração de hiperparâmetros de cada modelo treinado com os dados de 2010 – 2020.

Quadro 7 – Configuração de hiperparâmetros dos modelos de aprendizado de máquina

(continua)

Modelos	Hiperparâmetros	Descrição
RF	lag: 2	Janela deslizante
	n_estimators: 843	Número de árvores.
	criterion: squared_error	Medida da qualidade da divisão.
	bootstrap: True	Seleção de amostras para construir as árvores.
	min_samples_split: 7	Número mínimo de amostras para dividir o nó interno.
	min_samples_leaf: 4	Número mínimo de amostras em um nó folha.
	max_depth: None	Profundidade máxima da árvore, sendo com None os nós expandidos até que as folhas contêm menos de 7 amostras.
	max_features: 1	Todas as variáveis de entrada são consideradas na busca da melhor divisão.
GB	lag: 2	Janela deslizante
	n_estimators: 225	Número de árvores.
	loss: huber	Função de perda.
	alpha: 0,61222	Valor alpha da função de perda.
	criterion: friedman_mse	Medida da qualidade da divisão.
	min_samples_split: 7	Número mínimo de amostras para dividir o nó interno.
	min_samples_leaf: 3	Número mínimo de amostras em um nó folha.
	max_depth: 1	Profundidade máxima da árvore.
	max_features: 1	Todas as variáveis de entrada são consideradas na busca da melhor divisão.

Fonte: Elaborado pela autora (2025).

Quadro 7 – Configuração de hiperparâmetros dos modelos de aprendizado de máquina (conclusão)

Modelos	Hiperparâmetros	Descrição
SVM	lag: 3	Janela deslizante
	C: 63,800	Parâmetro de regularização.
	kernel: sigmoid	Função <i>kernel</i> .
	gamma: auto	Coefficiente da função <i>kernel</i> , (1/nº variáveis).
	epsilon: 0,17684	Margem de tolerância do hiperplano.
	max_iter: -1	Limite máximo de iterações

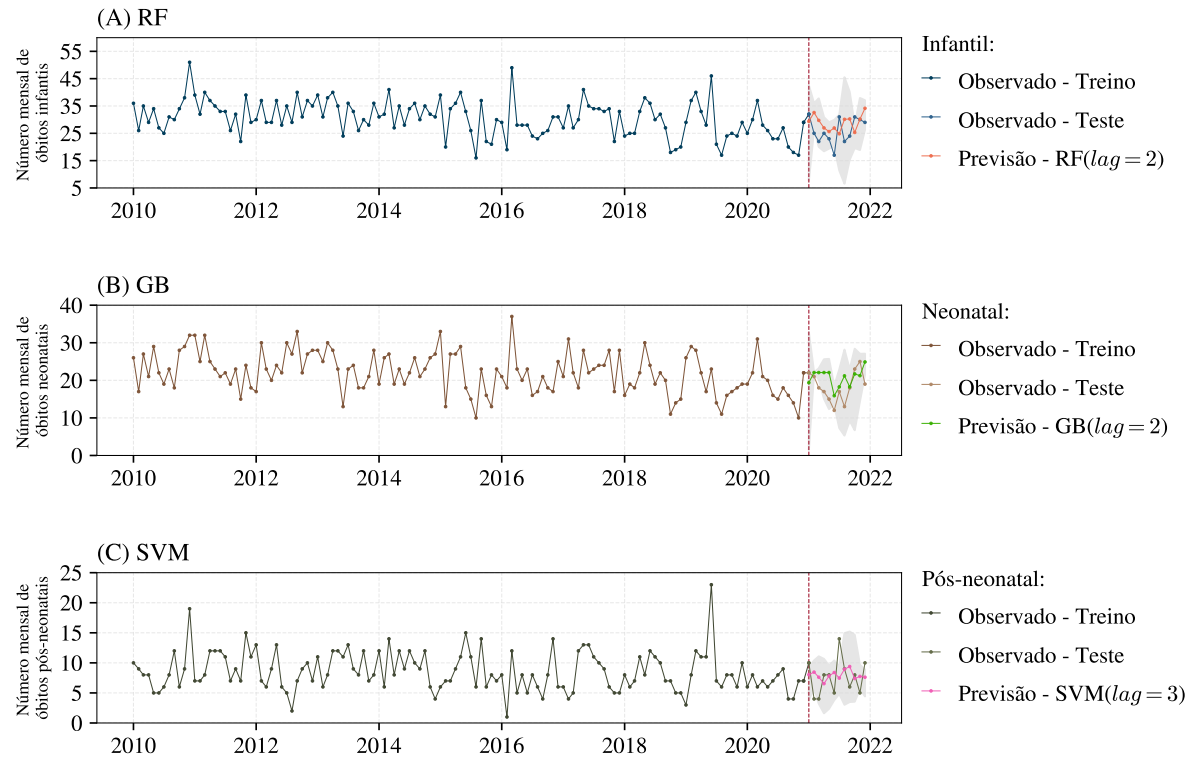
Fonte: Elaborado pela autora (2025).

Tabela 13 – Métricas de desempenho dos modelos com os dados de validação

Modelos	MAE	MSE	RMSE	MAPE	SMAPE	MPD
Infantil						
RF	5,343	36,367	6,030	0,226	0,202	1,356
Neonatal						
GB	3,742	17,326	4,162	0,224	0,199	0,898
Pós-neonatal						
SVM	2,653	9,998	3,162	0,424	0,354	1,282

Fonte: Elaborado pela autora (2025).

Figura 63 – Série temporal original com previsões de cada modelo validado para a mortalidade mensal infantil, neonatal e pós-neonatal



Fonte: Elaborado pela autora (2025).

A adequação e desempenho dos modelos foi avaliada utilizando as informações dos óbitos mensais de cada faixa etária do ano 2021. A Tabela 13 apresenta as métricas de avaliação de cada modelo em relação às previsões mensais do período de validação. Assim, os resultados obtidos nas métricas de erro de cada modelo treinado foram muito próximos dos valores obtidos nas métricas de erro da validação cruzada.

A Figura 63 apresenta a série temporal original do número mensal de óbitos infantis, neonatais e pós-neonatais, e as previsões do período de validação 2021 de cada faixa etária dos modelos RF (A), GB (B) e SVM (C).

Adicionalmente, a Tabela 14 apresenta os resultados do teste-t pareado das previsões de cada modelo treinado. Esta análise foi realizada com a finalidade de comparar estatisticamente o desempenho de cada modelo, testando as previsões dos modelos RF ($\bar{x} = 28,85$), GB ($\bar{x} = 20,56$) e SVM ($\bar{x} = 7,92$) com os dados originais de validação da mortalidade infantil ($\bar{x} = 25,92$), neonatal ($\bar{x} = 18,33$) e pós-neonatal ($\bar{x} = 7,58$).

Tabela 14 – Resultado do teste-T pareado

Modelos	valor-t	$t_{crítico}$	p-valor	Observação
Infantil				
RF	1,843	2,201	0,092	Aceitamos a H_0^a
Neonatal				
GB	2,098	2,201	0,060	Aceitamos a H_0
Pós-neonatal				
SVM	0,353	2,201	0,731	Aceitamos a H_0

(a) H_0 : Hipótese nula ($\mu_x = \mu_y$).

Fonte: Elaborado pela autora (2025).

Dos resultados do teste-t pareado (Tabela 14), a hipótese nula (H_0) foi aceita para os três modelos, demonstrando que os modelos desenvolvidos com técnicas de ML capturaram os padrões complexos dos dados. Portanto, as previsões de cada modelo treinado mostraram proximidade com os dados originais de cada faixa etária, com um nível de significância de 5%.

Deste modo, embora as previsões com modelos de séries temporais possam ser positivas para a tomada de decisões (Silva *et al.*, 2010), a sua aplicação ao problema da mortalidade de nascidos vivos ainda tem sido pouco abordada.

Por conseguinte, se comparar estes resultados com estudos publicados, observou-se que os modelos com técnicas de ML podem superar as técnicas clássicas (Chen; He; Wang, 2022; Rodriguez *et al.*, 2022; Ventura *et al.*, 2023), como ARIMA e SARIMA, demonstrando serem capazes de capturar padrões do conjunto de dados.

4.3 MODELO PARA IDENTIFICAÇÃO DE RISCO DE ÓBITO

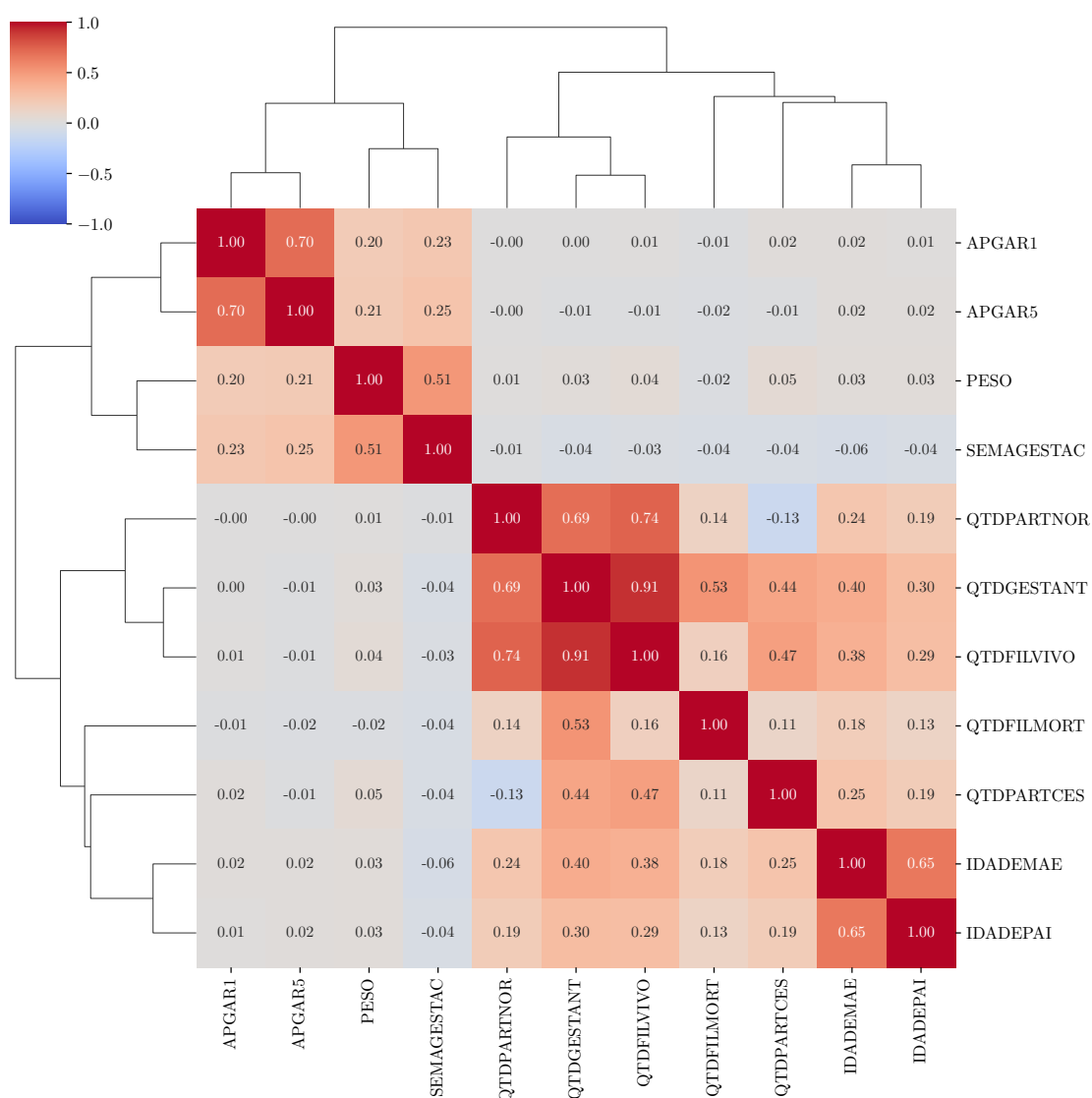
Esta seção está estruturada em três subseções principais. Na subseção 4.3.1, são apresentados os resultados da seleção de características e da análise de importância de variáveis. Na subseção

4.3.2, são apresentadas as avaliações de diferentes técnicas de ML durante o processo de experimentação e treinamento, abrangendo tanto o desenvolvimento de modelos binários para a identificação do risco de óbito infantil quanto de modelos de classificação multiclasse para a identificação dos riscos de óbito neonatal e pós-neonatal. Por fim, a subseção 4.3.3 apresenta os resultados do processo de validação dos modelos finais, contemplando tanto o modelo binário quanto o multiclasse.

4.3.1 Seleção e importância das características

Considerando que o sucesso e a eficiência dos modelos de ML estão diretamente relacionadas à qualidade das variáveis selecionadas, as análises estatísticas preliminares permitiram identificar a presença de relações significativas entre as características da mãe e do recém-nascido. As análises individuais das variáveis estão apresentadas no Apêndice K.

Figura 64 – Matriz de correlação de Pearson

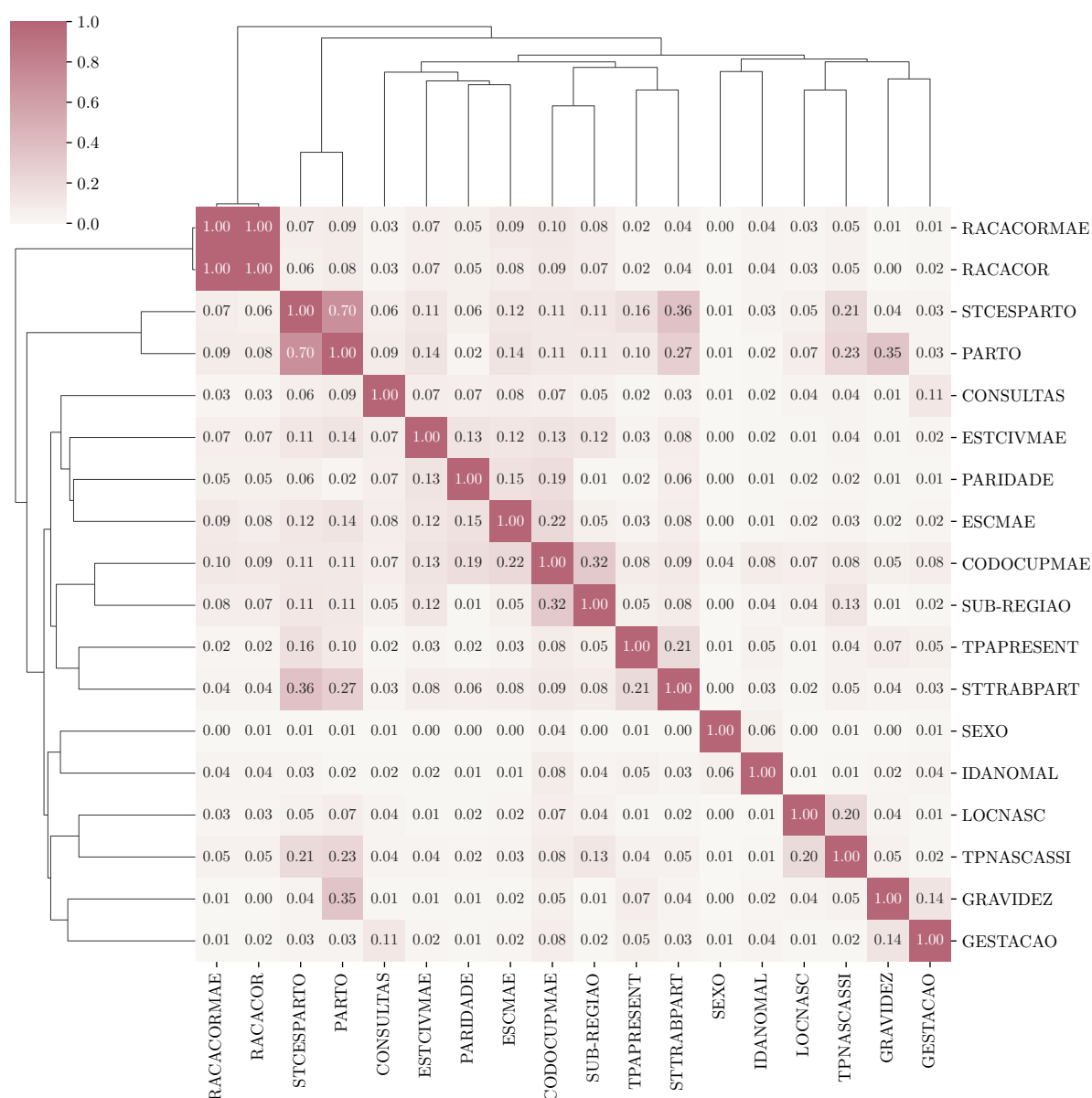


Fonte: Elaborado pela autora (2025).

A Figura 64 apresenta a matriz de correlação de Pearson das variáveis numéricas. A análise evidenciou a presença de uma correlação forte ($r > 0,70$) entre as variáveis QTDPARTNOR, QTDFILVIVO e QTDGESTANT, além de uma correlação moderada ($0.50 < r \leq 0,70$) entre as variáveis PESO, SEMAGESTAC, QTDPARTNOR, QTDGESTANT, QTDFILMORT, IDADEMAE e IDADEPAI.

Por outro lado, a Figura 65 apresenta a matriz do coeficiente V de Cramer (v_c), evidenciando uma associação perfeita significativa ($v_c = 1,00$) entre as variáveis RACACORMAE e RACACOR. Além disso, foi observada uma associação moderada ($v_c \leq 0,70$) entre essas variáveis STCESPARTO e PARTO.

Figura 65 – Matriz de coeficiente V de Cramér



Fonte: Elaborado pela autora (2025).

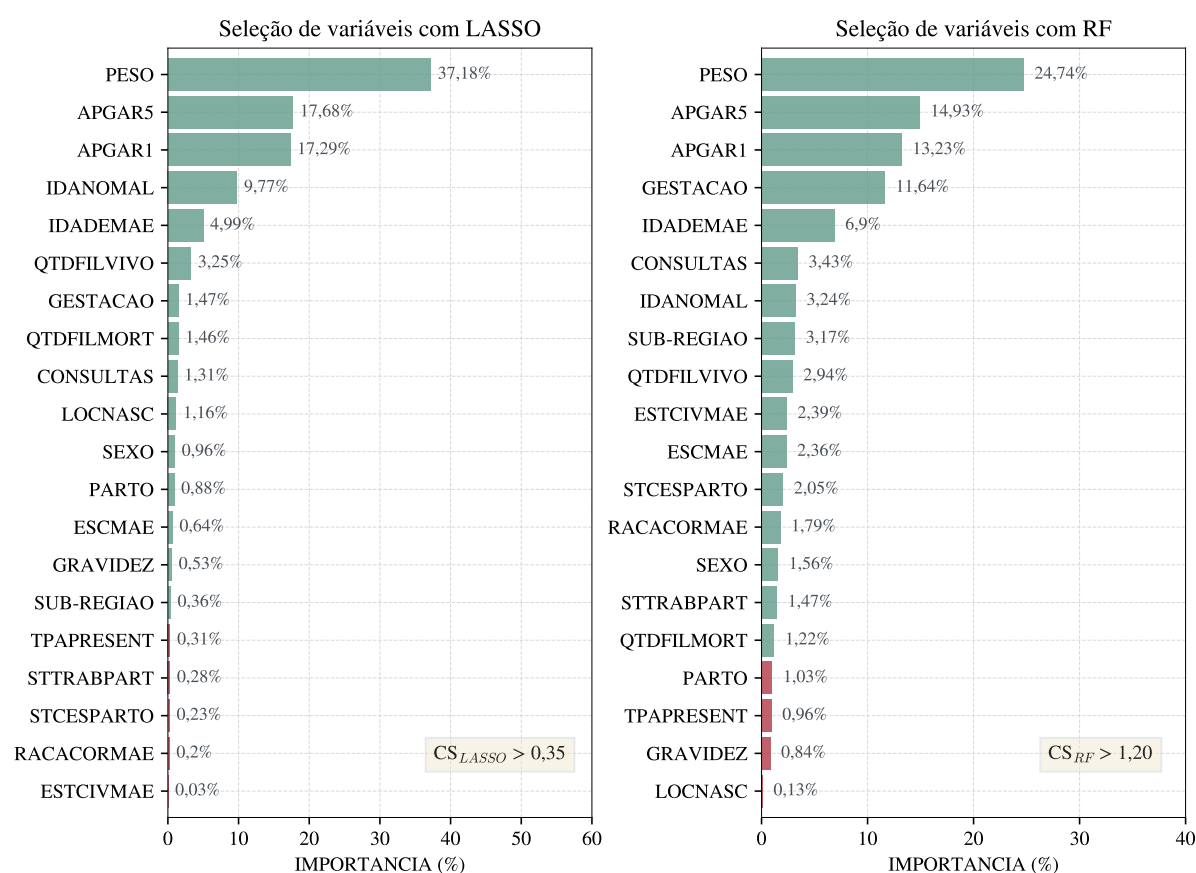
Com base nos resultados apresentados nas Figuras 64 e 65, juntamente com a análise da qua-

lidade dos dados descrita no Apêndice H, foi possível realizar uma exclusão de algumas variáveis do banco de dados. A variável RACACOR foi removida por apresentar informações equivalentes à variável RACACORMAE. Da mesma forma, as variáveis QTDGESTANT, QTDPARTNOR e STCESPARTO foram descartadas devido à colinearidade significativa ($r > 0,70$ e $v_c > 0,70$) com outras variáveis. A variável SEMGESTAC foi excluída por ser similar à variável GESTACAO, ambas representando informações sobre a idade gestacional. Adicionalmente, a variável TPNASCASSI foi desconsiderada por conter dados disponíveis apenas a partir de 2013 (Tabela 24). Por último, as variáveis IDADEPAI, QTDPARTCES, QTDPARTNOR, PARIDADE e CODOCUPMAE foram eliminadas devido à alta proporção de valores ausentes, superior a 30%.

De forma resumida, o CDC empregado para o treinamento dos modelos binário e multi-classe foi composto pelas variáveis de entrada: LOCNASC, ESTCIVMAE, TPAPRESENT, SEXO, PESO, APGAR1, APGAR5, RACACORMAE, IDANOMAL, ESCMAE, STCESPARTO, IDADEMAE, CONSULTAS, GRAVIDEZ, STTRABPART, SUB-REGIAO, QTDFILVIVO, QTDFILMORTO, GESTACAO e PARTO.

Em relação ao CDS, as variáveis que compuseram este conjunto de dados foram selecionadas com base na análise de importância de variáveis utilizando os métodos LASSO e RF.

Figura 66 – Importância de variáveis com LASSO e RF para modelo binário



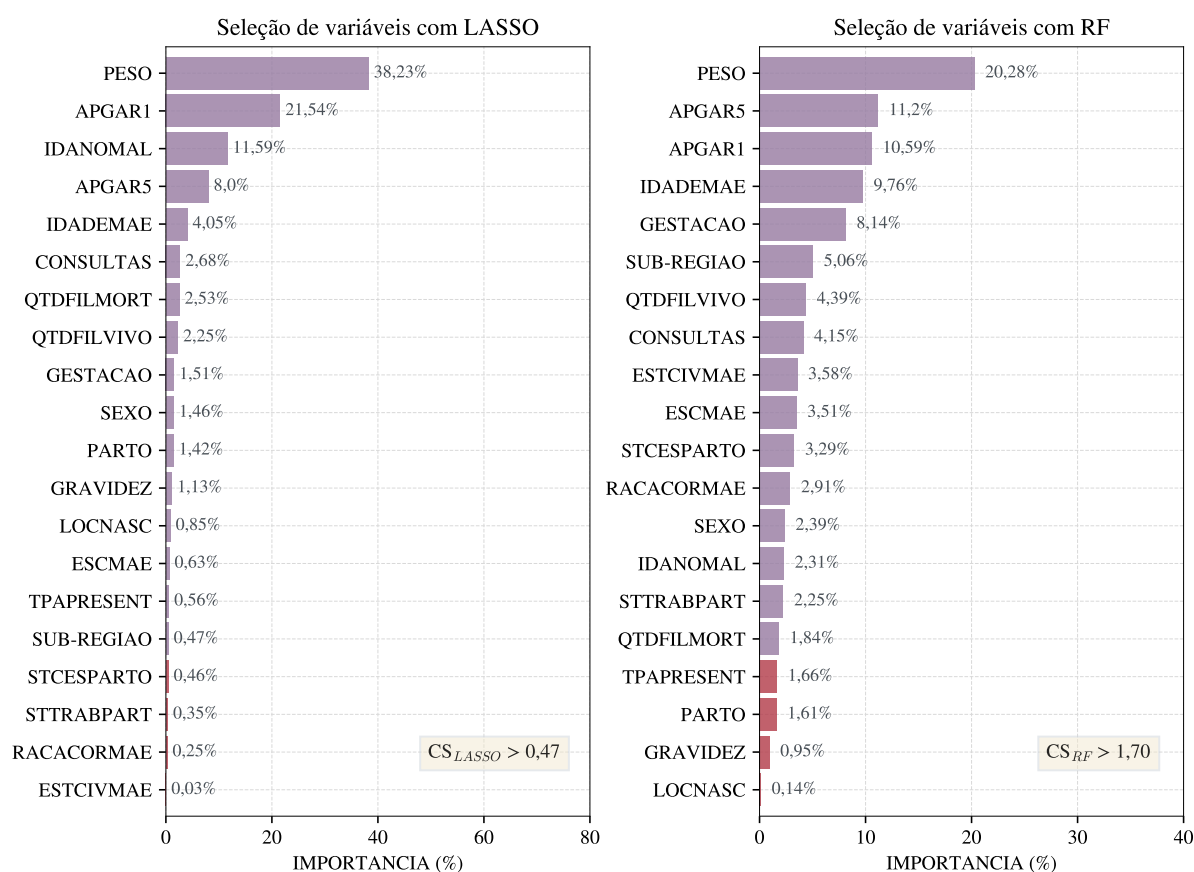
Fonte: Elaborado pela autora (2025).

A Figura 66 apresenta os resultados da análise de importância para o modelo binário,

mostrando que, com ambas as técnicas, a variável mais importante foi PESO, seguida por APGAR5 e APGAR1. Outras variáveis que também mostraram alta relevância foram IDADEMAE e IDANOMAL com LASSO, e GESTACAO com RF. Essas variáveis evidenciaram uma relevância significativa na predição da mortalidade infantil.

Da mesma forma, conforme mostrado na Figura 67, a variável PESO também foi a característica de maior importância para o modelo multiclasse. Outras variáveis relevantes foram, para o LASSO, APGAR1, IDANOMAL e APGAR5, enquanto, com a técnica RF, as variáveis mais importantes foram APGAR5, APGAR1 e IDADEMAE. Essas variáveis coincidem com as obtidas na análise binária, mas com diferentes graus de importância. Ao contrastar esses resultados com alguns estudos da literatura (Beluzo *et al.*, 2020; Batista *et al.*, 2021), observou-se similaridade com outros trabalhos que apontam essas variáveis como as de maior impacto nos modelos de predição.

Figura 67 – Importância de variáveis com LASSO e RF para modelo multiclasse



Fonte: Elaborado pela autora (2025).

Os resultados da análise de importância de variáveis, apresentados nas Figuras 66 e 67, indicam que a característica peso foi a de maior relevância, devido ao seu impacto significativo e à sua relação com complicações do recém-nascido, assim como com a idade gestacional (Paixao *et al.*, 2021; Bernardino *et al.*, 2022; Rodríguez *et al.*, 2023; Rodríguez *et al.*, 2024). Da mesma forma, as variáveis *apgar* demonstraram um impacto significativo na identificação do risco de

mortalidade, uma vez que refletem a saúde imediata do recém-nascido ao nascimento e sua adaptação ao ambiente externo (Santos *et al.*, 2012; Vilanova *et al.*, 2019).

Por outro lado, a presença de anomalias congênitas também está diretamente associada ao risco de mortalidade, pois esses problemas podem afetar significativamente o desenvolvimento físico e funcional do recém-nascido (WHO, 2018; Hug *et al.*, 2019). Já a idade materna está relacionada como um fator de risco tanto para o desenvolvimento do recém-nascido quanto para o parto, considerando que mães adolescentes ou em idade avançada apresentam maior probabilidade de enfrentar complicações (Santos *et al.*, 2012; Gaiva; Fujimori; Sato, 2016). Além disso, a idade materna também está associada a fatores que podem influenciar diretamente a mortalidade infantil, como baixo peso ao nascer, prematuridade e anomalias congênitas (Costa-Nobre *et al.*, 2021; Rodríguez *et al.*, 2023; Rodríguez *et al.*, 2024).

Nesta sequência, a identificação de características importantes com LASSO e RF permitiu selecionar variáveis relevantes relacionadas à mortalidade, tanto em termos lineares quanto não lineares. Para os modelos binários, o CDS foi composto pelas variáveis que apresentaram significância em ambas as técnicas $CS_{LASSO} > 0,35$ e $CS_{RF} > 1,20$ (Figura 66). Da mesma forma, para os modelos multiclasse, foram consideradas as variáveis com $CS_{LASSO} > 0,47$ y $CS_{RF} > 1,70$ (Figura 67). Essa abordagem reforçou a robustez do conjunto ao incluir apenas variáveis que demonstraram relevância simultânea nos dois métodos.

De maneira concisa, o CDS empregado para o treinamento dos modelos binário e multiclasse foi composto pelas variáveis de entrada: SEXO, PESO, IDANOMAL, SUB-REGIAO, APGAR1, APGAR5, GESTACAO, QTDFILMORT, QTDFILVIVO, CONSULTAS, IDADEMAE e ESCMAE.

Além disso, a Tabela 15 apresenta a quantidade de instâncias em cada subconjunto (treino/teste e validação) para os modelos binário e multiclasse. Observa-se que a classe "Vivo" possui a maior frequência em ambas as abordagens, indicando que os dados utilizados no treinamento estão desbalanceados.

Tabela 15 – Distribuição das instâncias nos conjuntos de treino/teste e validação para os modelos binário e multiclasse

Modelo	Tipo	Classe	Rótulos	Quantidade
Binário	Treino/teste	Vivo	0	329.186
		Óbito	1	3.726
	Validação	Vivo	0	30.595
		Óbito	1	302
Multiclasse	Treino/teste	Vivo	0	329.186
		Neonatal	1	2.688
		Pós-neonatal	2	1.038
	Validação	Vivo	0	30.595
		Neonatal	1	227
		Pós-neonatal	2	75

Fonte: Elaborado pela autora (2025).

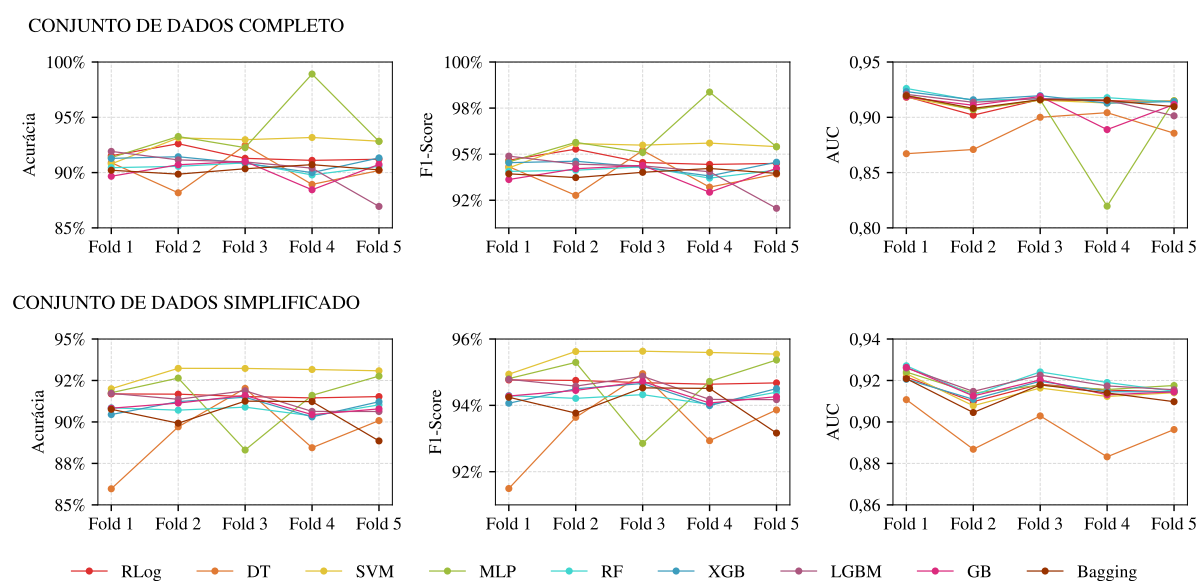
4.3.2 Avaliação do desempenho dos modelos de classificação

No treinamento dos modelos de predição utilizando ML, tanto para os modelos binários quanto para os multiclasse, observou-se que todas as técnicas testadas apresentaram um bom desempenho. No entanto, alguns resultados se destacaram mais do que outros, tanto no cenário de treinamento com o CDC quanto no treinamento com o CDS. A seguir, serão apresentados e discutidos os resultados obtidos para os modelos binários e multiclasse, com uma análise detalhada do desempenho em cada cenário.

• *Modelo Binário:*

A Figura 68 apresenta os resultados das métricas de acurácia, F1-score e AUC em cada dobra da validação cruzada 5-fold, observando-se o comportamento e desempenho de cada técnica de ML no treinamento do modelo binário com o CDC e o CDS.

Figura 68 – Resultados de acurácia, F1-score e AUC obtidos por meio de validação cruzada 5-fold para todas as técnicas de ML no modelo binário



Fonte: Elaborado pela autora (2025).

Conforme a Figura 68, para o CDC, independentemente das técnicas utilizadas, os valores da métrica de acurácia variaram entre 86,94% e 98,91%, o F1-score apresentou valores entre 92,06% e 98,37%, e a métrica AUC oscilou entre 81,96% e 92,61%.

Entre os modelos que se destacaram pelo seu desempenho, foram os treinados com SVM e RLog, pois apresentaram menor variabilidade entre as dobras, indicando maior consistência e equilíbrio na relação entre precisão e revocação (sensibilidade), além de mostrarem um bom desempenho geral na discriminação entre as classes "óbito" e "vivo". Por sua vez, os modelos com DT, MLP, GB e LGBM, embora apresentem métricas aceitáveis, possuem desempenho inferior e instável em cada dobra.

Por outro lado, os resultados mostraram que, ao utilizar apenas as variáveis mais importantes no CDS, os modelos treinados mantiveram desempenho muito próximo aos valores das métricas dos modelos treinados com o CDC. A acurácia variou entre 85,97% e 93,23%, a medida F1-score entre 91,49% e 95,63%, e a AUC entre 88,32% e 92,71%. Da mesma forma, os modelos treinados com SVM e RLog mostraram uma leve melhoria nas métricas de avaliação, além de apresentarem uma amplitude estreita de variação entre as dobras, indicando resultados consistentes em diferentes subconjuntos de dados.

Os modelos com as técnicas GB e LGBM apresentaram melhor desempenho com o CDS, possivelmente porque a redução de variáveis permitiu que as técnicas se concentrassem em capturar as interações não lineares e padrões complexos das variáveis mais relevantes, tornando os modelos mais eficazes e com melhor performance. Já os modelos treinados com as técnicas DT, MLP e Bagging apresentaram maior variabilidade e menor desempenho, mostrando que, para essas técnicas, a simplificação das variáveis reduz a diversidade de informações, afetando negativamente seu desempenho, principalmente no caso do DT.

Tabela 16 – Valores médios de desempenho dos modelos binários

Técnicas	Acur^a (IC^b_{95%})	Prec^c (IC_{95%})	Rev^d (IC_{95%})	F1^e (IC_{95%})	AUC^f (IC_{95%})
Conjunto de dados completo					
RLog	91,55 (±0, 76)	98,76 (±0, 03)	91,55 (±0, 76)	94,69 (±0, 42)	0,913 (±0, 008)
DT	90,13 (±2, 09)	98,74 (±0, 03)	90,13 (±2, 09)	93,88 (±1, 18)	0,886 (±0, 021)
SVM	92,58 (±1, 26)	98,76 (±0, 02)	92,58 (±1, 26)	95,27 (±0, 71)	0,914 (±0, 005)
MLP	93,72 (±3, 71)	98,58 (±0, 51)	93,72 (±3, 71)	95,82 (±1, 84)	0,895 (±0, 053)
RF	90,45 (±0, 50)	98,76 (±0, 02)	90,45 (±0, 50)	94,07 (±0, 28)	0,918 (±0, 006)
XGB	90,98 (±0, 74)	98,76 (±0, 02)	90,98 (±0, 74)	94,37 (±0, 41)	0,917 (±0, 005)
LGBM	90,28 (±2, 41)	98,76 (±0, 02)	90,28 (±2, 41)	93,97 (±1, 37)	0,914 (±0, 009)
GB	90,11 (±1, 32)	98,76 (±0, 02)	90,11 (±1, 32)	93,88 (±0, 75)	0,910 (±0, 015)
Bagging	90,28 (±0, 38)	98,76 (±0, 02)	90,28 (±0, 38)	93,97 (±0, 22)	0,914 (±0, 006)
Conjunto de dados simplificado					
RLog	91,58 (±0, 13)	98,76 (±0, 02)	91,58 (±0, 13)	94,70 (±0, 07)	0,916 (±0, 005)
DT	89,25 (±2, 78)	98,75 (±0, 02)	89,25 (±2, 78)	93,38 (±1, 59)	0,896 (±0, 014)
SVM	92,94 (±0, 66)	98,76 (±0, 02)	92,94 (±0, 66)	95,47 (±0, 37)	0,915 (±0, 007)
MLP	91,42 (±2, 25)	98,76 (±0, 02)	91,42 (±2, 25)	94,61 (±1, 27)	0,918 (±0, 004)
RF	90,77 (±0, 31)	98,76 (±0, 02)	90,77 (±0, 31)	94,25 (±0, 18)	0,920 (±0, 007)
XGB	90,94 (±0, 65)	98,76 (±0, 02)	90,94 (±0, 65)	94,34 (±0, 37)	0,916 (±0, 005)
LGBM	91,25 (±0, 73)	98,77 (±0, 02)	91,25 (±0, 73)	94,52 (±0, 41)	0,919 (±0, 006)
GB	90,96 (±0, 54)	98,76 (±0, 02)	90,96 (±0, 54)	94,36 (±0, 30)	0,917 (±0, 007)
Bagging	90,40 (±1, 27)	98,76 (±0, 02)	90,40 (±1, 27)	94,04 (±0, 72)	0,913 (±0, 008)

(a) Acur: Acurácia, (b) IC: Intervalo de confiança, (c) Prec: Precisão, (d) Rev: Revocação, (e) F1: F1-Score, (f) AUC: Área sob a Curva.

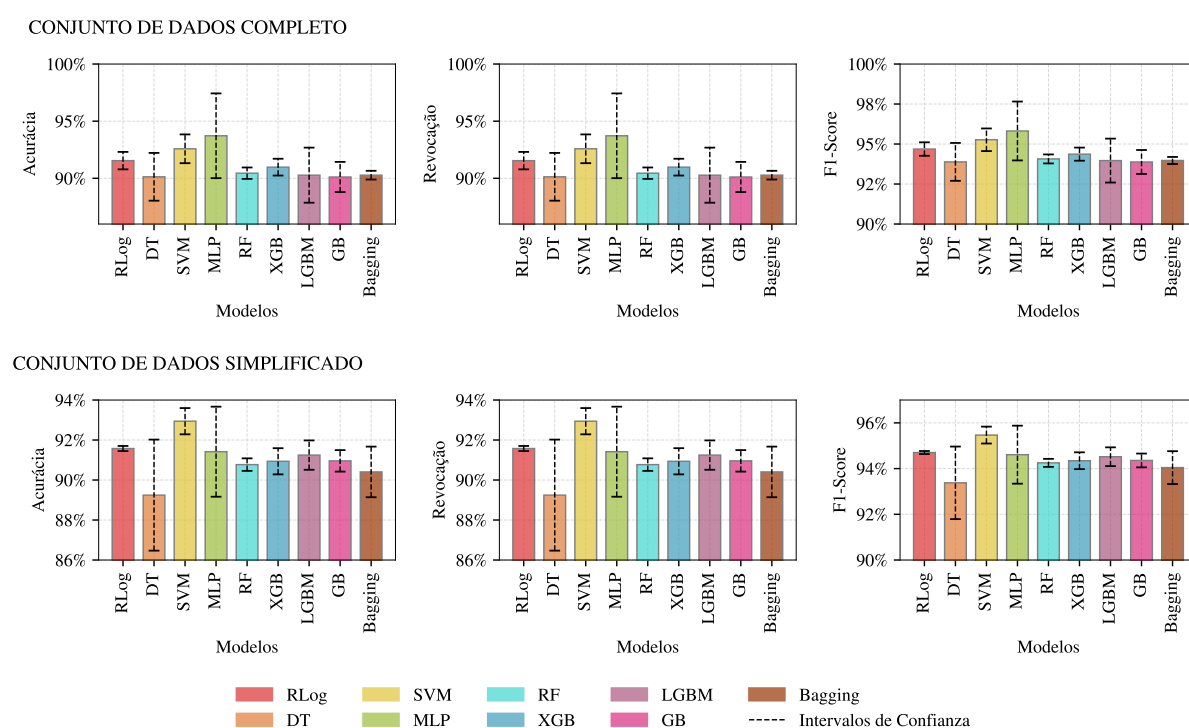
Fonte: Elaborado pela autora (2025).

A Tabela 16 e a Figura 69 apresentam os valores médios das métricas de desempenho dos modelos binários, calculados a partir dos resultados obtidos nas 5 dobras da validação cruzada, nos conjuntos de dados completo e simplificado. As métricas apresentadas

incluem acurácia, precisão, revocação, F1-score e AUC, com seus respectivos intervalos de confiança de 95%.

A verificação dos valores médios das métricas de avaliação mostradas na Figura 69 e na Tabela 16 confirmou que as técnicas testadas apresentaram desempenho consistente, superando 85% nas métricas (acurácia, precisão, revocação, F1-score e AUC). Observou-se uma diferença marginal e uma leve melhora no desempenho dos modelos treinados com o CDC e o CDS, sugerindo que a simplificação dos dados manteve a eficácia dos modelos, ao mesmo tempo em que reduziu a complexidade computacional.

Figura 69 – Valores médios de desempenho dos modelos binários

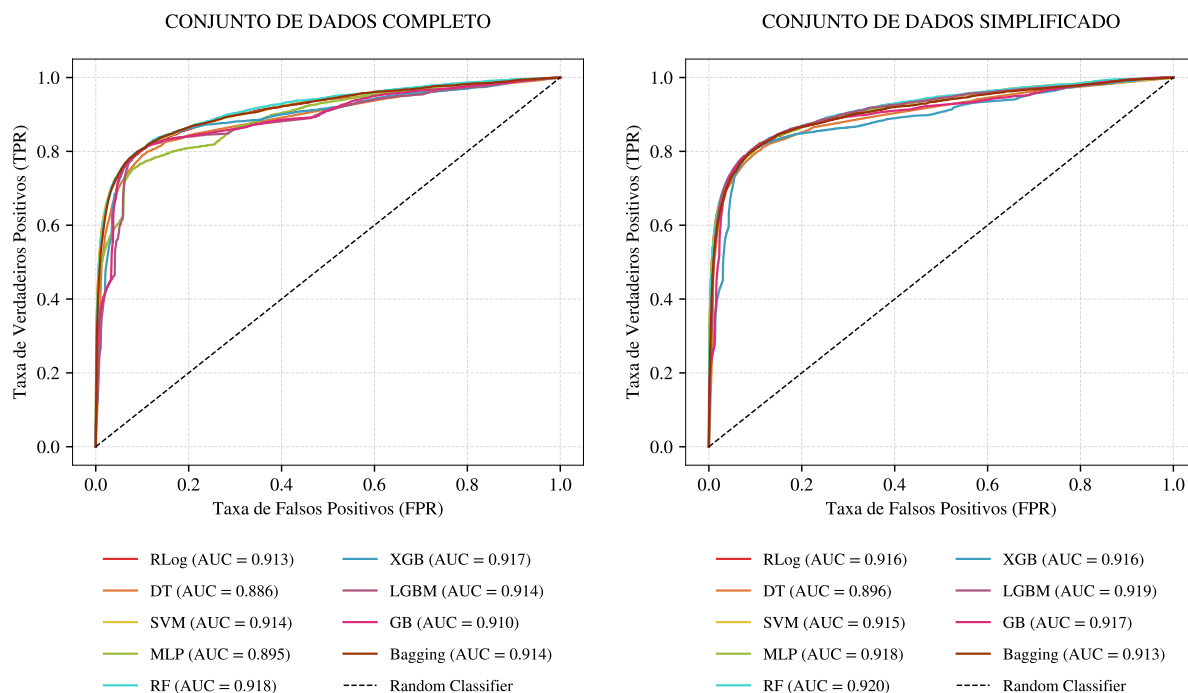


Fonte: Elaborado pela autora (2025).

Essa mesma diferença é observada na Figura 70, que apresenta as curvas ROC dos modelos treinados com o CDC e o CDS. De maneira semelhante aos valores médios das métricas de avaliação, as curvas ROC indicaram uma leve melhora no desempenho da maioria dos modelos treinados com o CDS.

Essa melhoria também pode ser corroborada na Tabela 17, que apresenta a eficiência média das métricas de avaliação dos modelos de classificação binária treinados. A fim de padronizar todas as métricas de avaliação, o valor do AUC foi ajustado para o intervalo de 0 a 100.

Figura 70 – Curvas ROC dos modelos de classificação binária



Fonte: Elaborado pela autora (2025).

Tabela 17 – Eficiência média dos modelos de classificação binária

Conjunto de dados completo			Conjunto de dados simplificado		
Modelo	Média	Desvio Padrão	Modelo	Média	Desvio Padrão
RLog	93,57	3,22	RLog	93,64	3,17
DT	92.29	4.10	DT	92.04	4.13
SVM	94.11	2.96	SVM	94.32	2.87
MLP	94.28	3.31	MLP	93.60	3.18
RF	93.11	3.49	RF	93.30	3.37
XGB	93.36	3.33	XGB	93.32	3.35
LGBM	92.93	3.59	LGBM	93.54	3.22
GB	92.77	3.69	GB	93.35	3.33
Bagging	92.93	3.59	Bagging	92.99	3.55

Fonte: Elaborado pela autora (2025).

Assim, com base nas evidências quantitativas e qualitativas extraídas das Figuras 69 e 70 e das Tabelas 16 e 17, pode-se inferir que a maioria dos modelos treinados com CDS apresentaram melhoria no desempenho. No entanto, os modelos treinados com DT ($Acur = 89,25\%$, $Prec = 98,75\%$, $Rev = 89,25\%$, $F1 = 93,38\%$ e $AUC = 0,896$) e Bagging ($Acur = 90,40\%$, $Prec = 98,76\%$, $Rev = 90,40\%$, $F1 = 94,04\%$ e $AUC = 0,913$) apresentaram elevados valores no desvios padrão ($S_{DT} = 4,13$ e $S_{Bagging} = 3,55$), o que sugere que seu desempenho é mais sensível, além de apresentarem menor consistência e maior variação, tanto entre as dobras quanto na eficiência média, sendo, assim, os modelos com desempenho inferior.

No caso do modelo com MLP ($Acur = 91,42\%$, $Prec = 98,76\%$, $Rev = 91,42\%$, $F1 = 94,61\%$ e $AUC = 0,918$), observou-se o melhor desempenho com as variáveis completas, mas com um desvio padrão médio ($S_{MLP} = 3,31$). Embora os modelos com MLP tenham apresentado valores médios superiores aos de outras técnicas, houve uma alta variação no desempenho entre as dobras da validação cruzada. Essas flutuações nos resultados podem refletir a sensibilidade do modelo a diferentes subconjuntos dos dados, o que o diferencia dos outros modelos baseados em árvores, bem como dos modelos como SVM e RLog.

Os modelos treinados com as técnicas de RF, XGB, LGBM e GB apresentaram desempenho médio, tanto com o CDC quanto com o CDS. Embora esses modelos tenham apresentado métricas de avaliação elevadas e competitivas, com consistência e estabilidade médias, destacaram-se entre eles os modelos treinados com RF e LGBM, os quais obtiveram valores de AUC superiores aos de todos os outros modelos. Adicionalmente, cabe ressaltar que o modelo treinado com LGBM demonstrou uma melhoria substancial em suas métricas de desempenho ao ser treinado com o CDS, apresentando maior estabilidade e posicionando-se como o terceiro modelo com os melhores valores nas métricas de avaliação ($Acur = 91,25\%$, $Prec = 98,77\%$, $Rev = 91,25\%$, $F1 = 94,52\%$ e $AUC = 0,919$).

Por outro lado, os modelos que apresentaram maior consistência e desempenho foram os treinados com as técnicas de RLog e SVM, tanto no cenário de treinamento com variáveis completas quanto no cenário em que foi utilizado o conjunto de dados com variáveis simplificadas. O modelo com SVM simplificado demonstrou o menor desvio padrão ($S_{SVM} = 2,87$), indicando maior estabilidade, tanto nas dobras da validação cruzada quanto nas métricas de avaliação média ($Acur = 92,94\%$, $Prec = 98,76\%$, $Rev = 92,94\%$, $F1 = 95,47\%$ e $AUC = 0,915$). O modelo treinado com RLog simplificado também exibiu baixo desvio padrão ($S_{RLog} = 3,17$), com os valores médios das métricas de desempenho ($Acur = 91,58\%$, $Prec = 98,76\%$, $Rev = 91,58\%$, $F1 = 94,70\%$ e $AUC = 0,916$) ligeiramente superiores às dos outros modelos.

Portanto, com base nos resultados obtidos nos análises, as técnicas de SVM e RLog apresentaram um equilíbrio entre alto desempenho, estabilidade e robustez, tanto utilizando o CDC quanto o CDS. Nos dois cenários apresentados observou-se um aumento no desempenho com o modelo simplificado. Especialmente, a técnica de SVM, quando treinada com menos variáveis, apresentou uma redução em sua complexidade, demonstrando desempenho superior e maior estabilidade em comparação com os outros modelos.

Além disso, para contrastar os resultados obtidos com as métricas e análises apresentadas, foi aplicado o teste de McNemar aos modelos que obtiveram as melhores métricas de avaliação (SVM e RLog simplificados), com o objetivo de avaliar a existência de diferenças entre seus erros de classificação. Os resultados do teste de McNemar para os modelos de

SVM e RLog são apresentados na Tabela 18. O valor crítico foi obtido a partir da tabela de distribuição qui-quadrado (ver Anexo F).

Tabela 18 – Resultados do teste de McNemar dos modelos de classificação binária

Modelos	X^{2a}	$X_{0,05;1}^b$	Descrição
RLog simplificado	2.601,21	3,84	Rejeite a H_0
SVM simplificado			

(a) X^2 : Estatística de McNemar, (b) $X_{0,05;1}$: Valor crítico com $\alpha = 0,05$ significância e 1 grau de liberdade.

Fonte: Elaborado pela autora (2025).

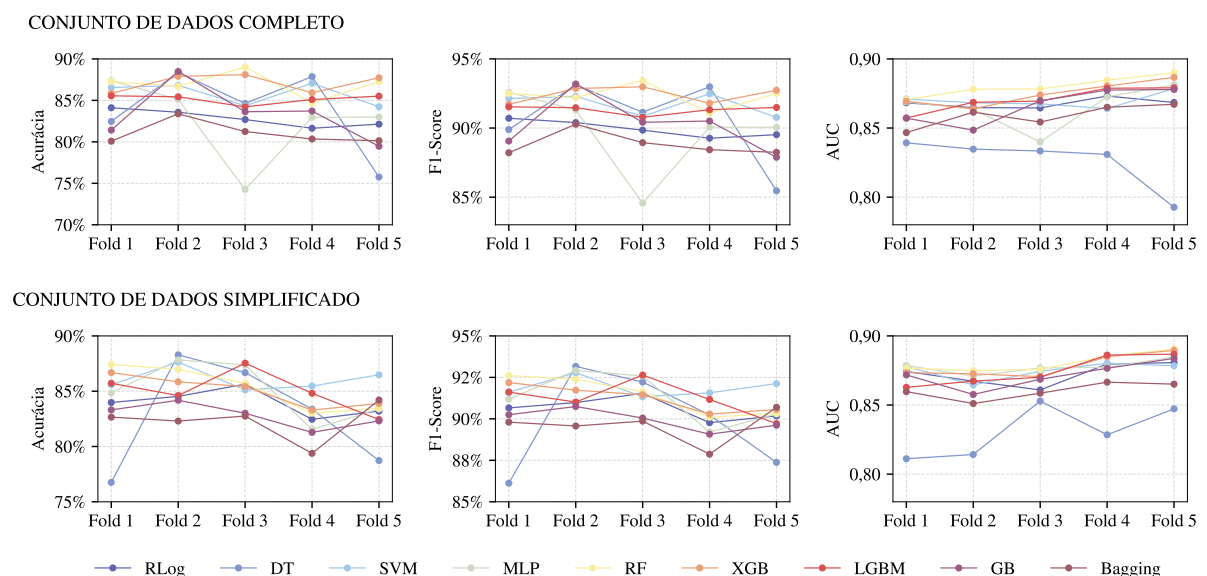
A Tabela 18 mostrou que o valor do estatístico calculado de McNemar foi inferior ao valor crítico ($X^2 < X_{0,05;1}$), o que, para um nível de significância de 5%, levou à rejeição da hipótese nula ($H_0: error_{M1} = error_{M2}$). Este resultado sugere a presença de diferenças estatisticamente significativas entre os erros de classificação dos modelos SVM e RLog.

Assim, ao priorizar a estabilidade e a generalização dos modelos, e considerando os resultados obtidos com o teste de McNemar, a técnica de SVM demonstrou ter capturado melhor os padrões complexos dos dados. Além disso, o SVM mostrou-se mais robusto em relação a dados não vistos, pois sua performance não estava limitada a um conjunto específico de dados, tornando-o menos suscetível ao sobreajuste e mais confiável em cenários reais.

• Modelo Multiclasse:

Em relação ao modelo multiclasse, a Figura 71 apresenta os resultados das métricas de acurácia, F1-score e AUC das dobras da validação cruzada 5-fold.

Figura 71 – Resultados de acurácia, F1-score e AUC obtidos por meio de validação cruzada 5-fold para todas as técnicas de ML no modelo multiclasse



Fonte: Elaborado pela autora (2025).

De forma semelhante ao modelo binário, foram observadas diferenças no desempenho entre os modelos treinados com o CDC e o CDS. Para o CDC, os valores da acurácia na validação cruzada variaram entre 74,26% e 89,02%, a medida F1-score oscilou entre 84,56% e 93,45%, e a AUC apresentou valores entre 79,27% e 88,99%.

No cenário de treinamento com as variáveis completas, os valores das métricas de desempenho da maioria das técnicas de ML apresentaram variações ao longo das dobras da validação cruzada. As técnicas que demonstraram maior estabilidade foram SVM, RLog, RF, XGB, GB, LGBM e Bagging. Em particular, os modelos de SVM, RF e XGB não apenas se destacaram pela estabilidade, mas também apresentaram as melhores métricas de desempenho. Por outro lado, os modelos com maior variação nas suas métricas de desempenho foram os treinados com as técnicas de DT e MLP.

Os resultados das métricas de avaliação dos modelos treinados com o CDS mostraram que alguns dos modelos apresentaram um desempenho ligeiramente superior ao dos modelos treinados com as variáveis completas. Neste caso, a acurácia variou entre 76,76% e 88,29%, a métrica de F1-score variou entre 86,12% e 93,16%, e a medida AUC oscilou entre 81,12% e 89,05%.

As métricas de avaliação dos modelos treinados com SVM, RF e XGB apresentaram menor variação entre as dobras, evidenciando maior estabilidade e equilíbrio no desempenho. No entanto, o RF e XGB simplificado apresentou uma leve queda em suas métricas de avaliação, ao contrário do SVM. Em contrapartida, os modelos treinados com as técnicas de DT e MLP mostraram maior variação entre as dobras da validação cruzada, além de uma redução nos valores de algumas das métricas avaliadas.

Assim, a Figura 72 e a Tabela 19 apresentam os valores médios das métricas de desempenho dos modelos multiclasse, calculados a partir das cinco dobras da validação cruzada, tanto para os modelos completos quanto para os modelos simplificados.

Tabela 19 – Valores médios de desempenho dos modelos multiclasse

(continua)					
Técnicas	Acur^a (IC^b_{95%})	Prec^c (IC_{95%})	Rev^d (IC_{95%})	F1^e (IC_{95%})	AUC^f (IC_{95%})
Conjunto de dados completo					
RLog	82,83 (±0, 76)	98,87 (±0, 02)	82,83 (±1, 26)	89,95 (±0, 74)	0,868 (±0, 004)
DT	83,82 (±2, 09)	98,91 (±0, 16)	83,82 (±6, 35)	90,52 (±3, 89)	0,826 (±0, 024)
SVM	85,80 (±1, 26)	98,87 (±0, 02)	85,80 (±1, 73)	91,71 (±1, 03)	0,870 (±0, 007)
MLP	82,54 (±3, 71)	98,88 (±0, 05)	82,54 (±6, 18)	89,72 (±3, 81)	0,865 (±0, 019)
RF	86,99 (±0, 50)	98,86 (±0, 05)	86,99 (±1, 92)	92,35 (±1, 06)	0,880 (±0, 009)
XGB	87,10 (±0, 74)	98,86 (±0, 04)	87,10 (±1, 39)	92,42 (±0, 76)	0,875 (±0, 011)
LGBM	85,17 (±2, 41)	98,87 (±0, 03)	85,17 (±0, 69)	91,32 (±0, 39)	0,871 (±0, 011)
GB	83,35 (±1, 32)	98,86 (±0, 06)	83,35 (±4, 18)	90,21 (±2, 46)	0,866 (±0, 016)
Bagging	81,04 (±0, 38)	98,84 (±0, 04)	81,04 (±1, 73)	88,82 (±1, 07)	0,859 (±0, 010)

(a) Acur: Acurácia, (b) IC: Intervalo de confiança, (c) Prec: Precisão, (d) Rev: Revocação, (e) F1: F1-Score, (f) AUC: Área sob a Curva.

Fonte: Elaborado pela autora (2025).

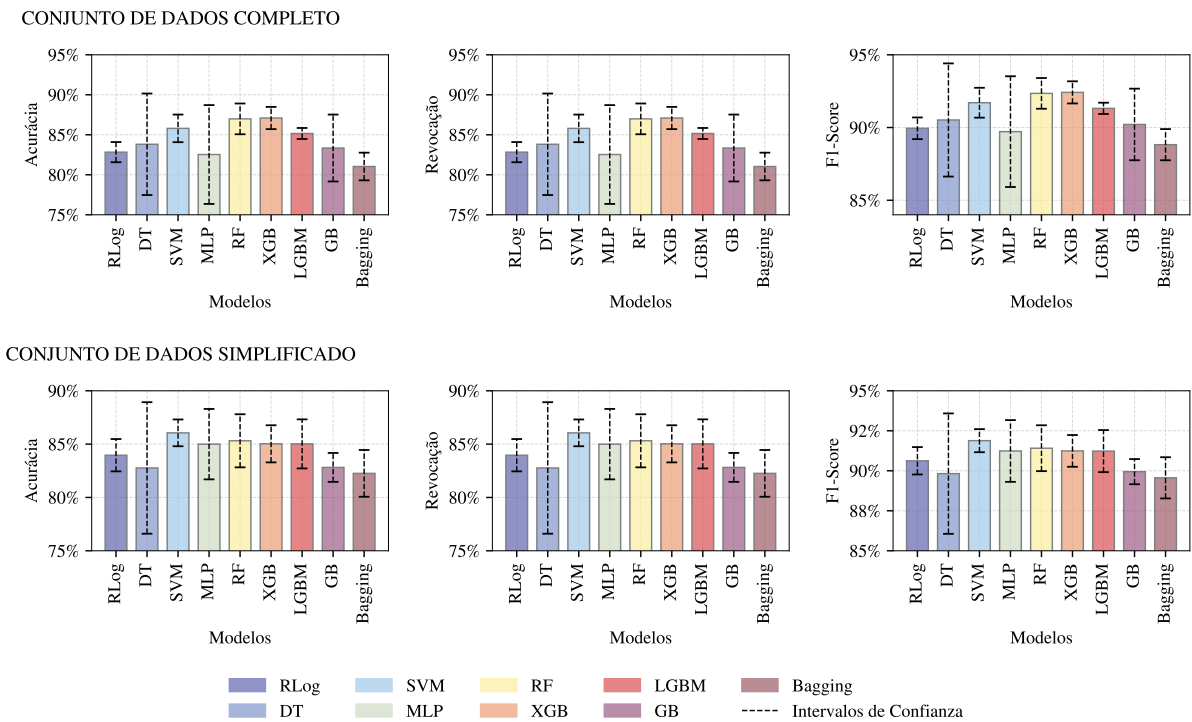
Tabela 19 – Valores médios de desempenho dos modelos multiclasse

(conclusão)					
Técnicas	Acur ^a (IC ^b _{95%})	Prec ^c (IC _{95%})	Rev ^d (IC _{95%})	F1 ^e (IC _{95%})	AUC ^f (IC _{95%})
Conjunto de dados simplificado					
RLog	83,96 (±0, 13)	98,89 (±0, 05)	83,96 (±1, 52)	90,63 (±0, 86)	0,873 (±0, 010)
DT	82,77 (±2, 78)	98,83 (±0, 05)	82,77 (±6, 16)	89,82 (±3, 76)	0,831 (±0, 023)
SVM	86,06 (±0, 66)	98,91 (±0, 02)	86,06 (±1, 26)	91,88 (±0, 72)	0,875 (±0, 008)
MLP	85,00 (±2, 25)	98,90 (±0, 03)	85,00 (±3, 30)	91,24 (±1, 93)	0,878 (±0, 006)
RF	85,31 (±0, 31)	98,89 (±0, 04)	85,31 (±2, 49)	91,41 (±1, 43)	0,881 (±0, 009)
XGB	85,03 (±0, 65)	98,88 (±0, 05)	85,03 (±1, 74)	91,24 (±0, 99)	0,878 (±0, 011)
LGBM	85,02 (±0, 73)	98,88 (±0, 06)	85,02 (±2, 30)	91,23 (±1, 31)	0,875 (±0, 014)
GB	82,82 (±0, 54)	98,89 (±0, 04)	82,82 (±1, 35)	89,95 (±0, 79)	0,872 (±0, 012)
Bagging	82,26 (±1, 27)	98,84 (±0, 05)	82,26 (±2, 19)	89,56 (±1, 29)	0,860 (±0, 008)

(a) Acur: Acurácia, (b) IC: Intervalo de confiança, (c) Prec: Precisão, (d) Rev: Revocação, (e) F1: F1-Score, (f) AUC: Área sob a Curva.

Fonte: Elaborado pelo autor (2025).

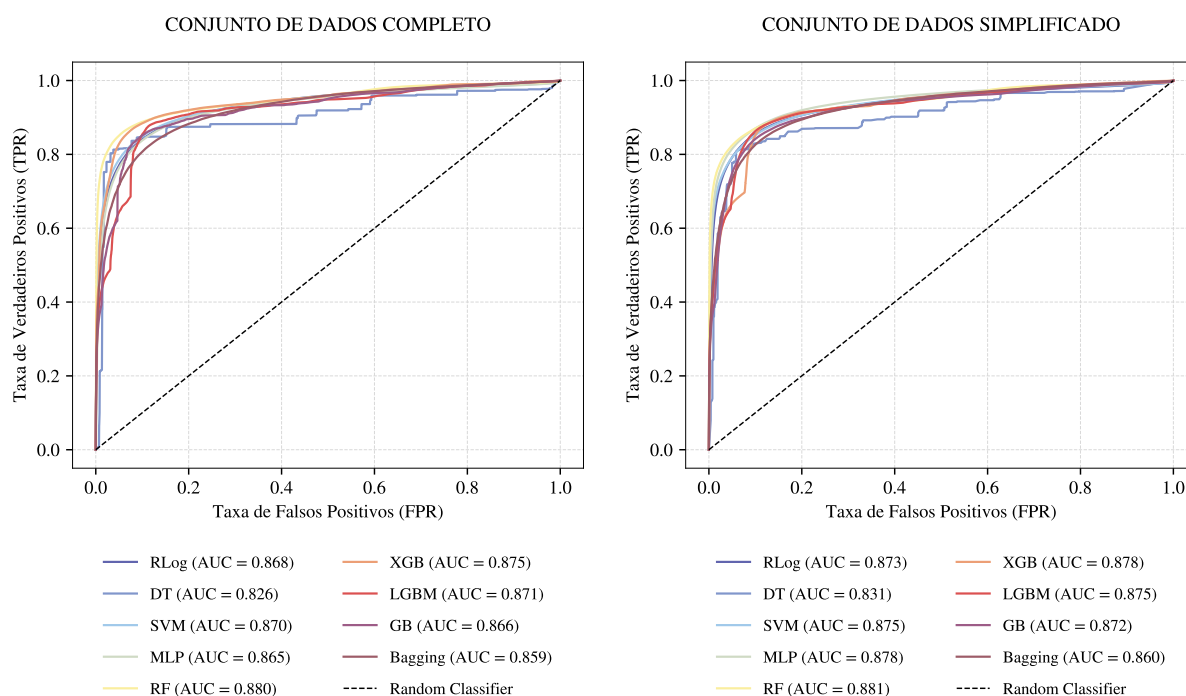
Figura 72 – Valores médios de desempenho dos modelos multiclasse



Fonte: Elaborado pela autora (2025).

Adicionalmente, a Figura 73 apresenta a curva dos modelos multiclasse, treinados com o CDC e o CDS.

Figura 73 – Curvas ROC dos modelos de classificação multiclasse



Fonte: Elaborado pela autora (2025).

Os resultados apresentados na Figura 72 e Tabela 19 corroboraram os achados da análise das métricas por dobra da validação cruzada (Figura 71), tanto para os conjuntos de dados CDC quanto CDS, além de evidenciar a variabilidade dos modelos baseados em DT e MLP.

Além disso, considerando que os modelos precisam aprender a distinguir mais de duas categorias, observou-se uma leve redução na revocação nos modelos multiclasse, em comparação com os modelos binários. Essa diminuição reflete o aumento na complexidade do modelo para identificar padrões discriminatórios entre as classes. Em algumas das técnicas testadas, a redução de variáveis gerou uma melhoria na eficiência dos modelos, enquanto em outras técnicas gerou uma leve queda nos valores das métricas, indicando uma possível perda de informação, o que pode comprometer a capacidade do modelo de discriminar as classes.

As principais técnicas que apresentaram uma melhoria na eficiência dos modelos com a redução de variáveis foram RLog ($Acur = 91,58\%$, $Prec = 98,89\%$, $Rev = 83,96\%$, $F1 = 90,63\%$ e $AUC = 0,873$) e SVM. Por outro lado, as técnicas RF e XGB apresentaram melhor eficiência quando treinadas com as variáveis completas.

Essas afirmações, além de estarem representadas na Tabela 19 e na Figura 72, são corroboradas na Tabela 20 que apresenta a eficiência média das métricas de avaliação e seu desvio padrão.

Tabela 20 – Eficiência media dos modelos de classificação multiclasse

Conjunto de dados completo			Conjunto de dados simplificado		
Modelo	Média	Desvio Padrão	Modelo	Média	Desvio Padrão
RLog	88,25	6,64	RLog	88,94	6,21
DT	87,94	6,88	DT	87,45	7,04
SVM	89,83	5,61	SVM	90,08	5,48
MLP	88,04	6,77	MLP	89,58	5,81
RF	90,64	5,10	RF	89,80	5,67
XGB	90,59	5,14	XGB	89,60	5,78
LGBM	89,52	5,80	LGBM	89,52	5,81
GB	88,47	6,46	GB	88,33	6,64
Bagging	87,13	7,35	Bagging	87,79	6,88

Fonte: Elaborado pela autora (2025).

Dessa forma, pode-se afirmar que, de acordo com o observado no processo de experimentação, os modelos que apresentaram melhor desempenho foram os modelos com RF ($Acur = 86,99\%$, $Prec = 98,86\%$, $Rev = 86,99\%$, $F1 = 92,35\%$ e $AUC = 0,880$) e XGB ($Acur = 87,10\%$, $Prec = 98,86\%$, $Rev = 87,10\%$, $F1 = 92,42\%$ e $AUC = 0,875$) treinados com o CDC e o modelo com SVM ($Acur = 86,06\%$, $Prec = 98,91\%$, $Rev = 86,06\%$, $F1 = 91,88\%$ e $AUC = 0,875$) treinado com o CDS.

A Tabela 21 apresenta os resultados do teste de McNemar, que foi utilizado nos modelos de RF, SVM e XGB, fim de avaliar a existência de diferenças entre seus erros de classificação.

Tabela 21 – Resultados do teste de McNemar dos modelos de classificação multiclasse

Modelos	X^{2a}	$X_{0,05;1}^b$	Descrição
RF completo	568,43	3,84	Rejeite a H_0
SVM simplificado			
XGB completo	11,11	3,84	Rejeite a H_0
RF completo			
XGB completo	607,68	3,84	Rejeite a H_0
SVM simplificado			

(a) X^2 : Estatística de McNemar, (b) $X_{0,05;1}$: Valor crítico com $\alpha = 0,05$ significância e 1 grau de liberdade.

Fonte: Elaborado pela autora (2025).

De acordo com as métricas de avaliação, o desempenho do modelo treinado com XGB completo é ligeiramente superior, em comparação com os modelos de RF e SVM simplificado. Além disso, a Tabela 21 mostra que as previsões do modelo com XGB são significativamente diferentes em relação aos modelos comparados, observando-se uma maior diferença nas previsões do modelo SVM do que no modelo RF.

Esses resultados do teste estatístico e das métricas de desempenho reforçam a escolha da técnica de XGB completo, pois, apesar das diferenças observadas em relação aos

outros modelos, essa técnica se mostrou eficaz em termos de precisão, variabilidade, generalização e eficiência computacional.

4.3.3 Validação dos resultados

Com base no processo de experimentação e seleção das melhores técnicas para os modelos de identificação de risco, o modelo binário simplificado com SVM e o modelo multiclasse completo com XGB foram treinados novamente, utilizando todas as informações dos dados de treino/teste (Tabela 15) do período 2010 – 2019.

O Quadro 8 apresenta a melhor configuração de hiperparâmetros dos modelos SVM e XGB.

Quadro 8 – Configuração de hiperparâmetros dos modelos binário e multiclasse

Modelo/Dados	Técnica	Hiperparâmetros	Descrição
Binário/ Simplificado	SVM	C: 14,20	Parâmetro de regularização.
		kernel: linear	Função <i>kernel</i> .
		gamma: scale	Coeficiente da função <i>kernel</i> , $(1/(n^o \text{ variáveis} \times \text{Variância}(\mathbf{x})))$
		tol: 0,0005	Tolerância para critério de parada.
		random_state: 42	Controlador de aleatoriedade reprodutível.
Multiclasse/ Completo	XGB	n_estimators: 179	Número de árvores.
		max_depth: 3	Profundidade máxima da árvore.
		learning_rate: 0,0102	Taxa de atualização dos parâmetros.
		min_child_weight: 9	Peso mínimo de divisão de folha.
		gamma: 0,4641	Perda mínima para divisão de nó
		random_state: 42	Controlador de aleatoriedade reprodutível.

Fonte: Elaborado pela autora (2025).

Os modelos treinados foram avaliados utilizando as informações de validação correspondentes ao período de 2020. Na Tabela 22, são apresentados os valores das métricas de avaliação dos modelos treinados para as abordagens binária e multiclasse.

Tabela 22 – Métricas de desempenho dos modelos com os dados de validação

Modelo/Dados	Técnicas	Acur ^a	Prec ^b	Rev ^c	F1 ^d	AUC ^e
Binário/Simplificado	SVM	92,88	98,94	92,88	95,53	0,934
Multiclasse/Completo	XGB	88,38	99,02	88,38	93,21	0,876

(a) Acur: Acurácia, (b) Prec: Precisão, (c) Rev: Revocação, (d) F1: F1-Score, (e) AUC: Área sob a Curva.

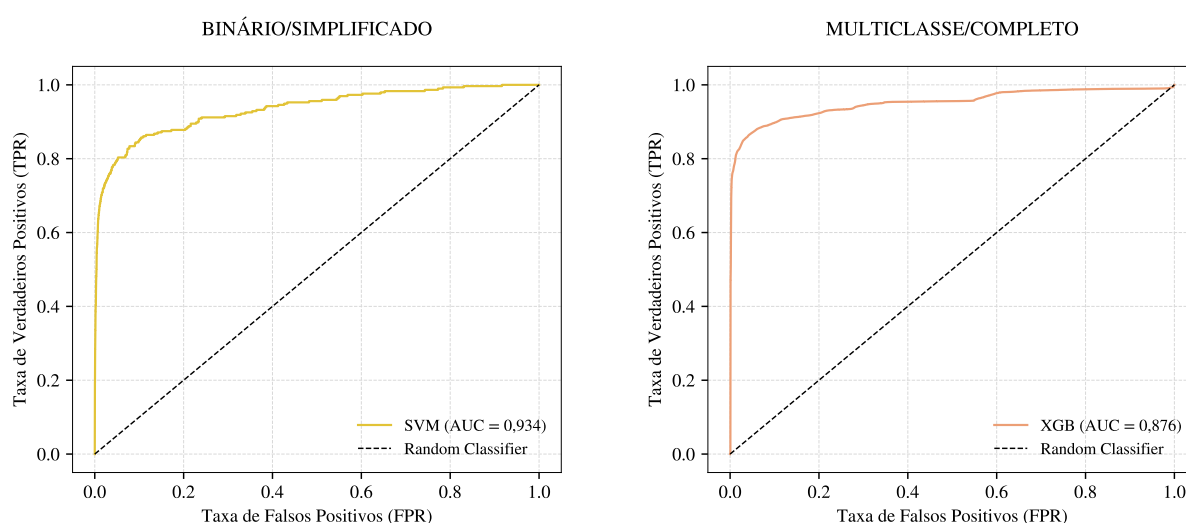
Fonte: Elaborado pela autora (2025).

Conforme as métricas de erro apresentadas na Tabela 22, os valores obtidos mostram-se muito próximos das métricas registradas no processo de experimentação. Isso foi observado tanto no modelo binário com SVM (Tabela 16), treinado com o CDS, quanto no modelo multiclasse com XGB (Tabela 19), treinado com o CDC.

A Figura 74 apresenta a curva ROC dos modelos binário e multiclasse treinados para validação, exibindo resultados muito semelhantes aos observados nas Figuras 70 e 73 referentes ao processo de experimentação.

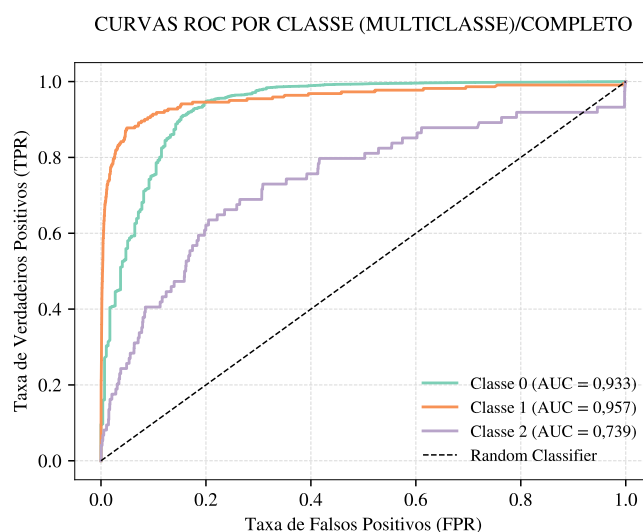
O modelo binário simplificado, implementado com SVM (AUC = 0,934) sugere uma alta capacidade de separação entre as duas classes, refletindo especificamente uma elevada capacidade discriminativa em identificar corretamente os casos de óbito infantil. O modelo multiclasse completo, utilizando a técnica XGB (AUC = 0,876), também reflete um desempenho robusto, embora ligeiramente inferior ao modelo binário.

Figura 74 – Curvas ROC dos modelos binário e multiclasse na validação



Fonte: Elaborado pela autora (2025).

Figura 75 – Curvas ROC por classe do modelo multiclasse na validação



Fonte: Elaborado pela autora (2025).

De forma geral, as curvas ROC da Figura 74 mantêm-se significativamente acima da linha

diagonal do classificador aleatório, evidenciando a habilidade dos modelos em capturar padrões relevantes nos dados.

Adicionalmente, a Figura 75 apresenta as Curvas ROC individuais para cada classe no modelo multiclasse completo com XGB, permitindo uma avaliação detalhada do desempenho discriminativo do classificador para as diferentes categorias: vivo (0), óbito neonatal (1) e óbito pós-neonatal (2). Essa representação facilita a análise específica da capacidade do modelo em distinguir cada classe, oferecendo uma visão mais aprofundada sobre a eficácia preditiva do modelo XGB. Como ilustrado na Figura 75, o modelo multiclasse com XGB apresentou uma alta capacidade de distinguir corretamente as classes 0 e 1. No entanto, a classe 1 ($AUC = 0,957$) obteve o melhor desempenho entre as classes, sugerindo que o modelo é altamente eficaz na identificação do óbito neonatal.

Por outro lado, a classe 2 exibiu uma $AUC = 0,739$, representando um desempenho consideravelmente inferior em comparação com as demais classes. Esse resultado sugere uma maior dificuldade do modelo em discriminar adequadamente os padrões associados ao óbito pós-neonatal, possivelmente devido à sobreposição de características associadas ao óbito neonatal e ao recém-nascido que sobreviveu.

Por conseguinte, ao contrastar os resultados apresentados nesta seção com os achados de alguns estudos da literatura (Quadro 2 e Figura 11), observa-se que a técnica RLog, sendo uma abordagem mais clássica, é apresentada como um modelo capaz de capturar padrões para identificar os nascidos vivos com risco de óbito (Batista *et al.*, 2021; Rahman *et al.*, 2021; Rahman *et al.*, 2022; Choi *et al.*, 2022; Na *et al.*, 2023), especialmente em trabalhos relacionados ao óbito infantil (Hajipour *et al.*, 2021; Rahman *et al.*, 2021; Rahman *et al.*, 2022; Yoneda *et al.*, 2023).

Neste estudo, os resultados das métricas de avaliação obtidos com RLog, assim como com a técnica RF (Jaskari *et al.*, 2020; Mfateneza *et al.*, 2022; Alsubaie; Omar; Riaz, 2023; Na *et al.*, 2023), são bastante adequados, tanto para modelos binários quanto para multiclasse. Não obstante, outras técnicas, como SVM e XGB, demonstraram uma melhor capacidade de adaptação para identificar recém-nascidos com risco de óbito.

Por outro lado, outros estudos indicam um bom desempenho das técnicas baseadas em redes neurais, como o MLP (Khurshid *et al.*, 2021; Do; Moon; Jin, 2022; Hsu *et al.*, 2021), para a previsão de óbito infantil e neonatal. No entanto, neste estudo, embora o MLP tenha apresentado valores elevados nas métricas de avaliação, ao analisar mais detalhadamente seu comportamento na validação cruzada, observou-se uma alta variabilidade, indicando baixa estabilidade e generalização da técnica, com desempenho menos consistente ao ser treinada e testada com diferentes subconjuntos de dados.

Por último, cabe ressaltar que os valores das métricas de desempenho do modelo binário com SVM para identificação de óbito infantil são superiores aos encontrados em outros estudos da literatura (Jaskari *et al.*, 2020; Rahman *et al.*, 2021; Hajipour *et al.*, 2021; Khurshid *et al.*, 2021; Rahman *et al.*, 2022; Mfateneza *et al.*, 2022).

Por outro lado, entre os estudos publicados analisados neste trabalho, não foram identificados trabalhos que abordassem a implementação de modelos para identificação do risco de óbito neonatal e pós-neonatal de forma integrada. foram encontrados estudos que implementaram, separadamente, modelos específicos para identificação do óbito neonatal (Cooper; Minneci; Deans, 2018; Hsu *et al.*, 2021; Batista *et al.*, 2021; Matsushita; Krebs; Carvalho, 2022; Choi *et al.*, 2022; Kim; Kim; Park, 2024).

Assim, ao comparar o valor de AUC do XGB da classe 1 (0,957), referente ao óbito neonatal no modelo multiclasse deste estudo, observa-se que ele é superior aos valores de AUC reportados na literatura para modelos treinados especificamente para identificar o óbito neonatal (Hsu *et al.*, 2021; Matsushita; Krebs; Carvalho, 2022; Choi *et al.*, 2022; Kim; Kim; Park, 2024). O valor de AUC do XGB no estudo de Batista *et al.* (2021) foi superior ao valor obtido neste estudo, possivelmente devido ao fato de o modelo ter sido implementado exclusivamente para a identificação do óbito neonatal, ao contrário do modelo multiclasse apresentado aqui, que abrange a identificação do risco de óbito neonatal e pós-neonatal.

5 CONCLUSÕES E RECOMENDAÇÕES PARA PESQUISAS FUTURAS

5.1 CONCLUSÕES E VERIFICAÇÃO DAS QUESTÕES E OBJETIVOS

A mortalidade do recém-nascido representa um indicador crucial do desenvolvimento e da saúde pública de um país, além de mostrar a qualidade da assistência recebida pela mãe e pelo recém-nascido. Este problema é de importância global, razão pela qual a redução da mortalidade infantil integrou os ODM, atingindo principalmente na redução da mortalidade pós-neonatal. Contudo, a persistência de elevados valores de mortalidade neonatal manteve esse desafio como uma prioridade nos atuais ODS com metas específicas para serem alcançados até 2030.

Sob essa perspectiva, ao longo do tempo, diferentes países têm implementado esforços significativos para reduzir suas taxas de mortalidade do nascido vivo, mas nem todos os países conseguiram a mesma diminuição. Observa-se que tanto países desenvolvidos quanto alguns em desenvolvimento apresentam atualmente uma TMI substancialmente inferior às taxas registradas no Brasil, incluindo aquelas taxas observadas no Estado de São Paulo e na RMVPLN. Esse cenário evidencia que, no contexto brasileiro, apesar dos progressos alcançados na redução da mortalidade infantil, a mortalidade neonatal persiste como um desafio considerável.

Num panorama geral, o esforço global para reduzir a mortalidade do recém-nascido, bem como outros problemas de saúde infantil, tem sido impulsionado pelo interesse crescente em aproveitar algumas ferramentas fornecidas pelo avanço da tecnologia e pela digitalização das informações. Em específico, tem havido um interesse na utilização das informações armazenadas de diferentes processos da área de saúde para serem analisadas, processadas e/ou desenvolver ferramentas para auxiliar na tomada de decisões e em consequência melhorar a assistência prestada.

Nesse sentido, a análise das informações relacionadas à mortalidade do nascido vivos por meio dos SIG e técnicas de ML representa uma abordagem promissora. Essas ferramentas permitem a extração de informações valiosas, que podem ajudar aos gestores em diferentes níveis de planejamento, seja estratégico, tático ou operacional. Além disso, os SIG e o ML possibilitam uma tomada de decisão mais fundamentada, favorecendo a priorização de ações, o monitoramento contínuo dos indicadores de mortalidade e a identificação precoce de nascidos vivos com risco de óbito. Adicionalmente, essas ferramentas têm o potencial de contribuir para a mitigação do impacto social e econômico decorrente da perda evitável de recém-nascidos, cuja dimensão humana transcende qualquer medida quantitativa.

Neste contexto, este estudo explora a aplicabilidade da ciência de dados e das técnicas de ML na análise do problema da mortalidade durante o primeiro ano de vida de recém-nascidos em uma região específica do Brasil. Especificamente, neste estudo foram exploradas as informações relacionadas aos óbitos de nascidos vivos de mães residentes na RMVPLN, por meio da integração de ferramentas de SIG para análises espaciais e da aplicação de técnicas de ML em

abordagens de predição e previsão.

Uma revisão sistemática e abrangente da literatura em bases de dados indexadas, como Scopus e PubMed, permitiu confirmar o potencial dos SIG e das técnicas de ML na análise de dados relacionados à mortalidade infantil. A análise bibliográfica evidenciou que os SIG e as técnicas estatísticas de análise espacial têm sido amplamente utilizados em estudos sobre mortalidade infantil, neonatal e pós-neonatal em diversas partes do mundo, com destaque para um interesse crescente no Brasil, devido às vantagens proporcionadas na identificação de padrões e áreas prioritárias para intervenção.

Os resultados dessa revisão permitiram identificar lacunas importantes na literatura, incluindo a ausência de análises específicas voltadas para a RMVPLN e outras regiões do país. Observou-se que muitas regiões do Brasil ainda carecem de análises detalhadas. Por exemplo, no que se refere à região geográfica deste estudo, embora existam investigações que abordem o Vale do Paraíba, a RMVPLN ainda não havia sido explorada.

A busca bibliográfica sobre o uso de ML na saúde infantil evidenciou que a aplicação dessas técnicas em abordagens de predição e previsão está em crescimento, com foco predominante na identificação de recém-nascidos com risco de óbito, especialmente nos óbitos neonatal e infantil. Do mesmo modo, essa análise também permitiu identificar as áreas geográficas onde os estudos foram conduzidos, a faixa etária analisada e os principais objetivos (associação, previsão ou identificação) das aplicações de ML.

Observou-se, ainda, que há uma escassez de estudos semelhantes aplicados à RMVPLN, bem como um número limitado de pesquisas brasileiras relacionadas diretamente a este trabalho. Esse cenário representa uma oportunidade significativa de investigação, não apenas para a RMVPLN, mas também para outras regiões, municípios e estados do Brasil, consolidando a relevância deste estudo no avanço científico e na formulação de políticas públicas baseadas em evidências.

Por outro lado, em relação à aplicabilidade dos SIG e à análise exploratória de dados, como a análise espacial, os resultados deste estudo indicaram que essas ferramentas podem ser efetivamente empregadas no desenvolvimento de soluções a nível estratégico. O uso dos dados fornecidos pelo SINASC e pelo SIM possibilitou a exploração e a identificação de informações relevantes. Além disso, a utilização das malhas cartográficas do IBGE, juntamente com a aplicação das técnicas de Moran, tanto univariadas quanto bivariadas, em análises locais e globais, permitiu a criação de mapas temáticos. Esses mapas desempenharam uma função importante na visualização das informações, identificando áreas geográficas com maior risco de mortalidade e fornecendo *insights* cruciais para a formulação de políticas públicas e ações voltadas à redução da mortalidade do nascido vivo.

Especificamente, os resultados da distribuição espacial das mortes infantis na RMVPLN mostraram que a redução das TMI durante o período 2010 – 2020 se deveu principalmente à redução do número de óbitos no período pós-neonatal.

A análise espacial global revelou significância estatística apenas para a TMN, sugerindo a possível presença de aglomerações espaciais na RMVPLN. A análise global não mostrou

significância estatística com a TMI e TMPN.

Por sua vez, a análise local identificou agrupamentos espaciais significativos de municípios com alto risco para a TMI, os quais coincidem com os municípios de maior risco para a TMN. Esses achados na análise local corroboraram os resultados obtidos na análise global. Foi observado que os municípios agrupados com altos valores de mortalidade infantil e neonatal estão localizados na região Norte Central da RMVPLN, além de indicar que o aglomerado de municípios com alta TMI é influenciado principalmente pela presença de altos valores na TMN. Por outro lado, a análise local da ocorrência de óbitos pós-neonatais não mostrou significância estatística em relação à presença de aglomerações de municípios com risco agravado na TMPN.

Adicionalmente, a análise dos dados disponíveis possibilitou a identificação de padrões de associação entre a mortalidade do recém-nascido e diversos determinantes associados. A análise exploratória inicial revelou variações no comportamento espacial e temporal da mortalidade do nascido vivo nas faixas etárias neonatal, pós-neonatal e infantil na região da RMVPLN. Observou-se uma oscilação não tão estável no número de óbitos ao longo dos anos, meses, quinzenas, semanas e dias. Além disso, a análise dos dados revelou a presença de valores ausentes em aproximadamente um quarto das variáveis.

A análise espacial bivariada, por sua vez, permitiu identificar associações significativas entre a mortalidade e uma série de determinantes associados, incluindo características dos pais, do recém-nascido, da assistência à saúde e do município. Em termos de mortalidade infantil (TMI), foi observada uma associação com a idade materna e paterna, tipo de parto e prematuridade. Para a mortalidade neonatal (TMN), além dessas variáveis, identificaram-se associações também com o estado civil, IPRS e PIB. A mortalidade pós-neonatal (TMPN) apresentou correlações significativas com o tipo de gravidez, IDH e PIB.

Esses achados constituem uma base importante para o desenvolvimento e treinamento de modelos preditivos utilizando técnicas de ML, permitindo identificar que as associações espaciais observadas podem ser atribuídas à presença de características semelhantes entre os nascidos vivos do mesmo município ou subregião. Essa homogeneidade nas características dos nascidos vivos em determinadas áreas pode influenciar os padrões de mortalidade, evidenciando a importância de considerar a dimensão espacial na modelagem preditiva.

Em relação à aplicação de ML, suas técnicas podem ser utilizadas para o desenvolvimento de modelos que auxiliem na tomada de decisões tanto a nível tático quanto operacional.

A nível tático, o modelo de previsão foi treinado, testado e validado com técnicas de regressão supervisionada de ML, como RF, SVM, MLP, XGB, LGBM, GB, DT e AdaBoost, com o objetivo de prever o número mensal de óbitos na região metropolitana. Estes foram comparados com modelos tradicionais de séries temporais, como ARIMA e SARIMA.

Para o treinamento dos modelos com ML, o banco de dados foi reestruturado ao formato de séries temporais do número mensal de óbito da faixa etária neonatal, pós-neonatal e infantil. As informações usadas nessa etapa relacionaram-se ao número de ocorrências mensais do período de 2010 a 2021, que foi usado para cada um dos processos. Concretamente, os dados de 132

meses de 2010 a 2020 foram usados para treinamento e teste com a técnica de validação cruzada contínua, e os dados dos 12 meses restantes de 2021 foram usados para o processo de validação.

Além disso, foram consideradas 11 configurações diferentes com relação ao número de meses anteriores usados para treinar os modelos. Esse procedimento gerou um processo iterativo no qual foram testados de 2 a 12 meses atrás para prever um novo valor. Durante o processo de treinamento foi realizada uma otimização de hiperparâmetros para identificar a melhor configuração para cada técnica testada.

Os resultados indicaram que cada técnica testada apresentou uma adequação diferente em relação ao padrão temporal das mortes mensais ocorridas nas faixas etárias neonatal, pós-neonatal e infantil.

Em resumo, a técnica de RF demonstrou capturar melhor o padrão temporal dos dados, com menor erro para a previsão mensal de mortes infantis, estimadas com base nas ocorrências dos dois meses anteriores. Para a previsão de óbitos neonatais mensais, o modelo com melhor desempenho foi treinado com GB, utilizando a estrutura do banco de dados dos dois meses anteriores para prever um novo valor para o mês seguinte. Já na previsão de mortalidade pós-neonatal, o modelo com melhores resultados nas métricas de avaliação foi o SVM, com os dados dos três meses anteriores. Nos resultados foi demonstrado que as técnicas de ML apresentaram um melhor ajuste nos dados, capturando o padrão temporal, no qual as previsões geradas apresentaram um erro menor em comparação com os modelos clássicos.

A nível operacional, foram realizados dois processos experimentais com o objetivo de treinar um modelo binário para identificação de risco de óbito infantil e um modelo multiclasse para identificação de risco de óbito neonatal ou pós-neonatal. Os modelos de predição foram treinados, testados e validados com técnicas de classificação supervisionada, como RLog, DT, SVM, MLP, RF, XGB, LGBM, Bagging e GB. Da mesma forma, foram otimizados seus hiperparâmetros em cada processo de treinamento.

Para o treinamento dos modelos para identificação de risco de óbito, os dados do SIM e do SINASC foram vinculados, permitindo a identificação dos nascidos vivos que evoluíram para óbitos. As informações referentes ao período de 2010 a 2019 foram utilizadas no processo experimental, enquanto as informações de 2021 foram usadas no processo de validação. No banco de dados, foi adicionada uma variável relacionada à dimensão espacial (SUB-REGIAO), com base na residência da mãe. Além disso, as características dos dados e sua importância foram analisadas para verificar se os modelos de predição apresentam um desempenho melhor com o CDC ou com o CDS.

De forma semelhante aos modelos de previsão, as técnicas testadas apresentaram um comportamento e variabilidade diferentes durante o processo de validação cruzada com 5 *folds*. Entre os melhores modelos treinados, foram selecionados aqueles que apresentaram menor variação nas métricas entre as dobras, sendo esses mais confiáveis para dados futuros, pois seu desempenho não depende de um conjunto específico. Assim, o modelo para a abordagem binária foi implementado com a técnica SVM e o CDS, enquanto o modelo multiclasse foi desenvolvido

com a técnica XGB e o CDC. Esses resultados indicam que as técnicas de SVM e XGB se destacaram como as mais adequadas para a predição de risco de mortalidade.

Por fim, os resultados obtidos em ambas as abordagens de predição e previsão foram validados por meio da avaliação de novas previsões com os conjuntos de dados de validação, correspondentes a dados não vistos em nenhum dos processos anteriores. Em ambos os casos, os resultados demonstraram que cada modelo gerou métricas de erro muito semelhantes às obtidas durante o processo experimental de treinamento e teste, confirmando que as previsões desses modelos estão muito próximas dos valores reais. Esses resultados foram confirmados por meio das métricas de desempenho e testes estatísticos específicos para cada abordagem, bem como por análises adicionais, incluindo o teste-t pareado e a curva ROC, garantindo confiabilidade, robustez e aplicabilidade prática dos resultados obtidos.

Dessa forma, conclui-se que as sistemáticas propostas neste trabalho podem ser aplicadas em diferentes níveis de planejamento, aprimorando a tomada de decisões e permitindo a implementação de medidas preventivas. Isso contribui para a melhoria da qualidade do cuidado ao recém-nascido e a garantia de sua sobrevivência.

5.2 PROPOSTA PARA FUTURAS PESQUISAS

Algumas propostas para pesquisas futuras são apresentadas a seguir:

No nível estratégico, a análise espacial demonstrou ser uma ferramenta poderosa; contudo, ainda existem diversas possibilidades para explorar a aplicabilidade dessas técnicas em diferentes contextos geográficos e socioeconômicos do Brasil. Futuros estudos poderiam investigar a influência de fatores ambientais, como a qualidade do ar e o acesso à água potável, sobre a mortalidade infantil.

No nível tático, o modelo apresentado mostrou-se uma abordagem eficaz para antecipar tendências e permitir uma alocação mais eficiente de recursos. No entanto, a precisão desses modelos pode ser aprimorada com a inclusão de variáveis externas, como campanhas de saúde pública, vacinação, surtos de doenças ou eventos climáticos.

No nível operacional, os modelos treinados podem direcionar intervenções de saúde de maneira eficaz. Estudos futuros podem explorar abordagens mais sofisticadas, além do uso de dados mais completos e variados, como informações provenientes de análises clínicas ou ultrassonografias. Além disso, os modelos poderiam ser ajustados para integrar dados em tempo real, o que permitiria a identificação de riscos emergentes e facilitaria intervenções rápidas.

Por último, seria relevante, testar a aplicabilidade dos modelos em diferentes contextos regionais, adaptando-os às peculiaridades de cada localidade. As pesquisas também poderiam concentrar-se na incorporação de abordagens de análise espacial local, com o objetivo de ajustar os modelos de risco com base em características geográficas, como o acesso a serviços de saúde e a infraestrutura disponível.

REFERÊNCIAS

- ABDULLAH, T.; ZAHID, M. S. M.; ALI, W. A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. **Symmetry**, Basel, v. 13, n. 12, p. 2439, 2021.
- ALMEIDA, E. **Econometria espacial aplicada**. Campinas: Alínea, 2012. 150–183 p.
- ALMEIDA, M. C. S.; GOMES, C. M. S.; NASCIMENTO, L. F. C. Análise espacial da mortalidade neonatal no estado de São Paulo, 2006-2010. **Revista Paulista de Pediatria**, São Paulo, v. 32, p. 374–380, 2014.
- ALMEIDA, M. F. B. *et al.* Perinatal factors associated with early deaths of preterm infants born in brazilian network on neonatal research centers. **Jornal de Pediatria**, Rio de Janeiro, v. 84, n. 4, p. 300–307, 2008.
- ALPAYDIN, E. **Machine learning**. Massachusetts: MIT Press, 2021.
- ALSUBAIE, H.; OMAR, M. H.; RIAZ, M. Study the efficiency of jean bourgeois pichat model to predict infant mortality compared to machine learning random forest. In: SUSTAINABLE DIGITAL TRANSFORMATION, 50., 2023, Sharjah. **Proceedings of Title of host publication 50th International Conference on Computers and Industrial Engineering**. Sharjah: Computers and Industrial Engineering, 2023. p. 522–531.
- ANDRADE, C. L. T.; SZWARCOWALD, C. L. Análise espacial da mortalidade neonatal precoce no município do Rio de Janeiro, 1995-1996. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1199–1210, 2001.
- ANSELIN, L. Local indicators of spatial association: Lisa. **Geographical Analysis**, Hoboken, v. 27, n. 2, p. 93–115, 1995.
- ANSELIN, L.; REY, S. J. **Modern spatial econometrics in practice: a guide to GeoDa, GeoDaSpace and PySAL**. Chicago: GeoDa Press LLC, 2014.
- ANSELIN, L.; SYABRI, I.; KHO, Y. GeoDa: an introduction to spatial data analysis. v. 38, p. 5–22, 2005.
- ANSELIN, L.; SYABRI, I.; SMIRNOV, O. Visualizing multivariate spatial correlation with dynamically linked windows. In: UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN, 2., 2002, Santa Bárbara. **Proceedings of the specialist meeting**. Hoboken: John Wiley and Sons, 2002. p. 1–20.
- ARECO, K. C. N.; KONSTANTYNER, T.; TADDEI, J. A. D. A. C. Secular trends in infant mortality by age-group and avoidable components in the state of São Paulo, 1996–2012. **Revista Paulista de Pediatria**, São Paulo, v. 34, p. 263–270, 2016.
- ASHORN, p. *et al.* The lancet small vulnerable newborn series: science for a healthy start. **The Lancet**, London, v. 396, n. 10253, p. 743–745, 2020.

Assembleia Legislativa do Estado de São Paulo. **Regiões metropolitanas de Piracicaba, Jundiaí e Rio Preto avançam na criação de planos integrados de desenvolvimento**. São Paulo: ALESP, 2022. Disponível em: <https://www.al.sp.gov.br/noticia/?id=441823>. Acesso em: 19 set. 2023.

AWOTUNDE, J. B. *et al.* Internet of medical things for enhanced smart healthcare systems. In: **IMPLEMENTATION OF SMART HEALTHCARE SYSTEMS USING AI, IOT, AND BLOCKCHAIN**. San Diego: Academic Press, 2023. p. 1–28.

BARES, M.; BOSSE, E. **Relational calculus for actionable knowledge**. Nova Iorque: Springer, 2022.

BARROS, C. I. M. *et al.* Infant mortality: surveillance, epidemiological characteristics and spatial distribution pattern in Recife, Brazil. **Acta Scientiarum Health Sciences**, Paraná, v. 41, p. 43997, 2019.

BATISTA, A. F. M. *et al.* Neonatal mortality prediction with routinely collected data: a machine learning approach. **BMC pediatrics**, Massachusetts, v. 21, p. 1–6, 2021.

BATTAGLIA, F. C.; LUBCHENCO, L. O. A practical classification of newborn infants by weight and gestational age. **The Journal of Pediatrics**, Nova Iorque, v. 71, n. 2, p. 159–163, 1967.

BEKKAR, M.; ALITOUCE, T. A. Imbalanced data learning approaches review. **International Journal of Data Mining and Knowledge Management Process**, Zurique, v. 3, n. 4, p. 15, 2013.

BELUZO, C. E. *et al.* Towards neonatal mortality risk classification: a data-driven approach using neonatal, maternal, and social factors. **Informatics in medicine unlocked**, Massachusetts, v. 20, p. 100398, 2020.

BENKE, K.; BENKE, G. Artificial intelligence and big data in public health. **International journal of environmental research and public health**, Basel, v. 15, n. 12, p. 2796, 2018.

BERNARDINO, F. B. S. *et al.* Trends in neonatal mortality in Brazil from 2007 to 2017. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 27, p. 567–578, 2022.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. Nova Iorque: Springer, 2006. v. 4.

BIVAND, R. S.; WONG, D. W. S. Comparing implementations of global and local indicators of spatial association. **Test**, Nova Iorque, v. 27, p. 716–748, 2018.

BONFIM, C. V. *et al.* Spatial analysis of inequalities in fetal and infant mortality due to avoidable causes. **Revista Brasileira de Enfermagem**, Rio de Janeiro, v. 73, 2020.

BRAHMA, D.; MUKHERJEE, D. Early warning signs: targeting neonatal and infant mortality using machine learning. **Applied Economics**, Pittsburgh, v. 54, p. 57–74, 2022.

Brasil. Secretaria de vigilância em saúde. portaria nro. 116, de 11 de fevereiro de 2009. **Sistema de Legislação da Saúde**, Ministério da Saúde, Brasília, 2009. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/svs/2009/prt0116_11_02_2009.html. Acesso em: 19 jul. 2022.

BRASIL. Lei no. 13.089, de 12 de janeiro de 2015. In: **ESTATUTO DA METRÓPOLE**. Brasília: Casa Civil, Presidência da República, 2015. Disponível em: <https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&data=13/01/2015&pagina=2>. Acesso em: 17 jul. 2022.

BRODAY, G. A.; KLUTHCOVSKY, A. C. G. C. Mortalidade infantil e estratégia saúde da família na 3 regional de saúde do Paraná, de 2005 a 2016. **Revista Paulista de Pediatria**, São Paulo, v. 40, p. 1–10, 2021.

BUHLER, H. F. *et al.* Spatial analysis of integrated health and environmental indicators for morbidity and mortality due to infant diarrhea in Brazil, 2010. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 30, p. 1921–1934, 2014.

CAPELLI, J. C. S. *et al.* Birth weight and factors associated with the prenatal period: a cross-sectional study in a maternity hospital of reference. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 19, n. 7, p. 2063–2072, 2014.

CHEN, T.; GUESTRIN, C. Xgboost. In: ASSOCIATION FOR COMPUTING MACHINER, 22., 2016, San Francisco. **Proceedings Of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining**. Nova Iorque: ACM, 2016. p. 785–794.

CHEN, Y.; HE, J.; WANG, M. A hybrid of long short-term memory neural network and autoregressive integrated moving average model in forecasting hiv incidence and morality of post-neonatal population in East Asia: global burden of diseases 2000–2019. **BMC Public Health**, London, v. 22, p. 1938, 2022.

CHIAVEGATTO, A. D. P. F. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, Brasília, v. 24, p. 325–332, 2015.

CHICCO, D. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, Londres, v. 21, n. 6, 2020.

CHOI, H. J. *et al.* Clinical characteristics of neonatal cholestasis in a tertiary hospital and the development of a novel prediction model for mortality. **EBio Medicine**, Amsterdam, v. 77, p. 103890, 2022.

CHRISTEN, P. **The data matching process**. Nova Iorque: Springer, 2012.

CLIFF, A. D.; ORD, K. **Spatial process: models and applications**. Londres: Pion Ltd, 1981.

COLODETTE, A. L. *et al.* Feature selection for identification of risk factors associated with infant mortality. In: COMPUTATIONAL ADVANCES IN BIO AND MEDICAL SCIENCE, 1., 2021, Cham. **Proceedings of International Conference on Computational Advances in Bio and Medical Sciences**. Nova Iorque: Springer, 2021. p. 92–102.

COOPER, J. N.; MINNECI, P. C.; DEANS, K. J. Postoperative neonatal mortality prediction using superlearning. **Journal of Surgical Research**, Filadélfia, v. 221, p. 311–319, 2018.

CORREIA, L. O. S.; PADILHA, B. M.; VASCONCELOS, S. M. L. Methods for assessing the completeness of data in health information systems in Brazil: a systematic review. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 19, n. 11, p. 4467–4478, 2014.

COSTA, M. S. C.; FIGUEIREDO, F. W. S. Relationship between income inequality, socioeconomic development, vulnerability index, and maternal mortality in Brazil, 2017. **BMC public health**, London, v. 21, p. 1842, 2021.

COSTA-NOBRE, D. T. *et al.* Clusters of cause specific neonatal mortality and its association with per capita gross domestic product: a structured spatial analytical approach. **Plos one**, California, v. 16, n. 8, p. e0255882, 2021.

CUPIDO, K.; JEVTIC, P.; PAEZ, A. Spatial patterns of mortality in the United States: a spatial filtering approach. **Insurance Mathematics and Economics**, Toronto, v. 95, p. 28–38, 2020.

DATASUS. **População residente na região metropolitana**. Brasília: Departamento de Informática do Sistema Único de Saúde, 2010. Disponível em: <http://tabnet.datasus.gov.br/cgi/ta bcgi.exe?ibge/cnv/popsp.def>. Acesso em: 18 jul. 2022.

DATASUS. **Informações de nascidos vivos e óbitos infantis**. Brasília: Departamento de Informática do Sistema Único de Saúde, 2023. Disponível em: <https://datasus.saude.gov.br/trans ferencia-de-arquivos>. Acesso em: 10 jul. 2022.

DEEPIKA, K.; CHOWHAN, S. A threat towards the neonatal mortality. In: COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE, 4., 2020, Valletta. **Proceedings of 4th International Conference on Advances in Computing and Data Sciences**. Nova Iorque: Springer, 2020. p. 56–65.

DEMIR, E. *et al.* The evolution of the field of legal medicine: a holistic investigation of global outputs with bibliometric analysis. **Journal of Forensic and Legal Medicine**, Londres, v. 69, p. 101885, 2020.

DEREJE, T. *et al.* Investigate risk factors and predict neonatal and infant mortality based on maternal determinants using homogenous ensemble methods. In: INFORMATION AND COMMUNICATION TECHNOLOGY, 7., 2021, Bahir Dar. **Proceedings of 7th International Conference on Information and Communication Technology for Development for Africa**. Nova Iorque: IEEE, 2021. p. 18–23.

DEVINE, O. J.; LOUIS, T. a.; HALLORAN, M. E. Empirical bayes methods for stabilizing incidence rates before mapping. **Epidemiology**, Cary, v. 5, n. 6, p. 622–630, 1994.

DIAS, B. A. S. *et al.* Spatial analysis of avoidable infant deaths in Espírito Santo, Brazil, 2006-2013. **Epidemiologia e Serviços de Saúde**, São Paulo, v. 28, n. 3, p. e2018111, 2019.

DILÉLIO, A. S. *et al.* Estrutura e processo na atenção primária à saúde das crianças e distribuição espacial da mortalidade infantil. **Revista de Saúde Pública**, São Paulo, v. 58, p. 21, 2024.

DO, H. J.; MOON, K. M.; JIN, H.-S. Machine learning models for predicting mortality in 7472 very low birth weight infants using data from a nationwide neonatal network. **Diagnostics**, Basel, v. 12, n. 3, p. 625, 2022.

DOUPE, P.; FAGHMOUS, J.; BASU, S. Machine learning for health services researchers. **Value in Health**, Nova Iorque, v. 22, n. 7, p. 808–815, 2019.

DUTRA, V. G. P. *et al.* Spatio-temporal analysis of infant mortality in the city of Rio de Janeiro, 2010–2019. **Revista Paulista de Pediatria**, São Paulo, v. 40, p. e2021061, 2022.

ECK, N. V.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, Nova Iorque, v. 84, n. 2, p. 523–538, 2010.

EMTU/SP. **Região Metropolitana do Vale do Paraíba e Litoral Norte**. São Paulo: Empresa Metropolitana de Transportes Urbanos de São Paulo, 2022. Disponível em: <https://www.emtu.sp.gov.br/emtu/institucional/quem-somos/vale-do-paraiba-e-litoral-norte.fss>. Acesso em: 15 jul. 2022.

ESPINOLA-SÁNCHEZ, M. *et al.* Prediction of neonatal death in pregnant women in an intensive care unit: application of machine learning models. **Heliyon**, Cambridge, v. 9, n. 10, p. e20693, 2023.

FACELI, K. *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2021.

FALAGAS, M. E. *et al.* Comparison of PubMed, scopus, web of science, and google scholar: strengths and weaknesses. **The FASEB journal**, Maryland, v. 22, n. 2, p. 338–342, 200.

FARIA, R.; SANTANA, P. Variações espaciais e desigualdades regionais no indicador de mortalidade infantil do estado de Minas Gerais, Brasil. **Saúde e Sociedade**, São Paulo, v. 25, n. 3, p. 736–749, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, Nova Iorque, v. 39, n. 11, p. 27–34, 1996.

FERRAZ, T. R.; NEVES, E. T. Risk factors for low birth weight in the in public maternities: a cross sectional study. **Revista Gaúcha de Enfermagem**, Porto Alegre, v. 32, p. 86–92, 2011.

FILHO, J. G. B. *et al.* Spatial distribution of the infant mortality rate and its principal determinants in the state of Ceará, Brazil, 2000-2002. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 23, n. 5, p. 1173–1185, 2007.

FONSECA, S. C. *et al.* Maternal education and age: inequalities in neonatal death. **Revista de Saúde Pública**, São Paulo, v. 51, p. 94, 2017.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI magazine**, v. 13, n. 3, p. 57–57, 1992.

FULLER, W. A. **Introduction to statistical time series**. Hoboken: John Wiley and Sons, 2009.

Fundação SEADE. **Produto interno bruto para os municípios de São Paulo**. São Paulo: Fundação Sistema Estadual de Análise de Dados, 2018. Disponível em: <https://repositorio.seade.gov.br/dataset/pib-municipal-2002-2018/resource/7a58161d-f687-4c15-8edf-0787a6b03245>. Acesso em: 18 jul. 2022.

Fundação SEADE. **Índice paulista de responsabilidade social**. São Paulo: Fundação Sistema Estadual de Análise de Dados, 2018. Disponível em: <http://www.iprs.seade.gov.br/>. Acesso em: 18 jul. 2022.

GAIVA, M. A. M.; FUJIMORI, E.; SATO, A. P. S. Maternal and child risk factors associated with neonatal mortality. **Texto e Contexto-Enfermagem**, Santa Catarina, v. 25, n. 4, p. e2290015, 2016.

GALICIA, A. *et al.* Multi-step forecasting for big data time series based on ensemble learning. **Knowledge-Based Systems**, Texas, v. 163, p. 830–841, 2019.

GARCIA, L. P. *et al.* What is the impact of national public expenditure and its allocation on neonatal and child mortality? a machine learning analysis. **BMC Public Health**, Londres, v. 23, p. 793, 2023.

GERHARDT, T. E. *et al.* **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

GHAVAMI, P. **Big data analytics methods**. Boston: Walter de Gruyter Inc., 2020.

GONÇALVES, A. C.; COSTA, M. C. N.; BRAGA, J. U. Análise da distribuição espacial da mortalidade neonatal e de fatores associados, em Salvador, Bahia, Brasil, no período 2000-2006. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 27, n. 8, p. 1581–1592, 2011.

GOOS, P.; MEINTRUP, D. **Statistics with JMP**: graphs, descriptive statistics and probability. Hoboken: John Wiley and Sons, 2015.

GUJARATI, D. N. **Basic econometrics**. Nova Jersey: Prentice Hall, 2022.

HAJIPOUR, M. *et al.* Predictive factors of infant mortality using data mining in Iran. **Journal of Comprehensive Pediatrics**, 's-Hertogenbosch, v. 12, n. 1, 2021.

HAN, J.; PEI, J.; TONG, H. **Data mining**: concepts and techniques. São Francisco: Morgan kaufmann, 2022.

HSU, J.-F. *et al.* Machine learning approaches to predict in-hospital mortality among neonates with clinically suspected sepsis in the neonatal intensive care unit. **Journal of Personalized Medicine**, Basel, v. 11, n. 8, p. 695, 2021.

HUG, L. *et al.* National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. **The Lancet Global Health**, Nova Iorque, v. 7, n. 6, p. e710–e720, 2019.

IBGE. **Tábuas completas de mortalidade**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2017. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9126-ta-buas-completas-de-mortalidade.html>. Acesso em: 12 abr. 2022.

IBGE. **Malha municipal**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2023. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html#:~:text=O%20que%20%C3%A9,e%20demais%20pesquisas%20do%20IBGE..> Acesso em: 08 jun. 2020.

INSAF, T. Z.; TALBOT, T. Identifying areas at risk of low birth weight using spatial epidemiology: a small area surveillance study. **Preventive Medicine**, Chicago, v. 88, p. 108–114, 2016.

Instituto de Pesquisa Econômica Aplicada. **Planilha índice de vulnerabilidade social e índice de desenvolvimento humano**. Brasília: IPEA, 2010. Disponível em: <http://ivs.ipea.gov.br/index.php/pt/planilha>. Acesso em: 18 jul. 2022.

IQBAL, F. *et al.* Prediction of mortality among neonates with sepsis in the neonatal intensive care unit: a machine learning approach. **Clinical Epidemiology and Global Health**, Indira Nagar, v. 24, p. 101414, 2023.

- IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. São Carlos: UICLAP, 2021.
- JAMES, G. *et al.* **An introduction to statistical learning with applications in python**. Nova Iorque: Springer, 2023.
- JASKARI, J. *et al.* Machine learning methods for neonatal mortality and morbidity classification. **IEEE Access**, Piscataway, v. 8, p. 123347–123358, 2020.
- JAVAID, M. *et al.* Significance of machine learning in healthcare: features, pillars and applications. **International Journal of Intelligent Networks**, Uttarakhand, v. 3, p. 58–73, 2022.
- JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, Nova Iorque, v. 349, n. 6245, p. 255–260, 2015.
- JUSTINO, D. C. P. *et al.* Assessment of the causes of infant morbidity and mortality in Brazil. **Mundo da Saúde**, São Paulo, v. 1, n. 45, p. 152–161, 2021.
- KAZEMBE, L. N.; KANDALA, N. B. Estimating areas of common risk in low birth weight and infant mortality in Namibia: a joint spatial analysis at sub-regional level. **Spatial and Spatio-Temporal Epidemiology**, Florida, v. 12, p. 27–37, 2015.
- KE, G. *et al.* Lightgbm: a highly efficient gradient boosting decision tree. In: ASSOCIATION FOR COMPUTING MACHINER, 31., 2017, Long Beach. **Proceedings of 31st Conference on Neural Information Processing Systems**. Nova Iorque: ACM, 2017. p. 1–9.
- KEFI, Z.; ALOUI, K.; NACEUR, M. S. New approach based on machine learning for short-term mortality prediction in neonatal intensive care unit. **International Journal of Advanced Computer Science and Applications**, Bradford, v. 10, n. 7, p. 583–591, 2019.
- KHANDWALA, Y. S. *et al.* Association of paternal age with perinatal outcomes between 2007 and 2016 in the United States: population based cohort study. **BMJ**, Londres, v. 363, p. k4372, 2018.
- KHURSHID, F. *et al.* Comparison of multivariable logistic regression and machine learning models for predicting bronchopulmonary dysplasia or death in very preterm infants. **Frontiers in Pediatrics**, Lausanne, v. 9, p. 759776, 2021.
- KIM, H. H.; KIM, J. K.; PARK, S. Y. Predicting severe intraventricular hemorrhage or early death using machine learning algorithms in VLBWI of the korean neonatal network database. **Scientific Reports**, Filadélfia, v. 14, p. 11113, 2024.
- KOVACS, D. *et al.* Developing practical clinical tools for predicting neonatal mortality at a neonatal intensive care unit in Tanzania. **BMC Pediatrics**, Massachusetts, v. 21, p. 537, 2021.
- KROPIWIEC, M. V.; FRANCO, S. C.; AMARAL, A. R. D. Factors associated with infant mortality in a brazilian city with high human development index. **Revista Paulista de Pediatria**, São Paulo, v. 35, n. 4, p. 391–398, 2017.
- KUMAR, S. N. *et al.* Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. **Reproductive Toxicology**, Lichfield, v. 94, p. 92–100, 2020.

KURINCZUK, J. J. *et al.* The contribution of congenital anomalies to infant mortality. **National Perinatal Epidemiology Unit, University of Oxford**, Londres, p. 2–3, 2010.

KWIATKOWSKI, D. *et al.* Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? **Journal of Econometrics**, Amsterdã, v. 54, n. 1-3, p. 159–178, 1992.

LAU, C. *et al.* Extremely low birth weight and infant mortality rates in the United States. **Pediatrics**, Nova Iorque, v. 131, n. 5, p. 855–860, 2013.

LAWN, J. E. *et al.* Every newborn: progress, priorities, and potential beyond survival. **The Lancet**, Londres, v. 384, n. 9938, p. 189–205, 2014.

LEAL, M. C.; SZWARCOWALD, C. L. Characteristics of neonatal mortality in the state of Rio de Janeiro, Brazil, in the 1980's: a spatio-temporal analysis. **Revista de Saúde Pública**, São Paulo, v. 31, n. 5, p. 457–465, 1997.

LEAL, M. D. C. *et al.* Saúde reprodutiva, materna, neonatal e infantil nos 30 anos do sistema único de saúde. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 23, p. 1915–1928, 2018.

LEE, J. *et al.* Predicting mortality risk for preterm infants using random forest. **Scientific Reports**, Filadélfia, v. 11, p. 7308, 2021.

LOURENÇO, B. P. *et al.* Identifying the determinants of infant and youth mortality in Portugal: a machine learning approach. In: INTERNATIONAL NEURAL NETWORKS SOCIETY, 2024, Yokohama. **Proceedings of International Joint Conference on Neural Networks**. Piscataway: IEEE, 2024. p. 1–9.

LOURENÇO, E. do C. *et al.* Impact variables on the decline in infant mortality in the state of São Paulo, Brazil: 1998-2008. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 19, n. 7, p. 2055, 2014.

MAKRIDAKIS, S.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. **Forecasting methods and applications**. Hoboken: John Wiley and sons, 2008.

MALHOTRA, A. *et al.* Emerging role of artificial intelligence, big data analysis and precision medicine in pediatrics. **Pediatric Research**, Jersey, v. 93, n. 2, p. 281–283, 2023.

MARINHO, C. S. R. *et al.* Millennium development goals: the impact of healthcare interventions and changes in socioeconomic factors and sanitation on under-five mortality in Brazil. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 36, n. 10, p. e00191219, 2020.

MARINONIO, A. S. S. *et al.* Clusters of preterm live births and respiratory distress syndrome-associated neonatal deaths: spatial distribution and cooccurrence patterns. **BMC Public Health**, Londres, v. 22, n. 1, p. 1–10, 2022.

MARSHALL, R. J. Mapping disease and mortality rates using empirical bayes estimators. **Journal of the Royal Statistical Society**, Cary, v. 40, n. 2, p. 283–294, 1991.

MARSLAND, S. **Machine learning: an algorithmic perspective**. Boca Raton: Taylor and Francis Group, 2015. ISBN 1-466-58328-2.

MARYS, P.; AUTRAN, M. Repositório datasus: organização e relevância dos dados abertos em saúde para a vigilância epidemiológica. **P2P E INOVAÇÃO**, Rio de Janeiro, v. 6, p. 50–59, 10 2019.

MATHEWS, T. J.; MACDORMAN, M. F.; THOMA, M. E. Infant mortality statistics from the 2013 period linked birth/infant death data set. **National Vital Statistics Reports**, California, v. 64, n. 9, p. 1–30, 2015.

MATSUSHITA, F. Y.; KREBS, V. L. J.; CARVALHO, W. B. Artificial intelligence and machine learning in pediatrics and neonatology healthcare. **Revista da Associação Médica Brasileira**, São Paulo, v. 68, p. 745–750, 2022.

MELLO, L. F.; TEIXEIRA, L. R.; MELLO, A. Y. I. População e desenvolvimento na Região Metropolitana do Vale do Paraíba e Litoral Norte de São Paulo. In: ASSOCIAÇÃO BRASILEIRA DE ESTUDOS POPULACIONAIS, 18., 2016, Águas de Lindóia. **Anais do XVIII Encontro Nacional de Estudos Populacionais**. São Paulo: Galoá, 2016. p. 1–25.

MELO, M. S. *et al.* Temporal trend, spatial analysis and spatiotemporal clusters of infant mortality associated with congenital toxoplasmosis in Brazil: time series from 2000 to 2020. **Tropical Medicine and International Health**, Hoboken, v. 28, n. 6, p. 476–485, 2023.

MFATENEZA, E. *et al.* Application of machine learning methods for predicting infant mortality in Rwanda: analysis of Rwanda demographic health survey 2014–15 dataset. **BMC Pregnancy and Childbirth**, Londres, v. 22, p. 388, 2022.

MOHR, D. L.; WILSON, W. J.; FREUND, R. J. **Statistical methods**. San Diego: Academic Press, 2021.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied statistics and probability for engineers**. Hoboken: John Wiley and sons, 2020.

MS-SVE. **Saúde Brasil 2019: uma análise da situação de saúde com enfoque nas doenças imunopreveníveis e na imunização**. Brasília: Ministério da Saúde, Secretaria de Vigilância Epidemiológica, 2019. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/saude_brasil_2019_analise_situacao.pdf. Acesso em: 12 jan. 2021.

MUNRO, B. H. **Statistical methods for health care research**. Filadélfia: Lippincott Williams and Wilkins, 2012. v. 6.

NA, J. Y. *et al.* Learning-based longitudinal prediction models for mortality risk in very-low-birth-weight infants: a nationwide cohort study. **Neonatology**, Columbus, v. 120, n. 5, p. 652–660, 2023.

NASCIMENTO, I. J. B. *et al.* Impact of big data analytics on people's health: overview of systematic reviews and recommendations for future studies. **Journal of Medical Internet Research**, Toronto, v. 23, n. 4, p. e27275, 2021.

NASCIMENTO, L. *et al.* Spatial analysis of neonatal mortality in Paraíba Valley, southeastern Brazil, 1999 to 2001. **Revista de Saúde Pública**, São Paulo, v. 41, p. 94–100, 2007.

NASCIMENTO, L. F. C.; ALMEIDA, M. C. S.; GOMES, C. D. M. S. Neonatal mortality and avoidable causes in the micro regions of São Paulo state. **Revista Brasileira de Ginecologia e Obstetrícia**, São Paulo, v. 36, p. 303–309, 2014.

NASCIMENTO, L. F. C.; MACHIN, A. B.; SANTOS, D. A. A. D. Are there differences in birth weight according to sex and associations with maternal exposure to air pollutants? a cohort study. **Sao Paulo Medical Journal**, São Paulo, v. 135, p. 347–354, 2017.

- NETO, O. L. M. *et al.* Differential patterns of neonatal and post-neonatal mortality rates in Goiânia, Brazil, 1992-1996: use of spatial analysis to identify high-risk areas. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1241–1250, 2001.
- NUSAIR, S. a. Testing the validity of purchasing power parity for asian countries during the current float. **Journal of Economic Development**, Coreia do Sul, v. 28, n. 2, p. 129–147, 2003.
- NYKIFORUK, C. I. J.; FLAMAN, L. M. Geographic information systems for health promotion and public health: a review. **Health promotion practice**, Los Angeles, v. 12, n. 1, p. 63–73, 2011.
- OBENSHAIN, M. K. Application of data mining techniques to healthcare data. **Infection Control e Hospital Epidemiology**, Nova Iorque, v. 25, n. 8, p. 690–695, 2004.
- OGALLO, W. *et al.* Identifying factors associated with neonatal mortality in Sub-Saharan Africa using machine learning. In: **AMIA ANNUAL SYMPOSIUM PROCEEDINGS**. Los Angeles: American Medical Informatics Association, 2020. p. 963–972.
- OH, M.-Y. *et al.* Machine-learning-based evaluation of the usefulness of lactate for predicting neonatal mortality in preterm infants. **Pediatrics and Neonatology**, Vietnã, 2024.
- ORTIZ, L. P.; OUSHIRO, D. A. Perfil da mortalidade neonatal no estado de São Paulo. **São Paulo em Perspectiva**, São Paulo, v. 22, n. 1, p. 19–29, 2008.
- PAIXAO, E. S. *et al.* Risk of mortality for small newborns in Brazil, 2011-2018: a national birth cohort study of 17.6 million records from routine register-based linked data. **The Lancet Regional Health-Americas**, Nova Iorque, v. 3, p. 100045, 2021.
- PARMEZAN, A. R. S.; SOUZA, V. M. a.; BATISTA, G. E. A. P. A. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. **Information sciences**, Virgínia, v. 484, p. 302–337, 2019.
- PETROPOULOS, F. *et al.* Forecasting: theory and practice. **International Journal of Forecasting**, Singapura, v. 38, n. 3, p. 705–871, 2022.
- PROBST, P.; BOULESTEIX, A.-L.; BISCHL, B. Tunability: importance of hyperparameters of machine learning algorithms. **The Journal of Machine Learning Research**, Nova Iorque, v. 20, n. 1, p. 1934–1965, 2019.
- PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. Novo Hamburgo: FEEVALE, 2013.
- PUGLIESE, R.; REGONDI, S.; MARINI, R. Machine learning-based approach: global trends, research directions, and regulatory standpoints. **Data Science and Management**, Sicília, v. 4, p. 19–29, 2021.
- RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. **Health information science and systems**, Nova Iorque, v. 2, p. 1–10, 2014.
- RAHMAN, A. *et al.* Machine learning algorithm for analysing infant mortality in Bangladesh. In: ASSOCIATION FOR COMPUTING MACHINER, 10., 2021, Melbourne. **Proceedings of 10th International Conference in Health Information Science**. Nova Iorque: ACM, 2021. p. 205–219.

- RAHMAN, A. *et al.* An assessment of random forest technique using simulation study: illustration with infant mortality in Bangladesh. **Health Information Science and Systems**, Nova Iorque, v. 10, p. 12, 2022.
- REYES, L.; MANALICH, R. Long-term consequences of low birth weight. **Kidney International**, Missouri, v. 68, p. S107–S111, 2005.
- RINTA-KOSK, O.-P. *et al.* In: COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING, 25., 2017, Bruges. **Proceedings of 25th European Symposium on Artificial Neural Networks**. Belgica: DBLP computer science, 2017. p. 193–198.
- RISTEVSKI, B.; SAVOSKA, S. Healthcare and medical big data analytics. In: **Applications of big data in healthcare: Theory and practice**. Nova Iorque: Academic Press, 2021. p. 85–112. ISBN 978-0-12-820203-6.
- RODRIGUES, M. *et al.* Using spatial analysis to identify areas vulnerable to infant mortality. **Revista Panamericana de Salud Pública**, Washington, v. 34, p. 36–40, 2013.
- RODRIGUES, M. *et al.* Análise espacial da mortalidade infantil e adequação das informações vitais: uma proposta para definição de áreas prioritárias. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 19, n. 7, p. 2047–2054, 2014.
- RODRIGUES, N. C. P. *et al.* Temporal and spatial evolution of maternal and neonatal mortality rates in Brazil, 1997-2012. **Jornal de Pediatria**, Rio de Janeiro, v. 92, n. 6, p. 567–573, 2016.
- RODRIGUEZ, E. Y. a. *et al.* A data-driven approach for neonatal mortality rate forecasting. In: EUROPEAN CONFERENCE ON GENETIC PROGRAMMING, 5., 2022, Lyon,. **Proceedings of the 5th International Conference on Informatics and Data-Driven Medicine**. Aquisgrán: CEUR Workshop Proceedings, 2022. p. 86–98.
- RODRÍGUEZ, E. C. A. *et al.* Padrões espaciais da prematuridade e seus determinantes na região metropolitana de São Paulo, Brasil, 2010-2019. **Revista Brasileira de Epidemiologia**, São Paulo, v. 27, p. e240008, 2024.
- RODRÍGUEZ, E. Y. A. *et al.* Comparison of adaptive neuro-fuzzy inference system (ANFIS) and machine learning algorithms for electricity production forecasting. **IEEE Latin America Transactions**, São Paulo, v. 20, n. 10, p. 2288–2294, 2022.
- RODRÍGUEZ, E. Y. A. *et al.* Spatial patterns of mortality in low birth weight infants at term and its determinants in the state of São Paulo, Brazil. **Revista Brasileira de Epidemiologia**, São Paulo, v. 26, p. 1–10, 2023.
- RONG, G. *et al.* Artificial intelligence in healthcare: review and prediction case studies. **Engineering**, Beijing, v. 6, n. 3, p. 291–301, 2020.
- SALIH, S. O. *et al.* Forecasting causes of death in northern Iraq using neural network. **Journal of Statistical Theory and Applications**, Paris, v. 21, n. 2, p. 58–77, 2022.
- SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, Nova Jersey, v. 3, n. 3, p. 210–229, 1959.

SANTOS, H. G. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. Orientador: Alexandre Dias Porto Chiavegatto Filho. 2018. Tese (Doutorado em Ciências) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018.

SANTOS, H. G. *et al.* Risk factors for infant mortality in a municipality in southern Brazil: a comparison of two cohorts using hierarchical analysis. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 28, n. 10, p. 1915–1926, 2012.

SARAVANOU, A. *et al.* Predictive modeling of infant mortality. **Data Mining and Knowledge Discovery**, 's-Hertogenbosch, v. 35, p. 1785–1807, 2021.

SAUGSTAD, O. D. Reducing global neonatal mortality is possible. **Neonatology**, Columbus, v. 99, n. 4, p. 250–257, 2011.

SAYEED, M. A. *et al.* On the interpretability of the SVM model for predicting infant mortality in Bangladesh. **Journal of Health, Population and Nutrition**, Dacca, v. 43, p. 170, 2024.

SCHOEPS, D. *et al.* Fatores de risco para mortalidade neonatal precoce. **Revista de Saúde Pública**, São Paulo, v. 41, p. 1013–1022, 2007.

SCHRAMM, J.; SZWARCOWALD, C. L. The Brazilian hospital system as a source of information to estimate stillbirth and neonatal mortality rates. **Revista de Saúde Pública**, São Paulo, v. 34, p. 272–279, 2000.

SHARIAT, N. *et al.* Neonatal mortality prediction in nicus: a machine learning approach. In: SCIENTIFIC INFORMATION DATABASE, 10., 2024, Teerã. **Proceedings of International Conference on Web Research**. Nova Iorque: IEEE, 2024. p. 231–237.

SHAW, P.; PACHPOR, K.; SANKARANARAYANAN, S. Explainable ai enabled infant mortality prediction based on neonatal sepsis. **Computer Systems Science and Engineering**, Henderson, v. 44, p. 311–325, 2023.

SHEIKHTAHERI, A. *et al.* Prediction of neonatal deaths in NICUs: development and validation of machine learning models. **BMC Medical Informatics and Decision Making**, Londres, v. 21, p. 131, 2021.

SHUKLA, V. v.; CARLO, W. A. Risk prediction for stillbirth and neonatal mortality in low-resource settings. **Newborn**, Nova Iorque, v. 1, n. 2, p. 215, 2022.

SHUKLA, V. v. *et al.* Social distancing during the covid-19 pandemic and neonatal mortality in the US. **JAMA Network Open**, Illinois, v. 7, n. 7, p. e2422995, 2024.

SIDNEY, S. Nonparametric statistics for the behavioral sciences. **The Journal of Nervous and Mental Disease**, Nashville, v. 125, n. 3, p. 497, 1957.

SILVA, A. A. M. *et al.* The epidemiologic paradox of low birth weight in Brazil. **Revista de Saúde Pública**, São Paulo, v. 44, p. 767–775, 2010.

SILVA, A. B. S. *et al.* Auto-regressive integrated moving average model: conceptual and methodological aspects and applicability in infant mortality. **Revista Brasileira de Saúde Materno Infantil**, Recife, v. 21, p. 647–656, 2021.

SILVA, A. F.; SILVA, J. D. P. Mortalidade infantil evitável em minas gerais: perfil epidemiológico e espacial.

SILVA, A. V. *et al.* Infant mortality in a Brazilian northeast state between 2009 and 2018: a temporal space approach. **Revista Brasileira de Saúde Materno Infantil**, Recife, v. 23, p. e20220032, 2023.

SILVA, L. P. *et al.* Avaliação da qualidade dos dados do sistema de informações sobre nascidos vivos e do sistema de informações sobre mortalidade no período neonatal, Espírito Santo, Brasil, de 2007 a 2009. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 19, p. 2011–2020, 2014.

SIN, J. **Commissioning and a population approach to health services decision-making**. São Francisco: Oxford University Press, 2020.

SIVARAJAN, a.; ADITYA, a. B.; SIVASANKAR, E. Comparing the predictive accuracy of machine learning algorithms for neonatal mortality risk classification. In: **ADVANCED MACHINE INTELLIGENCE AND SIGNAL PROCESSING**. Itanagar: Springer, 2022. p. 325–339.

SOARES, W. L. *et al.* Characterizing infant mortality using data mining: a case study in two Brazilian states Santa Catarina and Amapá. **Studies in Health Technology and Informatics**, Lansdale, v. 290, p. 772–776, 2022.

SOYIRI, I. N.; REIDPATH, D. D. An overview of health forecasting. **Environmental Health and Preventive Medicine**, Kyoto, v. 18, p. 1–9, 2013.

Statistical Analysis System. **SEMMA data mining methodology**. Cary: SAS, 2017. Disponível em: <https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jn8bbj1a2.htm>. Acesso em: 06 set. 2021.

SUZUKI, J.; SUZUKI, J. **Statistical learning with math and python**. Nova Iorque: Springer, 2021.

SWAMYNATHAN, M. **Mastering machine learning with python in six steps: a practical implementation guide to predictive data analytics using python**. Nova Iorque: Apress Berkeley, 2019.

TEJI, J. S. *et al.* NeoAI 1.0: Machine learning-based paradigm for prediction of neonatal and infant risk of death. **Computers in Biology and Medicine**, Nova Iorque, v. 147, p. 105639, 2022. ISSN 0010-4825.

TESEMA, G. A. *et al.* Incidence of infant mortality and its predictors in east africa using gompertz gamma shared frailty model. **Archives of Public Health**, Londres, v. 80, p. 1–12, 2022.

UN. **Millennium development goals report 2015**. Nova Iorque: United Nations, 2015. Disponível em: [https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20\(July%201\).pdf](https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf). Acesso em: 23 oct. 2021.

UN. **Transforming our world: the 2030 agenda for sustainable development**. Nova Iorque: United Nations, 2015. Disponível em: [https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20\(July%201\).pdf](https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf). Acesso em: 25 oct. 2021.

UNICEF. **UNICEF-WHO low birthweight estimates: levels and trends 2000-2015**. Nova Iorque: United Nations International Children's Emergency Fund, 2019. Disponível em: <https://www.unicef.org/media/96976/file/UNICEF-WHO-Low-Birthweight-estimates-2000-2015.pdf>. Acesso em: 14 set. 2023.

UNICEF. **Under-five mortality**. Nova Iorques: United Nations International Children's Emergency Fund, 2023. Disponível em: <https://data.unicef.org/topic/child-survival/under-five-mortality/>. Acesso em: 20 mar. 2024.

VENTURA, R. B. *et al.* Forecasting neonatal mortality in Portugal. **Engineering Proceedings: 9th International Conference on Time Series and Forecasting**, Gran Canaria, v. 39, p. 89, 2023.

VERCELLIS, C. **Business intelligence: data mining and optimization for decision making**. Hoboken: John Wiley and Sons, 2011.

VICTORA, C. G. *et al.* Maternal and child health in Brazil: progress and challenges. **The Lancet**, Londres, v. 377, n. 9780, p. 1863–1876, 2011.

VILANOVA, C. S. *et al.* The relationship between the different low birth weight strata of newborns with infant mortality and the influence of the main health determinants in the extreme south of Brazil. **Population health metrics**, Nova Iorque, v. 17, p. 1–12, 2019.

VILLAGRANA-BAÑUELOS, K. E. *et al.* Machine learning model based on lipidomic profile information to predict sudden infant death syndrome. **Healthcare**, Basel, v. 10, n. 7, p. 1303, 2022.

VILLAGRANA-BAÑUELOS, K. E. *et al.* Fatty chain acids risk factors in sudden infant death syndrome: a genetic algorithm approach. In: MEXICAN SOCIETY FOR ARTIFICIAL INTELLIGENCE, 19., 2020, Mexico. **Proceedings of 19th Mexican International Conference on Artificial Intelligence**. Cham: Springer, 22. p. 235–245.

WANG, Y.; KUNG, L.; BYRD, T. A. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. **Technological forecasting and social change**, Nova Iorque, v. 126, p. 3–13, 2018.

WEISSGERBER, T. L. *et al.* Data visualization, bar naked: a free tool for creating interactive graphics. **Journal of Biological Chemistry**, Maryland, v. 292, n. 50, p. 20592–20598, 2017.

WHO. **Global nutrition targets 2025: stunting policy brief**. Genebra: World Health Organization, 2014. Disponível em: https://apps.who.int/iris/bitstream/handle/10665/149020/WHO_NMH_NHD_14.5_eng.pdf?ua=1. Acesso em: 28 oct. 2022.

WHO. **Global health observatory data repository**. Genebra: World Health Organization, 2018. Disponível em: <https://apps.who.int/gho/data/node.main.ChildMort1002015?lang=en>. Acesso em: 14 set. 2023.

WHO. **Newborns: improving survival and well-being**. Genebra: World Health Organization, 2020. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality>. Acesso em: 18 oct. 2022.

WHO. **Using big data to inform health care: opportunities, challenges and considerations.** Genebra: World Health Organization, 2021. Disponível em: <https://www.who.int/europe/news/item/26-05-2021-using-big-data-to-inform-health-care-opportunities-challenges-and-considerations>. Acesso em: 12 ago. 2023.

WISE, P. H. Infant mortality. In: **INTERNATIONAL ENCYCLOPEDIA OF PUBLIC HEALTH**. 2nd ed. ed. Oxford: Academic Press, 2017. p. 216–221. ISBN 978-0-12-803708-9. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128036785002289>. Acesso em: 18 jul.2022.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: theory and practice. **Neurocomputing**, Amsterdã, v. 415, p. 295–316, 2020.

YANG, Y.-H. *et al.* Predicting early mortality and severe intraventricular hemorrhage in very-low birth weight preterm infants: a nationwide, multicenter study using machine learning. **Scientific Reports**, Filadélfia, v. 14, p. 10833, 2024.

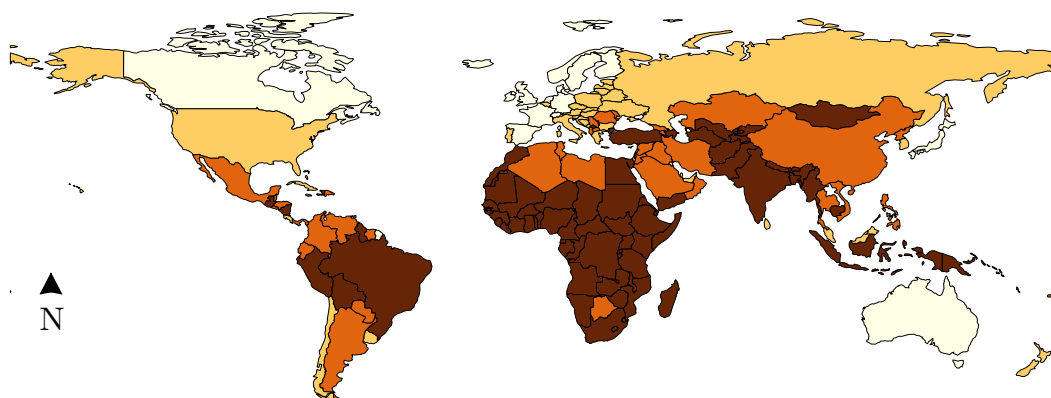
YONEDA, K. *et al.* Development of mortality prediction models for infants with isolated, left-sided congenital diaphragmatic hernia before and after birth. **Pediatric Pulmonology**, Nova Iorque, v. 58, p. 152–160, 2023.

ZHANG, Z.; XIAO, Q.; LUO, J. Infant death prediction using machine learning: a population-based retrospective study. **Computers in Biology and Medicine**, Nova Iorque, v. 165, p. 107423, 2023.

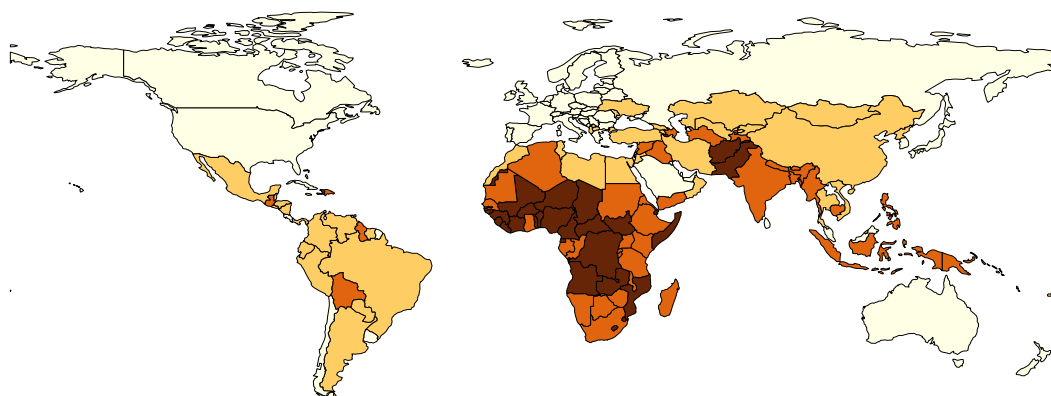
APÊNDICE A – DISTRIBUIÇÃO DAS TAXAS DE MORTALIDADE INFANTIL GLOBAL

Figura 76 – Distribuição das taxas de mortalidade infantil global, nos anos 1990, 2015 e 2020

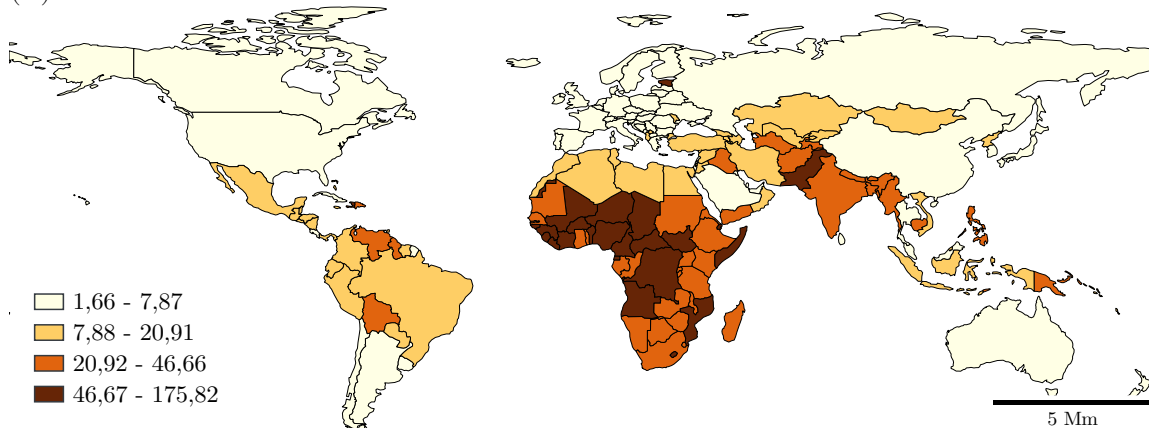
(A) Taxa de mortalidade infantil em 1990



(B) Taxa de mortalidade infantil em 2015



(C) Taxa de mortalidade infantil em 2020



Fonte: Elaborado pela autora (2025), com dados de UNICEF (2023).

APÊNDICE B – TENDÊNCIA DA ANÁLISE ESPACIAL E APRENDIZADO DE MÁQUINA NO CUIDADO DA SAÚDE

Foi realizada uma busca nas bases de dados PubMed e Scopus para identificar a tendência da aplicação de análise espacial e ML em pesquisas relacionadas ao cuidado da saúde. A busca foi refinada considerando os documentos publicados até 3 de novembro de 2024, incluindo apenas artigos originais e artigos de anais de eventos.

O Quadro 9 apresenta as palavras-chaves consideradas na busca e a quantidade de documentos encontrados em cada base de dados.

Quadro 9 – Análise espacial e aprendizado de máquina no cuidado da saúde

Palavras-chave	Scopus	PubMed
Análise espacial		
<i>TITLE-ABS-KEY("spatial analysis"AND ("health care"OR "healthcare"))</i>	2.476	521
ML e predição		
<i>TITLE-ABS-KEY("machine learning"AND ("health care"OR "healthcare"))</i>	24.107	10.010
ML e previsão		
<i>TITLE-ABS-KEY("machine learning"AND ("time series" OR "forecast") AND ("health care"OR "healthcare"))</i>	1.231	294

Fonte: Elaborado pela autora (2025).

Conforme o Quadro 9, a busca relacionada com análise espacial no cuidado da saúde resultou em 2.476 documentos recuperados de Scopus e 521 documentos de PubMed. Depois de uma filtragem dos documentos foram encontrados 420 artigos duplicados, resultando em 2.577 documentos com 41.921 citações acumuladas no período de 1972 a 2024. A tendência das publicações e citações da busca são apresentadas na Figura 77 (A).

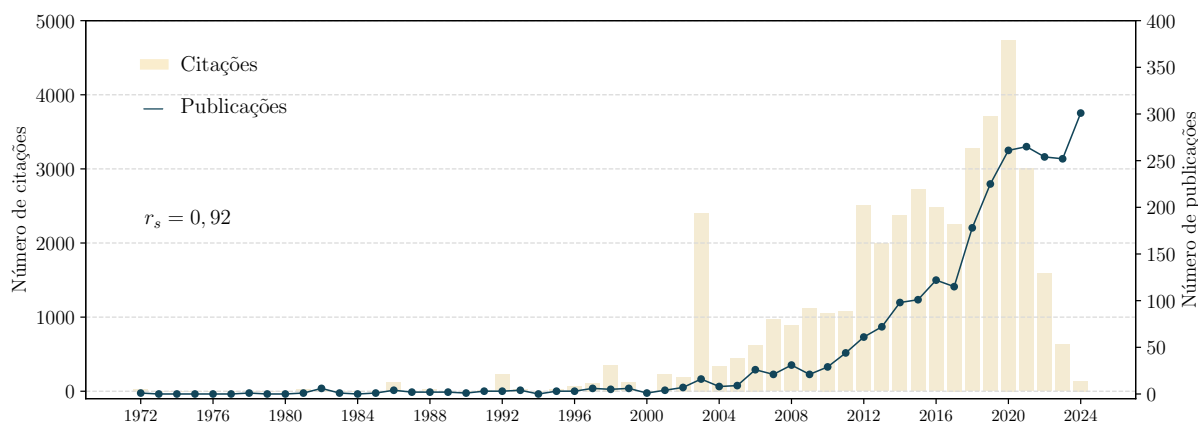
Conforme a Figura 77 (A) o número de publicações nesta área de estudo vem aumentando a cada ano, com uma taxa de crescimento anual de 9,86%. Também, foi observado um aumento no número de publicações a partir de 2006, com 96,12% (2.477) artigos publicados no período 2006 – 2024. Além disso, adotando-se um nível de significância estatística de 5%, o coeficiente de Spearman (r_s) mostrou uma correlação positiva entre o número total de publicações e o número total de citações ($r_s = 0,92$, p-valor < 0,001).

Do mesmo modo, na análise das publicações sobre ML no cuidado da saúde foram encontrados 24.107 documentos de Scopus e 10.010 de PubMed. Depois da filtragem de documentos semelhantes, encontraram-se 8.287 documentos duplicados, resultando em 25.830 documentos únicos com 259.191 citações acumuladas no período de 1994 a 2023. Na Figura 77 (B) apresenta-se a tendência das publicações e citações.

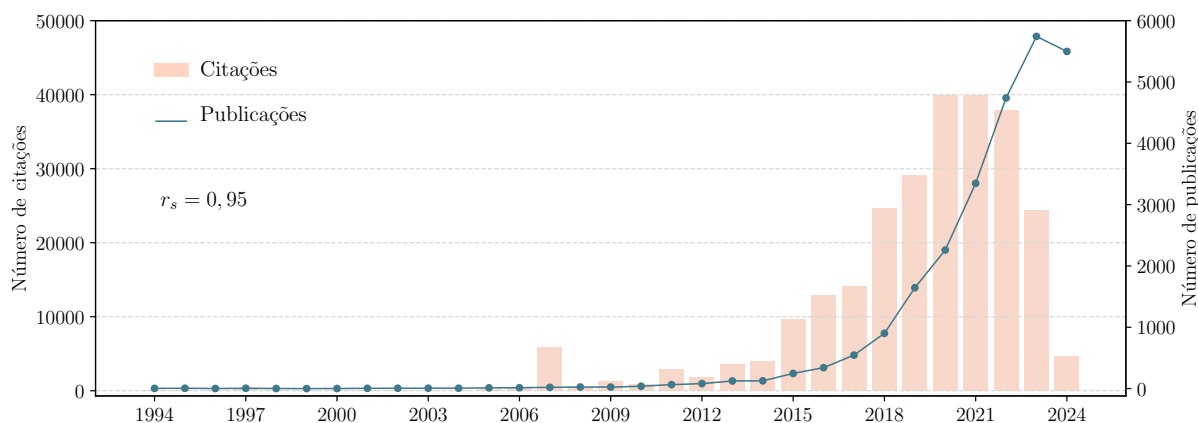
Na análise da distribuição anual das publicações observou-se um maior número de publicações nos últimos 20 anos, sendo que 99,86% (25.793) dos documentos foram publicados entre o

Figura 77 – Tendência das publicações e citações sobre (A) análise espacial, aprendizado de máquina para (B) predição e (C) previsão no cuidado da saúde

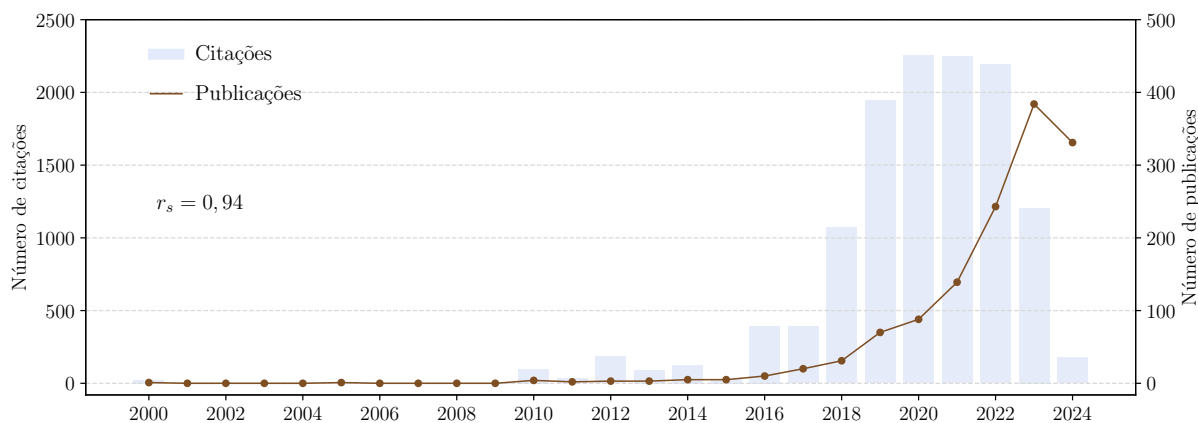
(A) Análise espacial



(B) Aprendizado de máquina para predição



(C) Aprendizado de máquina para previsão



Fonte: Elaborado pela autora (2025).

período de 2005 e 2024. Além disso, a taxa de crescimento anual das publicações nesta área de estudo foi de 26,76%. O número total de publicações e o número de citações no período analisado apresentaram uma associação significativa positiva ($r_s = 0,95, p - \text{valor} < 0,001$).

Por outro lado, entre as publicações de ML aplicados na previsão de eventos, foram encontrados 1.231 documentos recuperados de Scopus e 294 documentos de PubMed. Na filtragem do conjunto de dados foram encontradas 186 documentos duplicados, resultando em 1.339 documentos únicos com 12.466 citações acumuladas no período de 2000 a 2024.

A distribuição anual dos documentos de ML na previsão de eventos foram apresentados na Figura 77 (C), observando-se um aumento no número de publicações ao longo do tempo, com uma taxa de crescimento anual de 22,87%. Especificamente desde 2016, foi observado um aumento maior no número de artigos, sendo que 98,21% (1.316) dos documentos foram publicados entre 2016 e 2024.

Por último, o número de publicações e de citações nesta área de análise apresentaram uma associação significativa positiva ($r_s = 0,93, p - \text{valor} < 0,001$).

Assim, os gráficos da Figura 77 fornecem uma perspectiva visual sobre o interesse dos pesquisadores na aplicação da análise espacial e técnicas de ML neste campo de estudo. Portanto, espera-se que a aplicação destas técnicas no cuidado da saúde aumente, permitindo o desenvolvimento de diferentes ferramentas de apoio à tomada de decisão.

APÊNDICE C – CONCEITOS E TÉCNICAS PARA ANÁLISE EXPLORATÓRIA DE DADOS

A análise exploratória do banco de dados é a etapa inicial, na qual é possível identificar padrões e tendências nos dados por meio de técnicas estatísticas e gráficos de visualização, para entender o comportamento dos dados, identificar dados ausentes ou valores atípicos (*outliers*) (Faceli *et al.*, 2021). Dependendo do tipo de variável do conjunto de dados, pode-se realizar uma análise descritiva univariada com análises individuais, ou uma análise multivariada para avaliar suas relações. Seguidamente, podem ser aplicados métodos de transformação, padronização e/ou normalização para processar os dados (Goos; Meintrup, 2015).

ANÁLISE UNIVARIADA

A análise univariada do conjunto de dados é realizada utilizando as informações de cada variável $\mathbf{x}_j = \{x_1, x_2, x_3, \dots, x_n\}$, sendo cada x_i ($i = 1, 2, 3, \dots, n$) uma característica da variável $\mathbf{x}_j$ (Mohr; Wilson; Freund, 2021). A análise univariada das variáveis pode ser realizada avaliando-se a localização e espalhamento, além de utilizar métodos gráficos para visualização a distribuição dos dados:

- **Medidas de localização:**

As medidas de localização permitem descrever o conjunto de dados avaliando-se sua tendência central e distribuição dos dados. As medidas utilizadas serão a média, mediana ou segundo quartil (Q_2), primeiro quartil (Q_1), terceiro quartil (Q_3), valor mínimo e máximo. Estas medidas são descritas a seguir, conforme Faceli *et al.* (2021) e Mohr, Wilson & Freund (2021):

- *Média*: Medida muito sensível a valores atípicos, sendo considerada um bom indicador quando os dados da variável estão distribuídos simetricamente. A média denotado por $\bar{x}_j$ pode ser calculada pela Equação C.1:

$$média(\mathbf{x}_j) = \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_i \quad (C.1)$$

- *Mediana ou segundo quartil*: Definido como o valor que ocupa a posição central da variável que está sendo analisada, ordenada de forma crescente. Ao contrario da média, esta medida é menos sensível a valores discrepantes. A mediana é calculada conforme a Equação C.2:

$$mediana(\mathbf{x}_j) = \begin{cases} \frac{1}{2}(x_i + x_{i+1}), & \text{Se } n \text{ for par } (n = 2i) \\ x_{i+1}, & \text{Se } n \text{ for ímpar } (n = 2i + 1) \end{cases} \quad (C.2)$$

sendo n o número total de valores da variável $\mathbf{x}_j$. A mediana também é o segundo quartil (Q_2), permitindo que os dados sejam divididos em duas partes iguais (50:50).

- *Primeiro quartil*: Representa 25% do total dos dados, ou seja, os valores inferiores a Q_1 representam 25% dos dados e 75% dos valores restantes são superiores a Q_1 . O Q_1 é calculado pela Equação C.3 com os dados da variável ordenados de forma crescente:

$$Q_1 = \begin{cases} q = \frac{n+2}{4}, & \text{Se } n \text{ for par } (n = 2i) \\ q = \frac{n+3}{4}, & \text{Se } n \text{ for ímpar } (n = 2i + 1) \end{cases} \quad (\text{C.3})$$

sendo q um valor inteiro que indica a posição do Q_1 , dos valores ordenados de forma crescente da variável $\mathbf{x}_j$.

- *Terceiro quartil*: Semelhante ao primeiro (Q_1), o Q_3 é calculado pela Equação C.4 com os dados ordenados. O 25% dos dados são superiores a Q_3 e 75% restante são inferiores a Q_3 :

$$Q_3 = \begin{cases} q = \frac{3n+2}{4}, & \text{Se } n \text{ for par } (n = 2i) \\ q = \frac{3n+1}{4}, & \text{Se } n \text{ for ímpar } (n = 2i + 1) \end{cases} \quad (\text{C.4})$$

- *Valor mínimo e máximo*: Corresponde aos valores mínimo e máximo da variável de análise, sendo calculado pela Equação C.5:

$$\text{mínimo}(\mathbf{x}_j) = \min_{i=1,\dots,n} (x_i) \quad \wedge \quad \text{máximo}(\mathbf{x}_j) = \max_{i=1,\dots,n} (x_i) \quad (\text{C.5})$$

• **Medidas de espalhamento:**

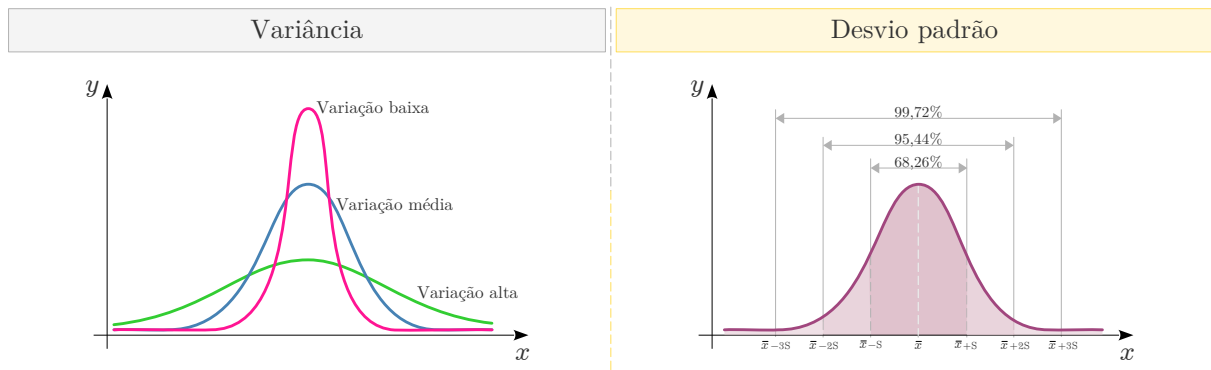
As medidas de espalhamento medem a dispersão dos valores da variável $\mathbf{x}_j$ do conjunto de dados, sendo as medidas mais utilizadas a variância e desvio padrão, apresentados na Figura 78:

- *Variância*: Utilizada para medir a dispersão dos dados, avaliando especificamente a diferença entre cada elemento da variável de estudo em relação à sua média (Mohr; Wilson; Freund, 2021). A variância é calculada pela Equação C.6:

$$\text{Variância}(\mathbf{x}_j) = S_j^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mathbf{x}}_j)^2 \quad (\text{C.6})$$

sendo $\bar{\mathbf{x}}_d$ a média dos valores da variável e o fator $n-1$ e a correção de Bessel da variância amostral (Faceli *et al.*, 2021). Além disso, o valor da variância não é negativo, e um valor baixo indica valores próximos da média amostral enquanto valores altos indicam maior variabilidade nos dados (Goos; Meintrup, 2015), conforme ilustrado na Figura 78.

Figura 78 – Medidas multivariadas de covariância e correlação



Fonte: Elaborado pela autora (2025).

- *Desvio padrão*: Esta medida é utilizada para comparar a dispersão dos dados, e é calculada pela raiz quadrada da variância (Equação C.6) (Faceli *et al.*, 2021), como apresentado na Equação C.7:

$$Desvio(\mathbf{x}_j) = S_j = \sqrt{S_j^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_j)^2} \quad (C.7)$$

Outra forma de análise é utilizar a média $\bar{x}_j$ e o desvio padrão S_j em conjunto, pois podem fornecer informações relevantes sobre a distribuição dos valores do conjunto de dados, como apresentado na Figura 78.

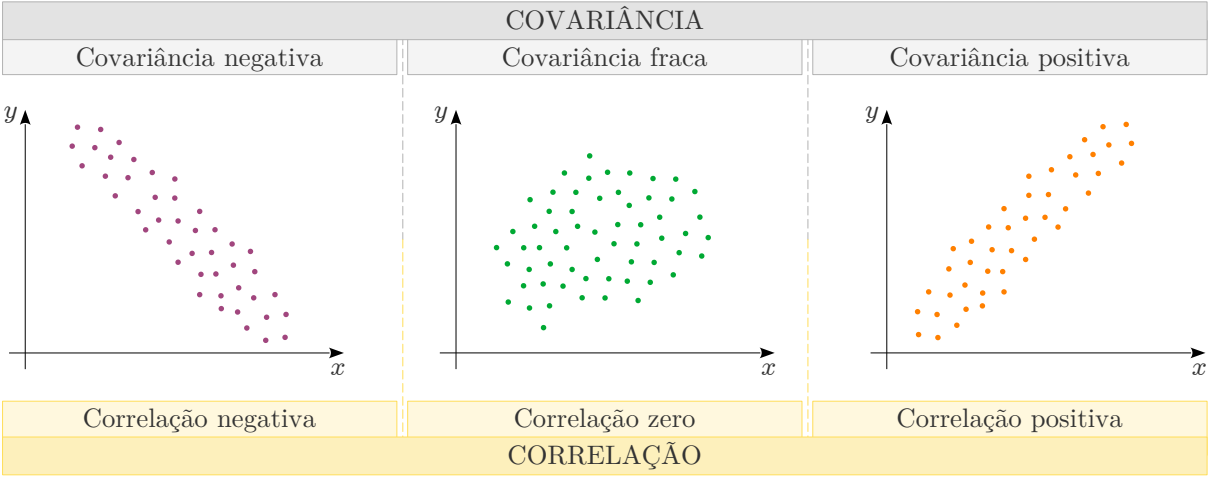
Em conformidade com Mohr, Wilson & Freund (2021) e a Figura 78, caso os dados em análise tenham distribuição normal, podem ser consideradas as seguintes avaliações:

- * $[\bar{x}_j \times S_j]$, contém aproximadamente 68,26% das observações;
- * $[\bar{x}_j \times 2S_j]$, contém aproximadamente 95,44% das observações e
- * $[\bar{x}_j \times 3S_j]$, contém aproximadamente 99,72% das observações, sendo $\bar{x}_j$ a média que descreve a localização e S_j o desvio padrão que descreve a dispersão de uma proporção de dados da variável $\mathbf{x}_j$.

ANÁLISE MULTIVARIADA

A maioria dos banco de dados contém mais de uma variável, deste modo, existem algumas medidas de espalhamento, como covariância e correlação, que podem ser utilizadas para analisar a relação linear entre elas (Goos; Meintrup, 2015; Faceli *et al.*, 2021). A Figura 79 apresenta de forma ilustrativa as medidas de covariância e correlação.

Figura 79 – Medidas multivariadas de covariância e correlação



Fonte: Elaborado pela autora (2025)

• **Medidas de espalhamento:**

- **Covariância:** A medida de covariância é calculada pela Equação C.8, medindo a covariância e seu grau de variação entre as variáveis $\mathbf{x}_j$ e $\mathbf{x}_k$:

$$Covariância(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{n - 1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{C.8}$$

sendo $\bar{x}_j$ e $\bar{x}_k$ o valor médio, x_{ij} e x_{ik} são elementos das variáveis, resultando que a diagonal da matriz é a variância de cada variável (Faceli *et al.*, 2021).

O resultado de $(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ pode ser positivo se os valores de x_{ij} e x_{ik} forem maiores que o valor médio ou menores que seu valor médio de cada variável. Caso contrário, o valor de $(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ é negativo se os valores de x_{ij} e x_{ik} são maiores e menores que sua média respectivamente (ou vice-versa). Assim, o valor da covariância da variável é negativo se a maioria das diferenças $(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ da variável for negativa, da mesma forma que a covariância positiva (Goos; Meintrup, 2015).

- **Correlação:** Também é uma medida de associação, que mede o grau de relacionamento e força entre duas variáveis (Mohr; Wilson; Freund, 2021). Duas medidas bastante conhecidas são a correlação de Pearson e Spearman, que variam entre -1 e $+1$, indicando relação direta ou inversa entre as variáveis (Sidney, 1957; Goos; Meintrup, 2015).

A correlação de Pearson (r) mede a relação linear entre duas variáveis contínuas (Goos; Meintrup, 2015; Mohr; Wilson; Freund, 2021). O coeficiente de correlação r

é calculada pela Equação C.9:

$$\text{Correlação}(\mathbf{x}_j, \mathbf{x}_k) = r = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}} \quad (\text{C.9})$$

Por outro lado, a correlação de Spearman (r_s) mede a relação monotônica entre duas variáveis contínuas ou ordinais (Sidney, 1957; Mohr; Wilson; Freund, 2021). O coeficiente de correlação r_s é calculada pela Equação C.10:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (\text{C.10})$$

sendo $d_i = x_{ij} - x_{ik}$ a diferença entre os intervalos das duas variáveis $\mathbf{x}_j$ e $\mathbf{x}_k$.

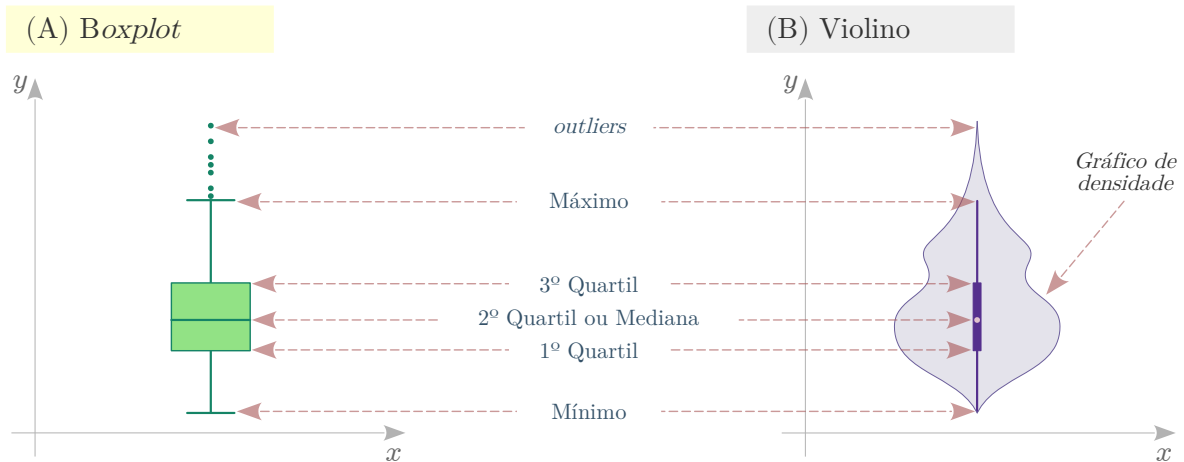
MÉTODOS GRÁFICOS DE VISUALIZAÇÃO

Existem diferentes métodos gráficos univariados ou multivariados que podem ser utilizados para explorar e visualizar o comportamento e distribuição dos dados quantitativos ou ordinais. Entre os métodos gráficos mais utilizados poder ser listado o gráfico de *boxplot*, gráfico de violino, gráfico de barras, gráfico de setores (ou pizza), gráfico temporal e gráfico de dispersão.

- **Gráfico de *boxplot* e violino:**

O gráfico de *boxplot*, também conhecido como gráfico de caixa, apresenta as informações relativas às cinco medidas como: valor mínimo, primeiro quartil, mediana, terceiro quartil e valor máximo, além de apresentar valores discrepantes (*outliers*) do conjunto de dados (Faceli *et al.*, 2021; Han; Pei; Tong, 2022), conforme ilustrado na Figura 80.

Figura 80 – Gráfico de *boxplot* e violino



Fonte: Elaborado pela autora (2025).

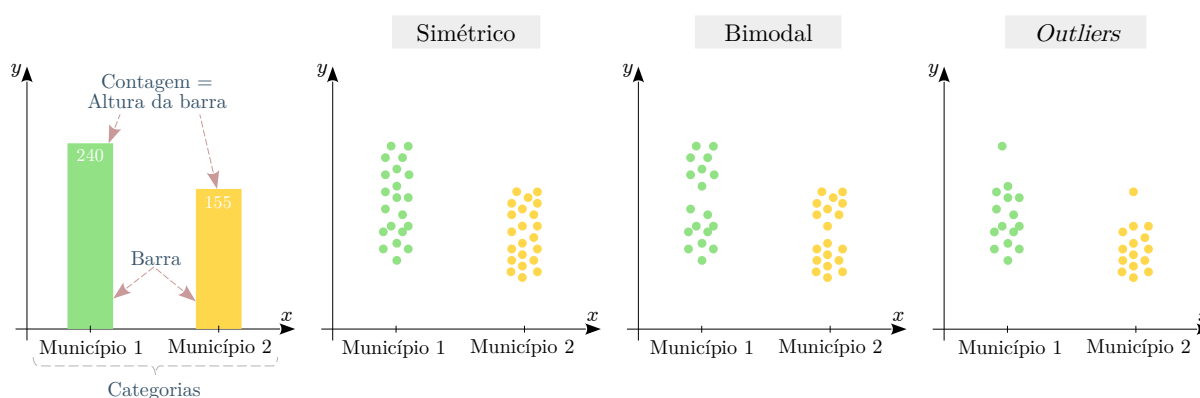
De forma semelhante ao gráfico de caixa, o gráfico de violino mostra informações semelhantes relacionadas à distribuição dos dados, como valores máximos, mínimos e quartis,

conforme mostrado na Figura 80. Complementarmente, o gráfico de violino apresenta a distribuição de dados mostrando a densidade de probabilidade da variável analisada, sendo mais eficaz para mostrar a forma da distribuição dos dados (Figura 80), como distribuições simétricas, bimodais (ou multimodal), além de apresentar valores atípicos (Weissgerber *et al.*, 2017).

- **Gráfico de barras:**

Estes gráficos estão projetados para variáveis categóricas mostrando dados contínuos da contagem e proporções de cada uma das categorias (Weissgerber *et al.*, 2017), como apresentado na Figura 81.

Figura 81 – Gráfico de barras



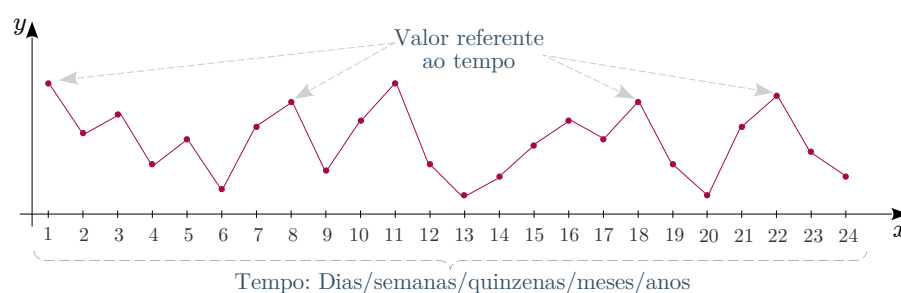
Fonte: Elaborado pela autora (2025).

Na literatura, alguns autores não recomendam o uso do gráficos de barras, pois um dos seus principais problemas é que não consideram ou ajudam a distinguir a distribuição dos dados (Weissgerber *et al.*, 2017; Han; Pei; Tong, 2022), conforme foi ilustrado na Figura 81.

- **Gráfico de linhas ou temporal:**

A Figura 82 apresenta um exemplo ilustrativo de gráfico temporal.

Figura 82 – Gráfico de linhas ou temporal



Fonte: Elaborado pela autora (2025).

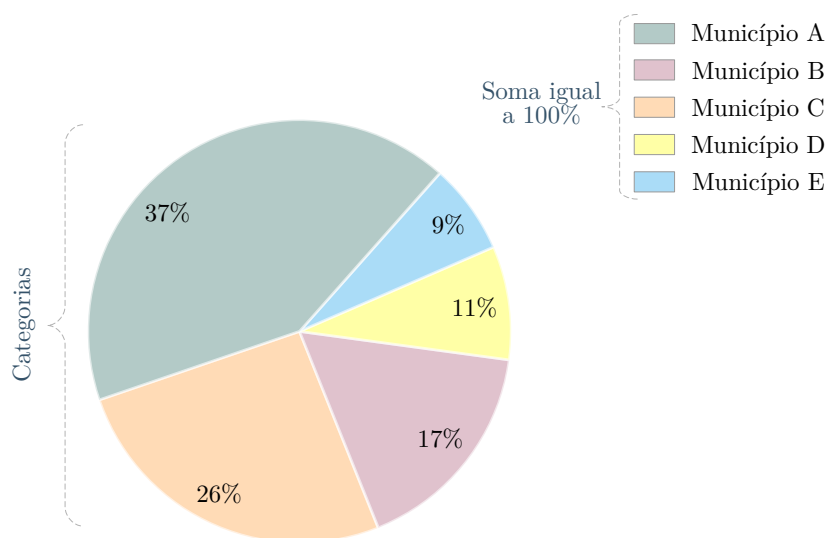
Este tipo de gráficos é bastante popular em diversas áreas de pesquisa e, especificamente em estudos da área de saúde é frequentemente utilizado para mostrar tendências longitudinais (Munro, 2012).

Como observado na Figura 82, os pontos relativos ao tempo são colocados no eixo horizontal e escala da variável analisada no eixo vertical, assim são traçadas linhas que conectam cada ponto do tempo (Munro, 2012).

- **Gráfico de setores:**

Também conhecido como gráfico de *pizza* é outro tipo de gráfico que apresenta a distribuição de dados, e possui formato circular dividido em proporções, cujas divisões estão relacionadas à frequência ou frequência relativa de cada categoria da variável (Faceli *et al.*, 2021; Mohr; Wilson; Freund, 2021). A Figura 83 apresenta um exemplo ilustrativo de gráfico de setores.

Figura 83 – Gráfico de setores

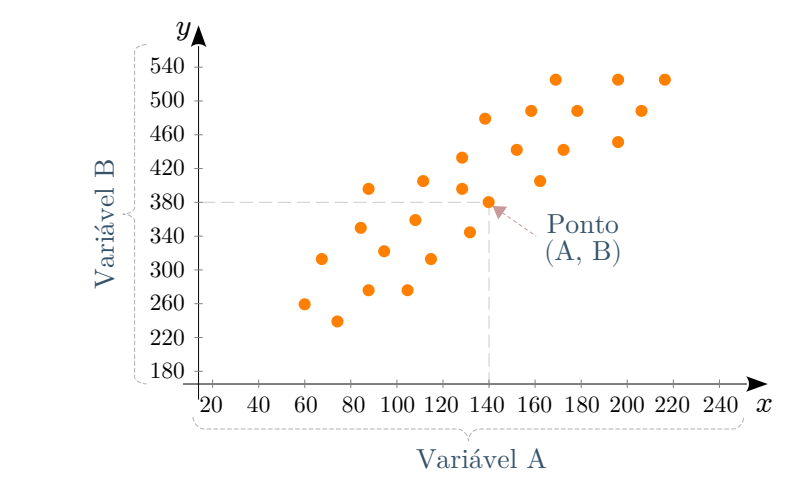


Fonte: Elaborado pela autora (2025).

- **Gráfico de dispersão:**

Também conhecido como *scatter plot* em inglês, é um dos gráficos mais utilizados para analisar e visualizar a associação entre duas variáveis (Goos; Meintrup, 2015; Mohr; Wilson; Freund, 2021). Além disso, este tipo de gráfico mostra a associação entre duas variáveis traçando vários pontos em um plano bidimensional (Faceli *et al.*, 2021), conforme apresentado na Figura 84.

Figura 84 – Gráfico de dispersão



Fonte: Elaborado pela autora (2025).

APÊNDICE D – CÓDIGO PARA CONVERSÃO DE ARQUIVOS DBC PARA CSV

Para a conversão dos arquivos obtidos dos sistemas SINASC e SIM no formato DBC¹ para o formato CSV², utilizou-se o software R e o ambiente RStudio v2023.06.1+524. O processo de conversão foi implementado com um código simples, conforme apresentado abaixo:

```

1      # Pacotes necessários
2      library(read.dbc)
3      library(readr)
4
5      # Função para converter DBC para CSV
6      dbc_to_csv <- function(sistema) {
7        if (sistema == "SINASC") {
8          anos <- 2010:2020
9          nm_sistema <- "DNSP"
10       } else if (sistema == "SIM") {
11         anos <- 2010:2021
12         nm_sistema <- "DOINF"
13       } else {
14         stop("Sistema inválido. Use 'SINASC' ou 'SIM'.")
15       }
16       for (ano in anos) {
17         arquivo_dbc <- paste0(nm_sistema, ano, ".dbc")
18         arquivo_csv <- paste0(sistema, "_", ano, ".csv")
19
20         dados <- read.dbc(arquivo_dbc) # Ler o arquivo DBC
21         dados <- data.frame(dados)
22         write_csv(dados, arquivo_csv) # Salvar como CSV
23
24         cat("Arquivo", arquivo_dbc, "convertido para", arquivo_csv, "\n")
25       }
26     }
27     # Converter arquivos SINASC
28     dbc_to_csv("SINASC")
29     # Converter arquivos SIM
30     dbc_to_csv("SIM")

```

O código apresentado realiza a leitura do arquivo DBC, utilizando a função `read.dbc()`, e em seguida, converte o arquivo para o formato CSV com a função `write_csv()`. Mais detalhes sobre o uso dessas funções podem ser encontrados na documentação do pacote `read.dbc`.

¹ Abreviação do termo inglês DataBase Container.

² Abreviação do termo inglês Comma Separated Values, que é um arquivo simples delimitando por vírgulas.

APÊNDICE E – SCRIPTS PARA VERIFICAÇÃO, CONSOLIDAÇÃO E PREPARAÇÃO DO BANCO DE DADOS

Para o desenvolvimento da pesquisa, foi utilizado o código Python v.3.9.7 no ambiente Jupyter Notebook v.6.4.6. O código utilizado para a Etapa 2 sobre a consolidação, filtragem, rotulagem, qualidade, limpeza e codificação dos bancos de dados pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/drive/folders/1rVDLcAbYahVvwwcvoxXz1UqEIQy8lZ9I?usp=drive_link.

As bibliotecas utilizadas na Etapa 2 para a manipulação e análise dos dados:

- Path de pathlib: Utilizada para a manipulação e controle dos caminhos de arquivos e diretórios.
- Pandas: Utilizado para gerenciamento e processamento dos dados.
- NumPy: Utilizado para processamento dos dados.
- Missingno: Utilizado para visualização e análise de dados ausentes.

VERIFICAÇÃO DA QUALIDADE

Inicialmente, foi realizado uma análise e verificação da qualidade dos conjuntos de dados do SIM e do SINASC. Especificamente, foi avaliada a presença de valores ausentes e de valores mal preenchidos. O código utilizado para a análise da qualidade dos dados pode ser acessado através do seguinte link: https://drive.google.com/file/d/1NaaAfnYtAotiRReKCz3KzaLt3CDJoCDF/view?usp=drive_link.

CONSOLIDAÇÃO E FILTRAGEM

Os bancos de dados dos sistemas SINASC e SIM fornecem informações de todos os municípios do estado de São Paulo. Portanto, foi necessário filtrar os dados para selecionar os registros dos municípios da RMVPLN.

```

1     import pandas as pd
2     import numpy as np
3     from pathlib import Path
4     # Código dos municípios da RMVPLN
5     municipios = [350970, 353170, 354820, 354860, 350850, 352020,
                   352440, 353800, 354600, 354990, 355410, 355480, 350250,
                   350860, 350995, 351340, 351840, 352660, 352720, 353850, 354075,
                   354190, 354430, 350315, 350350, 350490, 354960, 355200,
                   351360, 352490, 352630, 353230, 353560, 354230, 355000, 351050,
                   352040, 355070, 355540]
```

```

6     Nm_municipios = ['Campos do Jordão', 'Monteiro Lobato', 'Santo Antônio do Pinhal', 'São Bento do Sapucaí', 'Caçapava', 'Igaratá', 'Jacareí', 'Pindamonhangaba', 'Santa Branca', 'São José dos Campos', 'Taubaté', 'Tremembé', 'Aparecida', 'Cachoeira Paulista', 'Canas', 'Cruzeiro', 'Guaratinguetá', 'Lavrinhas', 'Lorena', 'Piquete', 'Potim', 'Queluz', 'Roseira', 'Arapeí', 'Areias', 'Bananal', 'São José do Barreiro', 'Silveiras', 'Cunha', 'Jambeiro', 'Lagoinha', 'Natividade da Serra', 'Paraibuna', 'Redenção da Serra', 'São Luiz do Paraitinga', 'Caraguatatuba', 'Ilhabela', 'São Sebastião', 'Ubatuba']
7
8     Municipios = pd.DataFrame({'Municipios': municipios,
9                                'Nm_municipios': Nm_municipios})
10
11    # Função para
12    def SIM_SINASC_VPP(base, ano):
13        path = (str(base) + "/" + str(base) + "_" + str(ano) + ".csv")
14        bd = pd.read_csv(path, low_memory=False)
15        bd = bd[bd.CODMUNRES.isin(Municipios)]
16        save_path = Path(str(base) + "/Mess-VPP_" + str(base) + "/" + str(base) + "_VPP_" + str(ano) + ".csv")
17        bd.to_csv(save_path)
18        return bd
19
20    for ano in range(2010, 2022): # 2010 a 2021
21        SIM_SINASC_VPP(base = "SIM", ano = ano)
22
23    for ano in range(2010, 2021): # 2010 a 2020
24        SIM_SINASC_VPP(base = "SINASC", ano = ano)

```

De modo similar, foi necessário adicionar uma nova coluna denominada ANO, com o objetivo de facilitar a análise dos dados.

```

1     from pathlib import Path
2
3     def load_SIM_SINASC(base, ano):
4         path = (str(base) + "/Mess-VPP_" + str(base) + "/" + str(base) + "_VPP_" + str(ano) + ".csv")
5         bd = pd.read_csv(path, low_memory=False)
6         return bd
7
8     dados_SIM = {}
9     print(f'Número de linhas e colunas por base de dados por ano (SIM)')
10
11    for ano in range(2010, 2022):
12        dados_SIM[ano] = load_SIM_SINASC(base="SIM", ano=ano)
13        dados_SIM[ano]['ANO'] = ano
14        print(f'- {ano}: {len(dados_SIM[ano].axes[1])} colunas e {len(dados_SIM[ano].axes[0])} linhas.')

```

```

15 dados_SINASC = {}
16 print(f'Número de linhas e colunas por base de dados por ano (
    SINASC)')
17 for ano in range(2010, 2021):
18     dados_SINASC[ano] = load_SIM_SINASC(base="SINASC", ano=ano)
19     dados_SINASC[ano]['ANO'] = ano
20     print(f'- {ano}: {len(dados_SINASC[ano].axes[1])} colunas e {len(
        dados_SINASC[ano].axes[0])} linhas.')

```

Foram também adicionadas as colunas MUNICÍPIOS e SUB-REGIÃO, que contêm o nome do município e sua classificação por sub-região.

```

1     # Sub-regiões (5):
2     Sr_saoJdC = [350850, 352020, 352440, 352490, 353560, 354600,
3         353170, 354990]
4     Sr_taub = [352630, 353230, 350970, 353800, 354230, 354860, 355000,
5         354820, 355410, 355480]
6     Sr_guar = [350250, 354430, 351360, 353850, 354075, 351840, 352720,
7         350995, 350860]
8     Sr_cruz = [350315, 350490, 351340, 350350, 355200, 352660, 354190,
9         354960]
10    Sr_litNor = [351050, 352040, 355070, 355540]
11
12    def Munic(database):
13        lst = []
14        datab= database.CODMUNRES
15        for i in range(0,len(datab)):
16            for j in range(0,len(Municipios.Municipios)):
17                if datab[i] == Municipios.Municipios[j]:
18                    lst.append(Municipios.Nm_municipios[j])
19            database['MUNICIPIOS'] = lst
20        return database
21
22    def SubR(database):
23        lst = []
24        datab= database.CODMUNRES
25        for i in range(0,len(datab)):
26            SubReg = [Sr_saoJdC, Sr_taub, Sr_guar, Sr_cruz, Sr_litNor]
27            for a in range(0,len(SubReg)):
28                sr = SubReg[a]
29                for j in range(0,len(sr)):
30                    if datab[i] == sr[j]:
31                        if a == 0:
32                            SubRegiao = 'São José dos Campos'
33                        elif a == 1:
34                            SubRegiao = 'Taubaté'
35                        elif a == 2:
36                            SubRegiao = 'Guaratinguetá'
37                        elif a == 3:
38                            SubRegiao = 'Cruzeiro'
39                        elif a == 4:
40                            SubRegiao = 'Lins'
41                        lst.append(SubRegiao)

```

```

33     elif a == 2:
34         SubRegiao = 'Guaratinguetá'
35         lst.append(SubRegiao)
36     elif a == 3:
37         SubRegiao = 'Cruzeiro'
38         lst.append(SubRegiao)
39     elif a == 4:
40         SubRegiao = 'Litoral Norte'
41         lst.append(SubRegiao)
42     database['SUB-REGIAO'] = lst
43     return database
44
45     for ano in range(2010, 2022):
46         dados_SIM[ano] = Munic(dados_SIM[ano])
47         dados_SIM[ano] = SubR(dados_SIM[ano])
48
49     for ano in range(2010, 2021):
50         dados_SINASC[ano] = Munic(dados_SINASC[ano])
51         dados_SINASC[ano] = SubR(dados_SINASC[ano])

```

Em seguida, como cada banco de dados anual apresenta um número diferente de colunas, foi realizado um processo de seleção de características, seguido pela consolidação dos dados em um único documento, tanto para o SIM quanto para o SINASC.

```

1     col_sim = ['CODESTAB', 'TIPOBITO', 'DFOBITO', 'HORAOBITO', 'DTNASC', '
        IDADE', 'SEXO', 'RACACOR', 'CODMUNRES', 'LOCOCOR', 'CODMUNOCOR', '
        IDADEMAE', 'ESMAE', 'ESMAE2010', 'OCUPMAE', 'QTDFILVIVO', '
        QTDFILMORT', 'GRAVIDEZ', 'SEMAGESTAC', 'GESTACAO', 'PARTO', '
        PESO', 'TPMORTEOCO', 'ASSISTMED', 'CIRURGIA', 'LINHAA', 'LINHAB
        ', 'LINHAC', 'LINHAD', 'LINHAI', 'CAUSABAS', 'CIRCOBITO', '
        CAUSAMAT', 'TPOBITOCOR', 'MORTEPARTO', 'CODMUNNATU', '
        ESMAEAGR1', 'ANO', 'MUNICIPIOS', 'SUB-REGIAO']
2     col_sinasc = ['CODESTAB', 'CODMUNNASC', 'LOCNASC', 'IDADEMAE', '
        ESTCIVMAE', 'ESMAE', 'CODOCUPMAE', 'QTDFILVIVO', 'QTDFILMORT',
        'CODMUNRES', 'GESTACAO', 'GRAVIDEZ', 'PARTO', 'CONSULTAS', '
        DTNASC', 'HORANASC', 'SEXO', 'APGAR1', 'APGAR5', 'RACACOR', '
        PESO', 'IDANOMAL', 'CODANOMAL', 'CODMUNNATU', 'DTNASCMAE', '
        RACACORMAE', 'QTDGESTANT', 'QTDPARTNOR', 'QTDPARTCES', '
        IDADEPAI', 'DTULTMENST', 'SEMAGESTAC', 'CONSPRENAT', 'MESPRENAT
        ', 'TPAPRESENT', 'STTRABPART', 'STCESPARTO', 'TPROBSON', '
        RACACOR_RN', 'RACACORN', 'ESMAE2010', 'TPNASCASSI', '
        ESMAEAGR1', 'PARIDADE', 'ANO', 'MUNICIPIOS', 'SUB-REGIAO']
3
4     def SelecaoColunasSINASC(database, year):
5         coluna = col_sinasc.copy()
6         if year == 2010: not_columns = ['RACACOR_RN', 'RACACORN', '
            ESMAE2010', 'TPNASCASSI', 'ESMAEAGR1', 'PARIDADE']

```

```

7     elif year == 2011: not_columns = ['RACACORN', 'ESMAE2010', '
      TPNASCASSI', 'ESMAEAGR1', 'PARIDADE']
8     elif year == 2012: not_columns = ['RACACOR_RN', 'TPNASCASSI', '
      ESMAEAGR1', 'PARIDADE']
9     elif year == 2013: not_columns = ['CONSPRENAT', 'TPROBSON', '
      RACACOR_RN', 'PARIDADE']
10    else: not_columns = ['RACACOR_RN', 'RACACORN']
11
12    for i in not_columns:
13        coluna.remove(i)
14        database = database[coluna]
15    return database
16
17    for ano in range(2010, 2022):
18        dados_SIM[ano] = SelecaoColunasSIM(dados_SIM[ano], year = ano)
19
20    for ano in range(2010, 2021):
21        dados_SINASC[ano] = SelecaoColunasSINASC(dados_SINASC[ano], year =
          ano)
22
23    # Calcular o número total de linhas para o SINASC e SIM
24    n_sinasc = sum(len(df.axes[0]) for df in dados_SINASC.values())
25    print(f'Número de dados do SINASC período 2010-2020: {n_sinasc}')
26
27    n_sim = sum(len(df.axes[0]) for df in dados_SIM.values())
28    print(f'Número de dados do SIM período 2010-2021: {n_sim}')
29
30    #Concatenar os dados
31    SINASC_1020 = pd.concat(dados_SINASC.values(), axis=0, ignore_
      index=True)
32    SINASC_1020.to_csv(Path('SINASC/Mess-VPP_SINASC/SINASC_VPP_1020.
      csv'))
33
34    SIM_1020 = pd.concat(dados_SIM.values(), axis=0, ignore_index=True
      )
35    SIM_1020.to_csv(Path('SIM/Mess-VPP_SIM/SIM_VPP_1020.csv'))

```

ROTULAGEM

Com relação ao banco de dados do SIM, procedeu-se à rotulação dos dados, identificando a faixa etária do óbito, classificando-os como neonatal (precoce e tardio), pós-neonatal e infantil. A rotulação foi realizada com base na coluna IDADE (mais informações no ANEXO E).

```

1    def load_SIM_SINASC_F(base):
2    path = (str(base) + "/Mess-VPP_" + str(base) + "/" + str(base) + "_VPP_1020.
      csv")

```

```

3     bd = pd.read_csv(path, low_memory=False)
4     return bd
5
6     SINASC_1020 = load_SIM_SINASC_F(base = "SINASC")
7     SIM_1020 = load_SIM_SINASC_F(base = "SIM")
8
9     def FaixaObito(database):
10        Ob_Tempo = database['IDADE'].astype(str).str[-2:] # primeiros .str
            [:2]
11        Ob_IDTempo = []
12        for ob in database['IDADE']:
13            #Columna do tempo: minutos, hora, dia, mes
14            if (ob < 100) and (ob >= 0):
15                Ob_IDTempo.append(0) #'Minutos'
16            elif (ob >= 100) and (ob < 200):
17                Ob_IDTempo.append(1) #'Hora'
18            elif (ob >= 200) and (ob < 300):
19                Ob_IDTempo.append(2) #'Dia'
20            elif (ob >= 300) and (ob < 400):
21                Ob_IDTempo.append(3) #'Mes'
22            elif (ob >= 400):
23                Ob_IDTempo.append(4) #'Ano'
24            #Coluna para tipo de óbito
25            database['Ob_IDTempo'] = Ob_IDTempo
26            database['Ob_Tempo'] = Ob_Tempo
27            database['Ob_IDTempo'] = database['Ob_IDTempo'].astype(str)
28            database['Ob_Tempo'] = database['Ob_Tempo'].astype(float)
29            Faixa_ob = []
30            Faixa_ob_name = []
31            for h, k in zip(database['Ob_IDTempo'], database['Ob_Tempo']):
32                if (h == 0) or (h == 1):
33                    Faixa_ob.append(0)
34                    Faixa_ob_name.append('Neontatal Precoce')
35                elif (h == 2) and (k <= 6):
36                    Faixa_ob.append(0)
37                    Faixa_ob_name.append('Neontatal Precoce')
38                elif (h == 2) and (k > 6) and (k <= 27):
39                    Faixa_ob.append(1)
40                    Faixa_ob_name.append('Neonatal Tardio')
41                elif (h == 2) and (k > 27):
42                    Faixa_ob.append(2)
43                    Faixa_ob_name.append('Pós-Neonatal')
44                elif (h == 3) or (h == 4):
45                    Faixa_ob.append(2)
46                    Faixa_ob_name.append('Pós-Neonatal')
47            database['Faixa_ob'] = Faixa_ob
48            database['Faixa_ob_name'] = Faixa_ob_name

```

```
49     return database
50
51     SIM_1020 = FaixaObito(SIM_1020)
52     SIM_1020.to_csv(Path('SIM/Mess-VPP_SIM/SIM_VPP_1020_FaixaObito.csv'
                          ''))
```

LIMPEZA, PADRONIZAÇÃO E CODIFICAÇÃO DE DADOS

Após a identificação de valores ausentes e incorretamente preenchidos, prosseguiu-se com a limpeza e padronização de cada conjunto de dados, conforme a descrição das variáveis nos Anexos D e E. Além disso, foram codificadas as variáveis CODOCUPMAE, CODANOMAL e CAUSABAS.

O código utilizado para a limpeza e padronização dos dados pode ser acessado através do seguinte link: https://drive.google.com/file/d/17Jns_KFICTnAvm7HcQJI-0qELXQP7Hja/view?usp=drive_link.

APÊNDICE F – IMPLEMENTAÇÃO PARA REFORMATÇÃO EM SUBCONJUNTOS DE DADOS

O código utilizado para a Etapa 3 sobre a reformatação em subconjuntos de dados pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/drive/folders/1uRlmjMTbtnVloghEZAIUlvXyWIUzqIZs?usp=drive_link.

- `Path` de `pathlib`: Utilizada para a manipulação e controle dos caminhos de arquivos e diretórios.
- `Pandas`: Utilizado para gerenciamento e processamento dos dados.
- `NumPy`: Utilizado para processamento dos dados.
- `RecordLinkage`: Utilizado para realizar a vinculação de registros entre diferentes conjuntos de dados.

ANÁLISE ESPACIAL

O código utilizado para a estruturação do subconjunto de dados para análise espacial pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/file/d/1oRalwK59seb-_wDr_hdqu2z0LzF8ILB2/view?usp=drive_link.

MODELO DE PREVISÃO

O código utilizado para a estruturação do subconjunto de dados para modelo de previsão pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/file/d/1hldHALpbE6CfqftXKN0GFeZBMUyyvh37/view?usp=drive_link.

MODELO DE PREDIÇÃO

A integração das bases de dados foi realizada com base na conexão entre as informações do SINASC e do SIM. A vinculação dos conjuntos de dados foi efetuada por meio de um conjunto de regras, utilizando a biblioteca `recordlinkage` e as variáveis-chave: `DTNASC`, `PESO`, `IDADEMAE` e `GESTACAO`. O código utilizado para a estruturação do subconjunto de dados para modelo de predição pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/file/d/1FLv7mXuwZQQRk7eOFFQR8fCBS5L5f3Nc/view?usp=drive_link.

APÊNDICE G – CÓDIGOS PARA EXPLORAÇÃO DOS SUBCONJUNTOS DE DADOS

O código utilizado para a exploração dos subconjuntos de dados pode ser acessado através do Google Colaboratory no seguinte link: https://drive.google.com/drive/folders/1HAz1Qj29jh1SvqbnIv1jz4QL7G6e7Fml?usp=drive_link.

ANÁLISE ESPACIAL

As bibliotecas utilizadas na análise exploratória espacial são listadas a seguir:

- `Pandas`: Usada para a manipulação de estruturas de dados, como `DataFrames`, facilitando a análise e o tratamento dos dados.
- `Numpy`: Usada para manipulação de arrays.
- `Matplotlib.pyplot`: Utilizada para a criação de gráficos e visualizações de dados.
- `Seaborn`: Utilizado para geração de gráficos para visualização estatística avançada.
- `Geopandas`: Empregada para manipulação e análise de dados geoespaciais. Utilizado em conjunto com a biblioteca `pandas`.
- `Matplotlib.patches` e `matplotlib.colors`: Utilizadas para adicionar formas geométricas e ajustar a paleta de cores nos gráficos de visualização.
- `Matplotlib_scalebar.scalebar`: Empregada para adicionar uma barra de escala nos mapas gerados.

O código utilizado para a análise e a geração de gráficos para visualização pode ser acessado através do seguinte link: https://drive.google.com/file/d/1gal6JaUAaCY0zysIQ2616nI80KWEJnKu/view?usp=drive_link.

MODELO DE PREVISÃO

Além das bibliotecas de `pandas`, `numpy`, `matplotlib`, `seaborn` e `patches`, seguem as bibliotecas adicionais utilizadas para a análise de séries temporais:

- `Rectangle` de `matplotlib.patches`: Utilizada para criar retângulos nos gráficos de visualização.
- `Timestamp` de `pandas`: Utilizada para manipulação de dados temporais facilitando operações como indexação e análise de séries temporais.

- `ListedColormap` de `matplotlib.colors`: Utilizada para a criação de mapas de cores personalizados.

O código utilizado para a análise e a geração de gráficos para visualização pode ser acessado através do seguinte link: https://drive.google.com/file/d/1TE0tXw4JNrsMzHkQu0GC473Fo8bZnQn1/view?usp=drive_link.

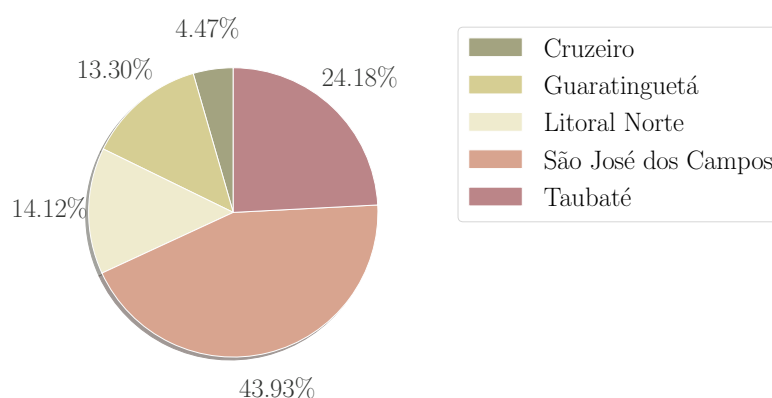
MODELO DE PREDIÇÃO

O código utilizado para a análise e a geração de gráficos para visualização pode ser acessado através do seguinte link: .

APÊNDICE H – EXPLORAÇÃO E VERIFICAÇÃO DA QUALIDADE DOS DADOS

A análise individual das informações coletadas do sistema SINASC revelou que foram registrados 363.809 nascidos vivos no período 2010 – 2020, especificamente nos 39 municípios pertencentes às 5 sub-regiões da RMVPLN. Na Figura 85 foi apresentada a proporção de dados coletados do SINASC por sub-região da RMVPLN.

Figura 85 – Proporção de dados coletados dos nascidos vivos do período 2010 – 2020

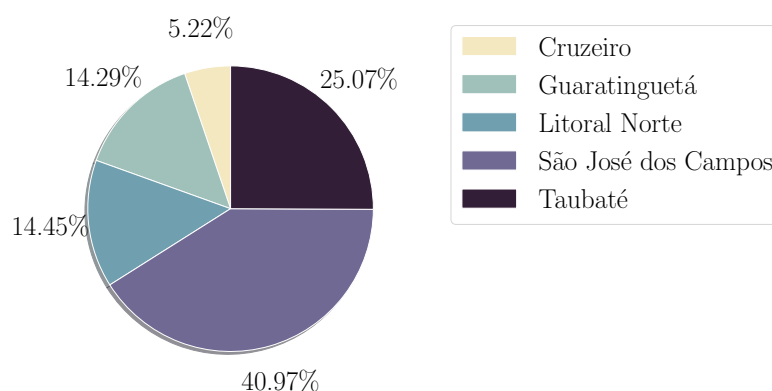


Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Aproximadamente 68,11% dos registros coletados corresponderam às sub-regiões de São José dos Campos (43,93%) e Taubaté (24,18%), com 159.817 e 87.967 registros, respectivamente. Sendo estes, as sub-regiões com maior número de registros do conjunto de dados. Sob outra perspectiva, as sub-regiões de Litoral Norte (14,12%), Guaratinguetá (13,30%) e Cruzeiro (4,47%) apresentaram o menor número de registros no banco de dados, com 51.358, 48.401 e 16.266 instâncias, respectivamente.

Por outro lado, do sistema SIM foram coletados 4.352 registros de óbitos do período 2010 – 2021. Na Figura 86 foi apresentada a proporção de dados coletados do SIM por sub-região da RMVPLN, do período 2010 – 2021.

Figura 86 – Proporção de dados coletados dos óbitos do período 2010 – 2021



Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Assim, conforme apresentado na Figura 86, e de forma semelhante ao observado no banco de dados do SINASC, no SIM as sub-regiões com maior número de dados corresponderam a São José dos Campos (40,97%) e Taubaté (25,07%), com 1.783 e 1.091 registros de óbitos, respectivamente. Pelo contrário, as sub-regiões Litoral Norte (14,45%), Guaratinguetá (14,29%) e Cruzeiro (5,22%) apresentaram o menor número de registros no banco de dados, com 629, 622 e 227 registros de óbitos, respectivamente.

Por outro lado, estudos anteriores, como o de (Correia; Padilha; Vasconcelos, 2014), destacam a presença de campos incompletos nos bancos de dados dos sistemas SINASC e SIM, comprometendo a robustez das análises. Diante disso, foi realizada uma análise da qualidade dos dados coletados da RMVPLN, visando identificar valores ausentes ou mal preenchidos em diversas características ou variáveis essenciais, como IDADEMAE, SEXO, DTNASC, GESTACAO, GRAVIDEZ, ESCMAE, CONSULTAS, APGAR1, APGAR5, entre outros.

Na análise de valores ausentes das características relacionadas aos nascidos vivos, identificou-se que 36 de 43 variáveis apresentaram dados faltantes, conforme apresentado na Tabela 23 e Figura 87.

Tabela 23 – Quantidade e proporção de valores ausentes por variável dos nascidos vivos

Variáveis	MV^a	MV (%)	Variáveis	MV	MV (%)
CODOCUPMAE	287.857	79,12	QTDFILMORT	78.741	21,64
IDADEPAI	223.574	61,45	TPAPRESENT	78.314	21,53
DTULTMENST	181.026	49,76	DTNASCMAE	78.291	21,52
PARIDADE	132.504	36,42	QTDFILVIVO	43.691	12,01
QTDPARTNOR	129.876	35,70	ESCMAE	7.583	2,08
QTDPARTCES	124.530	34,23	RACACOR	6.861	1,89
CONSPRENAT	110.885	30,48	APGAR1	3.750	1,03
TPROBSON	106.603	29,30	GESTACAO	3.750	1,03
QTDGESTANT	105.241	28,93	APGAR5	3.593	0,99
ESCMAEAGR1	104.139	28,62	IDANOMAL	2.333	0,64
TPNASCASSI	101.301	27,84	ESTCIVMAE	1.577	0,43
STCESPARTO	93.293	25,64	CODESTAB	1.192	0,33
MESPRENAT	85.113	23,39	HORANASC	429	0,12
ESCMAE2010	83.882	23,06	CONSULTAS	385	0,11
RACACORMAE	82.222	22,60	PARTO	359	0,10
CODMUNNATU	81.274	22,34	GRAVIDEZ	338	0,09
STTRABPART	80.826	22,22	SEXO	49	0,01
SEMAGESTAC	79.955	21,98	PESO	18	0,005

(a) MV: Valores ausentes (*missing values*).

Fonte: Elaborado pela autora (2025).

Como pode ser observado na Figura 87 e Tabela 23, 51,16% (22) das variáveis apresentaram mais de 20% de dados faltantes. As 5 variáveis com maior quantidade de valores ausentes foram: CODOCUPMAE (79,12%), IDADEPAI (61,45%), DTULTMENST (49,76%), PARIDADE (36,42%) e QTDPARTNOR (35,70%), referente à ocupação materna, idade paterna, data da última menstruação, número total de filhos anteriores (vivos e mortos) e número de partos vaginais

Tabela 24 – Evolução da disponibilidade de variáveis nos registros do SINASC por ano (2010 – 2020)

Variável	(conclusão)										
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
QTDFILVIVO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DTNASCMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPAPRESENT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QTDFILMORT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEMAGESTAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STTRABPART	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CODMUNNATU	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RACACORMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ESMAE2010			✓	✓	✓	✓	✓	✓	✓	✓	✓
MESPENAT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STCESPARTO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPNASCASSI				✓	✓	✓	✓	✓	✓	✓	✓
ESMAEAGR1				✓	✓	✓	✓	✓	✓	✓	✓
QTDGESTANT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPROBSON	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
CONSPENAT	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
QTDPARTCES	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QTDPARTNOR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PARIDADE					✓	✓	✓	✓	✓	✓	✓
CODOCUPMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DTULTMENST	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IDADEPAI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Fonte: Elaborado pela autora (2025).

Da análise da Tabela 24 foi identificado que algumas variáveis não estiveram presentes em todos os anos avaliados. Especificamente, as variáveis de TPNASCASSI, ESCMAEAGR1, TPROBSON, CONSPENAT e PARIDADE apresentaram uma maior proporção de valores ausentes de acordo com a Tabela 23 e a Figura 87.

Por sua vez, embora os dados dos óbitos (SIM) sejam utilizados para rotular o conjunto de dados dos nascidos vivos (SINASC) com o objetivo de identificar os recém-nascidos que vieram a óbito, também foi analisada a ocorrência de valores faltantes nos dados do SIM.

Como resultado, identificaram-se 28 de 35 variáveis com valores ausentes. A Tabela 25 e Figura 88 apresentam os resultados da análise de ausência de dados das variáveis do conjunto de dados de óbitos.

Tabela 25 – Quantidade e proporção de valores ausentes por variável dos óbitos

(continua)					
Variáveis	MV ^a	MV (%)	Variáveis	MV	MV (%)
CIRCOBITO	4.243	97,50	LINHAD	3.288	75,55
TPMORTEOCO	4.127	94,83	LINHAI I	2.529	58,11
TPOBITOCOR	4.049	93,04	OCUPMAE	2.375	54,57
CIRURGIA	3.982	91,50	ESMAEAGR1	2.320	53,31

(a) MV: Valores ausentes (*missing values*).

Fonte: Elaborado pela autora (2025).

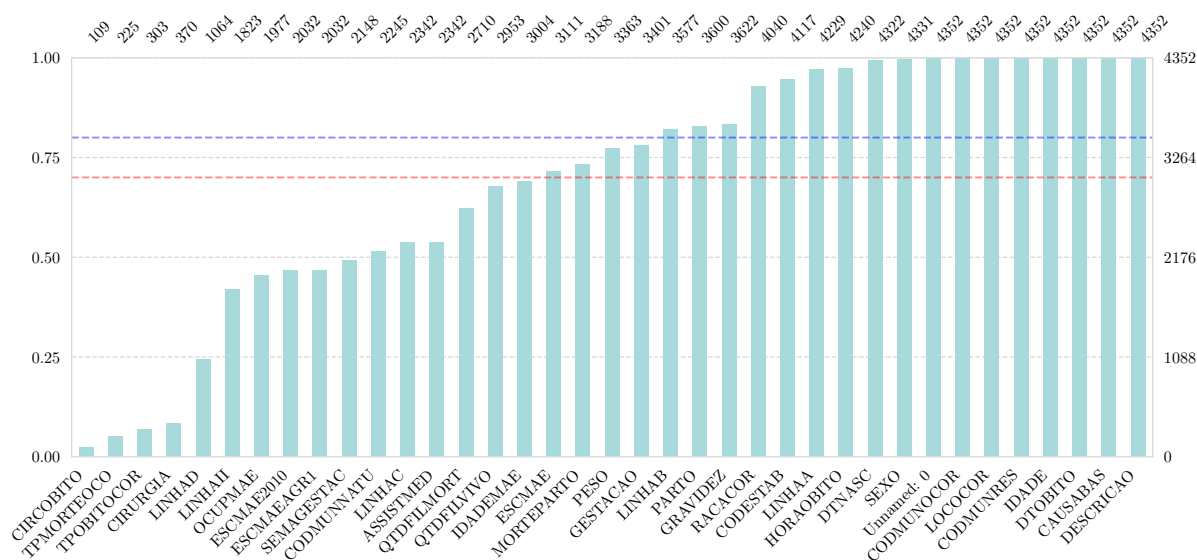
Tabela 25 – Quantidade e proporção de valores ausentes por variável dos óbitos
(conclusão)

Variáveis	MV ^a	MV (%)	Variáveis	MV	MV (%)
ESMAE2010	2.320	53,31	PESO	989	22,73
SEMAGESTAC	2.204	50,64	GESTACAO	951	21,85
CODMUNNATU	2.107	48,41	LINHAB	775	17,81
LINHAC	2.010	46,19	PARTO	752	17,28
ASSISTMED	2.010	46,18	GRAVIDEZ	730	16,77
QTDFILMORT	1.642	37,73	RACACOR	312	7,17
QTDFILVIVO	1.399	32,15	CODESTAB	235	5,40
IDADEMAE	1.348	30,97	LINHAA	123	2,83
MORTEPARTO	1.164	26,75	HORAOBITO	112	2,57
ESMAE	1.103	25,34	DTNASC	30	0,69

(a) MV: Valores ausentes (*missing values*).

Fonte: Elaborado pela autora (2025).

Figura 88 – Identificação de dados ausentes do banco de dados SIM



Fonte: Elaborado pela autora (2025).

Da Figura 88 e Tabela 25, observou-se que 57,14% das variáveis apresentaram mais de 20% de valores ausentes, sendo as variáveis de CIRCUBITO (97,50%), TPMORTEOCO (94,83%), TPOBITOCOR (93,04%), CIRURGIA (91,50%) e LINHAD (75,55%) as 5 variáveis com maior ausência de dados. Mais informação das variáveis em Quadro 6.

Do mesmo modo, avaliou-se a evolução anual da disponibilidade de variáveis dos registros do SIM, conforme apresentado na Tabela 26.

Tabela 26 – Evolução da disponibilidade de variáveis nos registros do SIM por ano (2010-2020)

Variável	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
DTOBITO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IDADE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEXO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CAUSABAS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CODMUNRES	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LOCOCOR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CODMUNOCOR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DTNASC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HORAOBITO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LINHAA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CODESTAB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RACACOR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GRAVIDEZ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PARTO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LINHAB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GESTACAO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PESO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ESCMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MORTEPARTO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IDADEMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QTDFILVIVO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QTDFILMORT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ASSISTMED	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LINHAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CODMUNNATU					✓	✓	✓	✓	✓	✓	✓
SEMAGESTAC			✓	✓	✓	✓	✓	✓	✓	✓	✓
ESCMAE2010		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ESCMAEAGR1			✓		✓	✓	✓	✓	✓	✓	✓
OCUPMAE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LINHAI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LINHAD	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CIRURGIA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPOBITOCOR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPMORTEOCO			✓	✓	✓	✓	✓	✓	✓	✓	✓
CIRCOBITO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

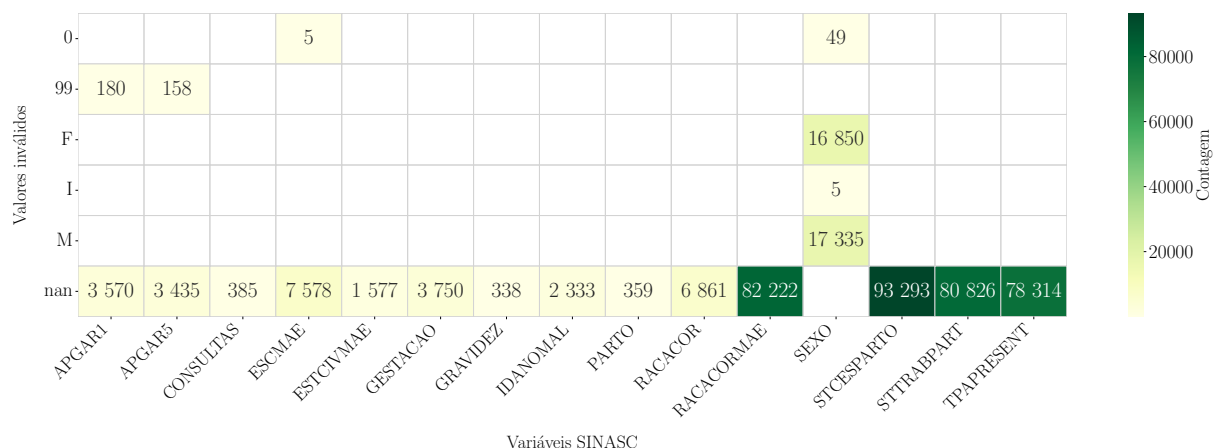
Fonte: Elaborado pela autora (2025).

Nos registros do SIM observou-se que as variáveis de CODMUNNATU, SEMAGESTAC, ESCMAE2010, ESCMAEAGR1 e TPMORTEOCO não estiveram presentes em todos os anos do período de 2010 a 2020, sendo uma das possíveis causas da alta proporção de valores ausentes nas variáveis identificadas.

Por outro lado, a análise dos dados também permitiu identificar características com valores mal preenchidos, o que pode afetar significativamente a qualidade e a confiabilidade de análises.

A Figura 89 apresenta um mapa de calor ilustrando a quantidade de valores mal preenchidos ou inválidos nas variáveis do SINASC, sendo que as cores mais claras representam uma menor ocorrência desses erros.

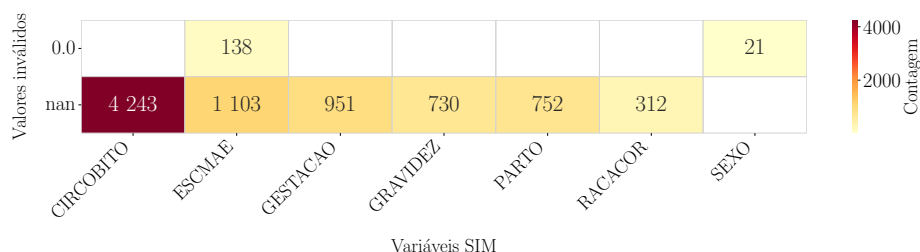
Figura 89 – Mapa de calor de valores inválidos das variáveis SINASC



Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Adicionalmente, na Figura 90 é exibido um mapa de calor correspondente às variáveis mal preenchidas das variáveis do SIM. Além disso, em ambos os conjuntos de dados, foram encontrados os valores 9 ou 99, inseridos pelo especialista para indicar que a informação correspondente foi ignorada.

Figura 90 – Mapa de calor de valores inválidos das variáveis SIM



Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

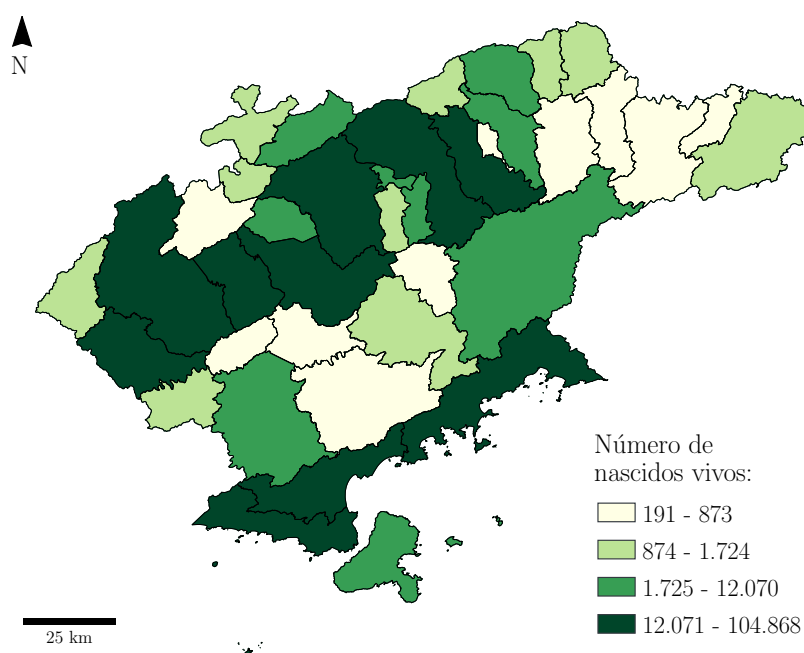
O preenchimento inadequado ou ausência de valores no conjunto de dados do SINASC e SIM pode ser devido a diversas causas, como erros de digitação, falta de capacitação dos profissionais de saúde ou responsáveis, desconhecimento da informação para responder e formulários confusos ou interfaces pouco amigáveis, que podem levar a erros na coleta de dados, entre outros (Correia; Padilha; Vasconcelos, 2014; Beluzo *et al.*, 2020).

Assim, tanto a Figura 89 quanto a Figura 90 oferecem uma visão abrangente da qualidade dos dados coletados, facilitando a identificação das características mais afetadas. A identificação dessas características permite orientar as análises corretivas, com o objetivo de reduzir erros significativos quando os dados são utilizados em análises espaciais ou no desenvolvimento de modelos com técnicas de ML, evitando assim comprometer a precisão das inferências.

APÊNDICE I – EXPLORAÇÃO DOS NASCIDOS VIVOS E ÓBITOS NA REGIÃO

De acordo com as informações coletadas da RMVPLN no SINASC, no período 2010-2020 ocorreram 363.809 nascimentos. A Figura 91 ilustra a distribuição espacial do número total de nascidos vivos do período analisado, apresentado por município da RMVPLN.

Figura 91 – Distribuição do número de nascidos vivos do período 2010 – 2020



Fonte: Elaborado pela autora (2025).

Conforme está na Figura 91, 25,64% (10) dos municípios apresentaram mais de 12.070 nascimentos, porém o município de São José dos Campos (0,29%) apresentou a maior quantidade, com 104.868 nascidos vivos. Além de São José dos Campos, os próximos três municípios com maior número de nascidos vivos foram Taubaté (0,12%), Jacareí (0,09%) e Pindamonhangaba (0,07%) com 43.306, 34.350 e 25.113 nascidos vivos.

Sob outra perspectiva, 25,64% (10) dos municípios apresentaram menos de 874 nascimentos no período analisado. O município de Arapeí (< 0,01%) apresentou a menor quantidade de nascimentos, com 191 nascidos vivos, seguido pelos municípios de Monteiro Lobato (< 0,01%), Lagoinha (< 0,01%) e Areias (< 0,01%) com 421, 423 e 481 nascidos vivos, respectivamente. Já 48,72% (19) dos demais municípios tiveram entre 874 e 12.070 nascidos vivos.

Na Figura 92 apresenta-se a distribuição do número de nascidos vivos anualmente por município, mostrando-se que o número de nascidos vivos entre os municípios apresentaram uma oscilação próxima ao longo do tempo.

Figura 92 – Distribuição anual do número de nascidos vivos



Fonte: Elaborado pela autora (2025).

Ademais, a Figura 92 mostrou uma leve queda no número de nascidos vivos nos últimos dois anos (2019 e 2020) entre os municípios da RMVPLN.

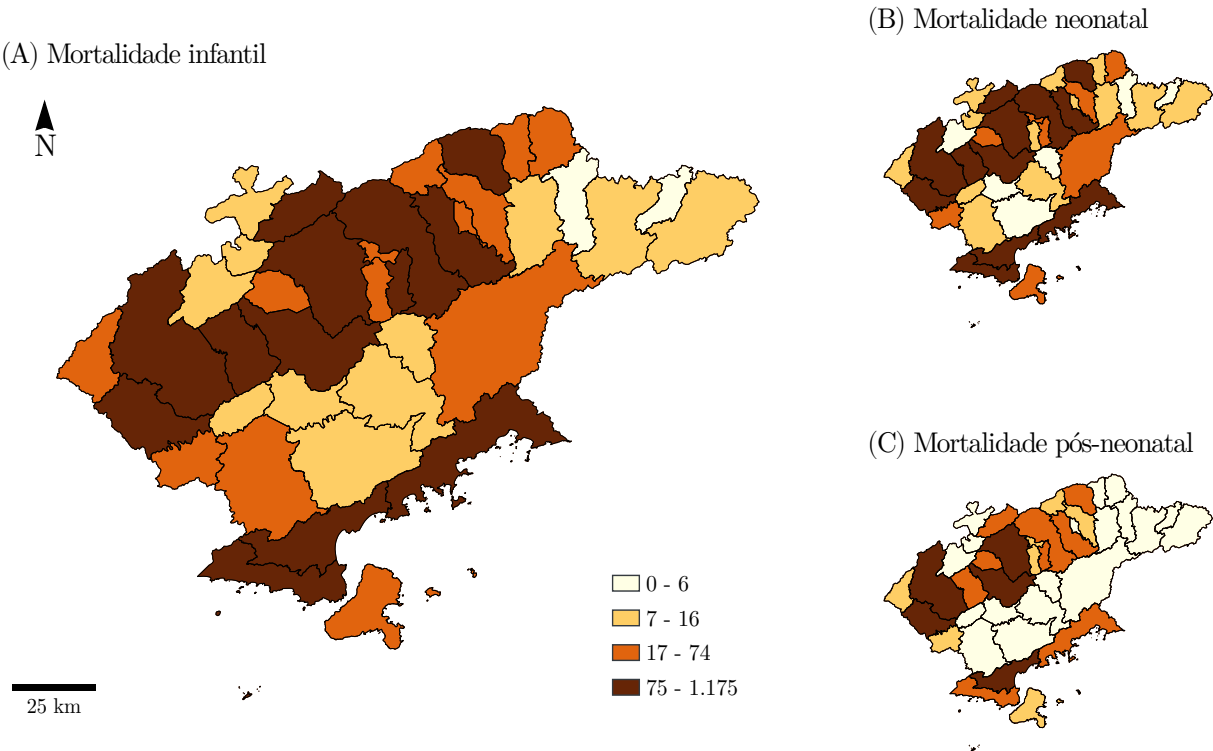
Por outro lado, as informações coletadas do SIM mostraram que no período 2010–2021 ocorreram 4.352 óbitos infantis (com menos de uma ano de idade), sendo que, destes óbitos 71,99%(3.133) ocorreram no período neonatal e 28,01% (1.219) no período pós-neonatal.

A distribuição espacial do número de óbitos infantis, neonatais e pós-neonatais para cada município da RMVPLN foi apresentada na Figura 93.

Na mortalidade infantil, Figura 93 (A), 33,33% (13) dos municípios apresentaram mais de 75 mortes entre 2010 e 2021. Assim, de forma semelhante ao observado na análise dos nascidos vivos, o município de São José dos Campos (0,27%) apresentou a maior quantidade de mortes de nascidos vivos, com 1.175 óbitos infantis. Seguidamente, os municípios de Taubaté (0,12%), Jacareí (0,08%) e Pindamonhangaba (0,07%) também apresentaram o maior número de óbitos infantis com 527, 372 e 317 mortes, respectivamente.

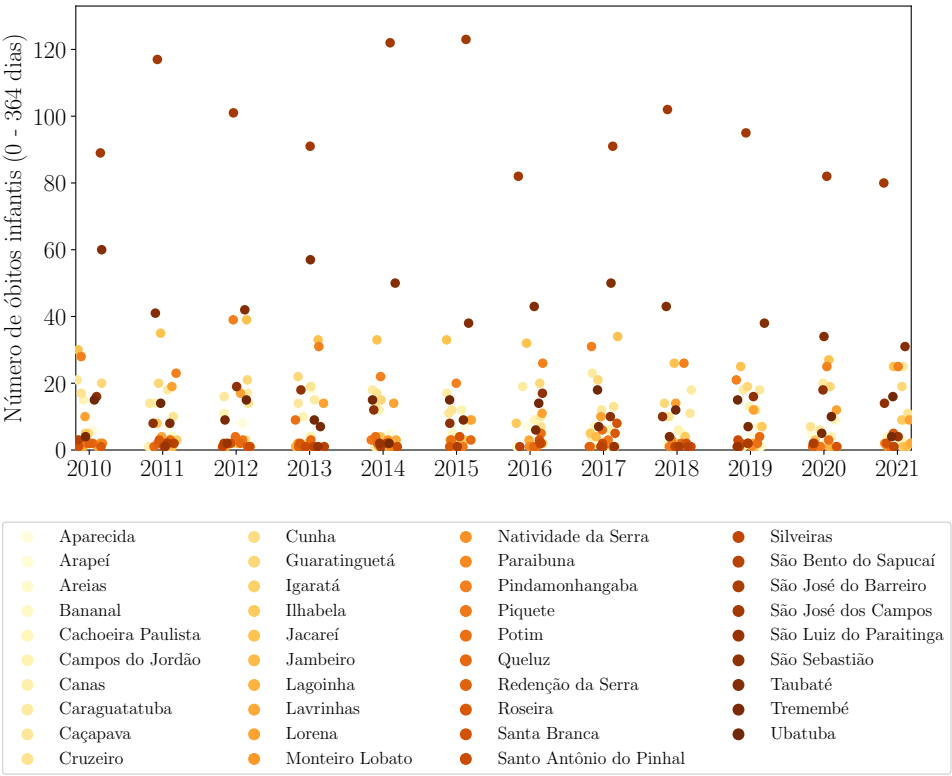
Por outro lado, 5,13% (2) dos municípios apresentaram menos de 6 óbitos, sendo estes os municípios de Arapeí (<0,01%) e Areias (<0,01%) com 3 óbitos cada um. Os 61,54% (24) dos municípios restantes apresentaram entre 7 e 74 óbitos (Figura 93).

Figura 93 – Distribuição do número de óbitos infantis, neonatais e pós-neonatais do período 2010 – 2021



Fonte: Elaborado pela autora (2025).

Figura 94 – Distribuição anual do número de óbitos infantis



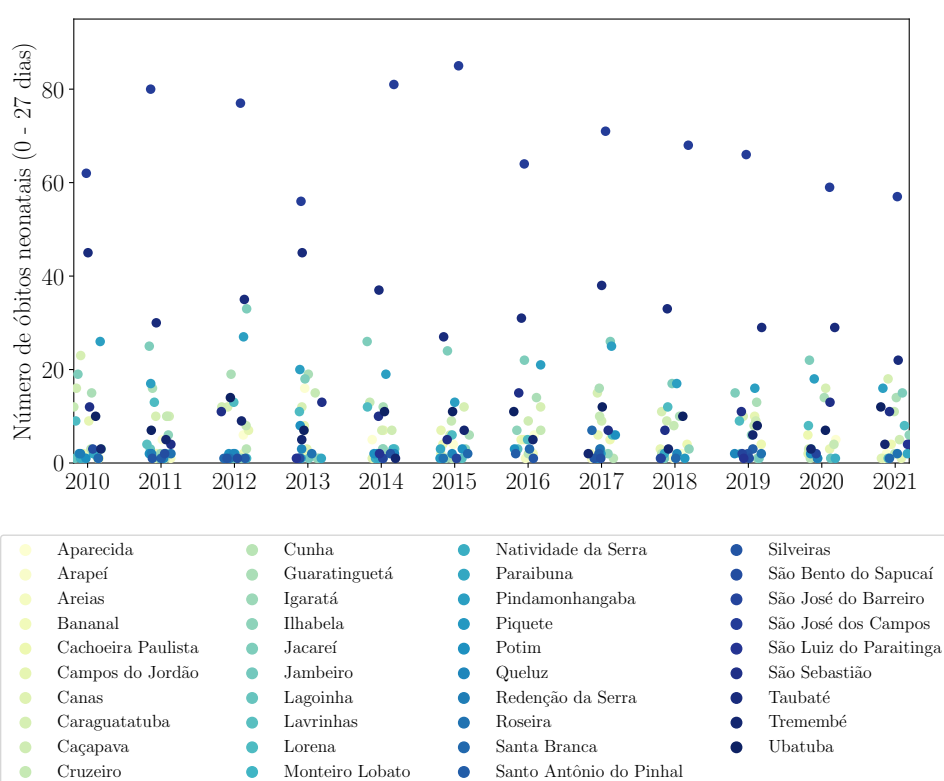
Fonte: Elaborado pela autora (2025).

Na Figura 94 apresenta-se a distribuição anual do número de óbitos infantis entre os municípios da RMVPLN, observando-se que, diferentemente dos nascidos vivos, os óbitos infantis não apresentaram uma variação constante.

A análise da mortalidade na faixa etária neonatal mostrou que, conforme observado na Figura 93 (B), 30,77% (12) dos municípios apresentaram mais de 75 óbitos. No intervalo de 7 e 74 óbitos estiveram 53,85% (21) dos municípios, e 15,38% (6) dos municípios apresentaram menos de 6 óbitos neonatais.

Além disso, conforme observado na Figura 95, a mortalidade neonatal também apresentou uma oscilação instável no número anual de óbitos no período analisado.

Figura 95 – Distribuição anual do número de óbitos neonatais

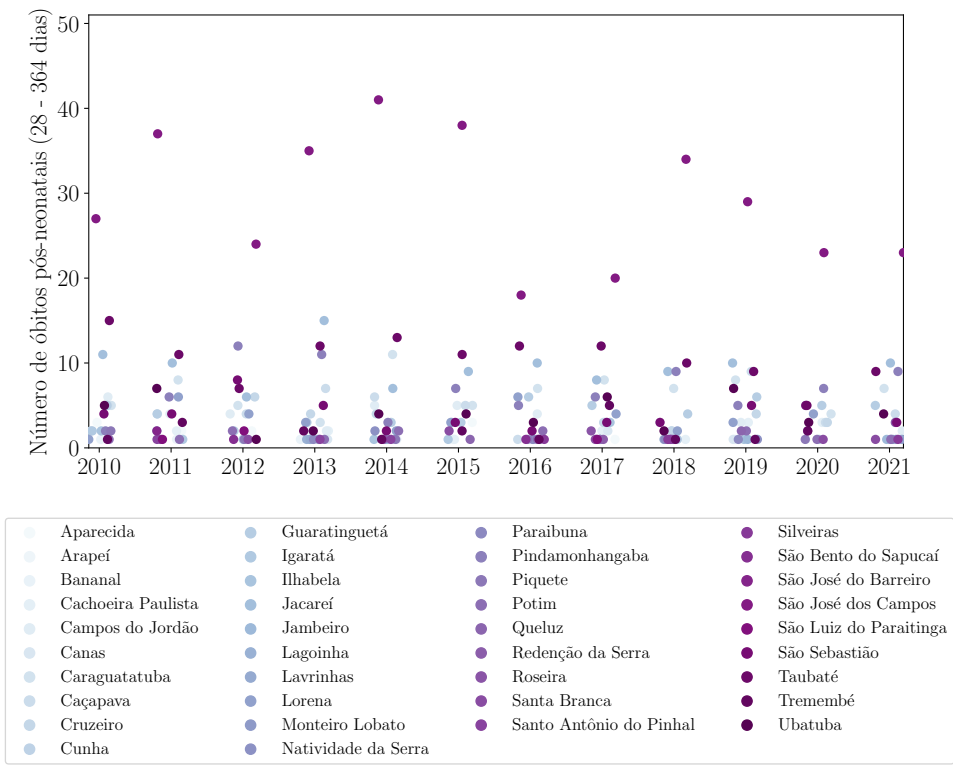


Fonte: Elaborado pela autora (2025).

Por outro lado, na faixa etária pós-neonatal, conforme observado na Figura 93 (C), 12,82% (5) dos municípios apresentaram mais de 75 óbitos, e entre o intervalos 7 e 74 estiveram 41,03% (16) dos municípios. Já o 46,15% (18) dos municípios restantes apresentaram abaixo de 6 óbitos pós-neonatais.

Do mesmo modo que na mortalidade infantil e neonatal, a distribuição anual do número óbitos pós-neonatais entre os municípios da RMVPLN apresentou uma variação instável, conforme pode ser observado na Figura 96.

Figura 96 – Distribuição anual do número de óbitos pós-neonatais



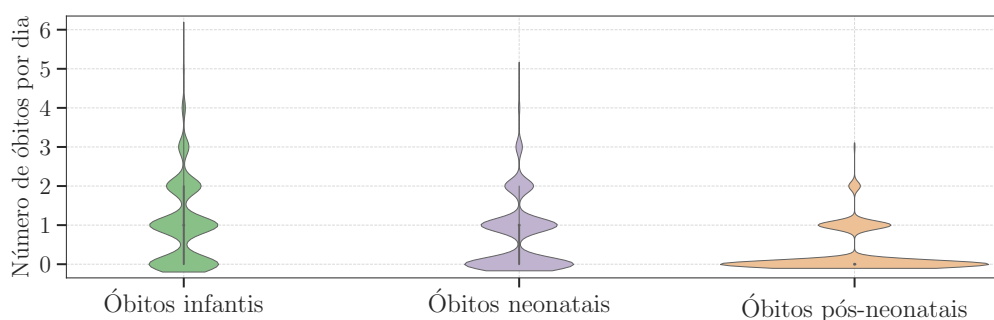
Fonte: Elaborado pela autora (2025).

APÊNDICE J – PADRÃO TEMPORAL DA MORTALIDADE DO NASCIDO VIVO

Sob outra perspectiva, como os dados coletados pertencem à declaração de óbito do recém-nascido, as informações relativas à data do óbito permitiram a estruturação temporal do número de óbitos. Os dados foram estruturados temporalmente em número de óbitos por dia, semana, quinzena, e mês, para a RMVPLN e para as sub-regiões de São José dos Campos, Taubaté, Guaratinguetá, Cruzeiro e Litoral Norte.

Em relação à distribuição do número de óbitos na RMVPLN, a Figura 97 apresenta o gráfico de violino do número de óbitos diários, avaliados pelas faixas etárias: infantil, neonatal e pós-neonatal.

Figura 97 – Distribuição do número de óbitos diários por faixa etária, período 2010 – 2021



Fonte: Elaborado pela autora (2025).

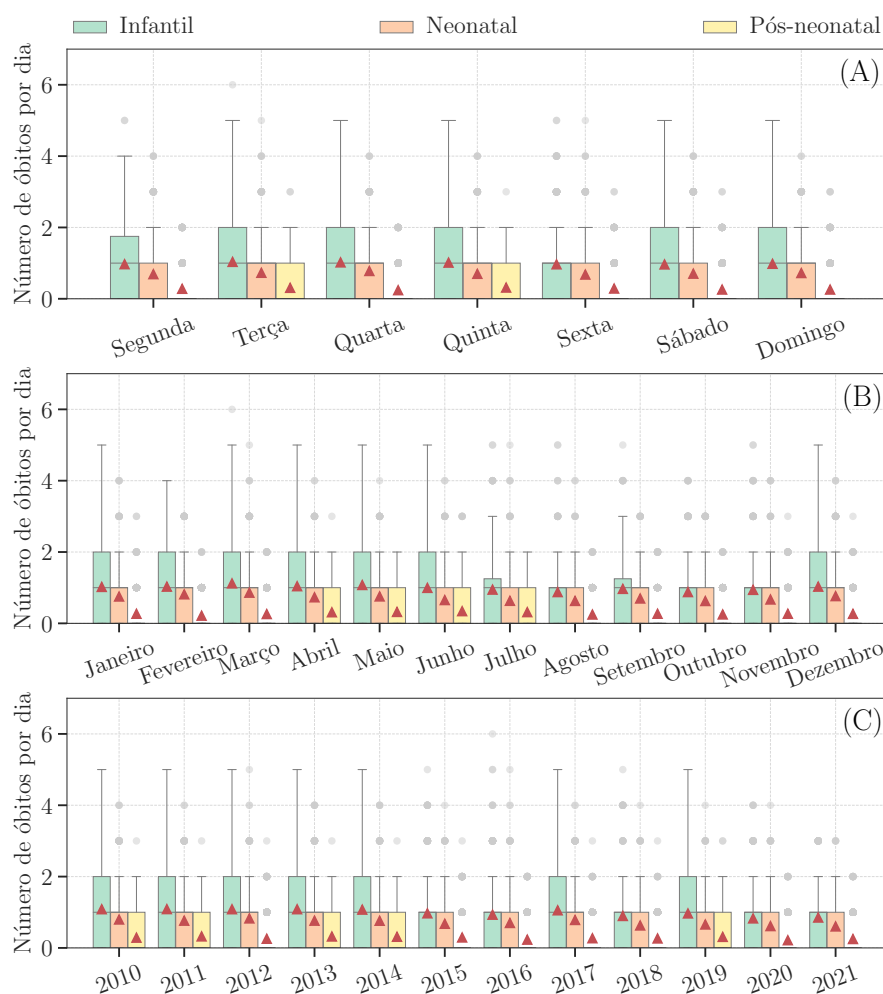
A Figura 97 mostra que o número de óbitos em cada faixa etária apresentou mais de um grupo de frequência, identificando-se assim uma distribuição de frequência multimodal.

Adicionalmente, a fim de observar o comportamento dos óbitos diários, a Figura 98 apresenta o gráfico de caixa do número de óbitos diários agrupados por dia da semana (A), mês (B) e ano (D).

Da Figura 98 foram observados valores *outliers* (discrepantes) em cada um dos grupos analisados, com sobreposição principalmente entre o limite inferior e o primeiro quartil dos gráficos de caixa. Em específico, na análise dos dados diários de óbitos infantis agrupados por dia da semana foi identificado um padrão temporal semelhante entre cada dia da semana, Figura 98 (A), no entanto, foi observada uma aparente diminuição nas sextas-feiras.

De forma semelhante, os óbitos neonatais apresentaram um padrão temporal diário equivalente durante a avaliação semanal, diferentemente das mortes pós-neonatais, que apresentaram um aumento às terças-feiras e quintas-feiras, em relação ao número de mortes diárias.

Figura 98 – Distribuição do número de óbitos diários agrupados por dia da semana, mês e ano



Fonte: Elaborado pela autora (2025).

Em relação ao número de óbitos diários agrupados por mês, Figura 98 (B), na mortalidade infantil foi observada uma variação semelhante entre os meses de dezembro e janeiro, com uma leve caída em fevereiro, seguida de uma variação mais constante entre os meses de março, abril, maio e junho. Entre os meses de julho e novembro a mortalidade infantil apresentou uma menor variação.

Por outro lado, na mortalidade neonatal, além de apresentar valores *outliers*, observou-se uma variação mais constante entre o número de óbitos diários quando avaliada por mês. Na mortalidade pós-neonatal foi observada menor variação mensal em relação ao número de óbitos diários, com aumento entre os meses de abril e julho.

A análise do número diário de óbitos agrupados por ano, da Figura 98 (C), mostrou que nas três faixas etárias foram observados *outliers*. Na mortalidade infantil foi observada uma variação aparentemente constante entre os anos de 2010 e 2014, e os anos seguintes apresentaram variação mais instável com queda entre 2015, 2016, 2018, 2020 e 2021, além de mostrar aumento em 2017 e 2019.

Da mesma forma, não foram observadas variações consideráveis na mortalidade neonatal,

ao contrário do número diário de mortes pós-neonatais, dado que apresentaram comportamento semelhante entre 2010 – 2014 e 2019, e uma diminuição entre os anos 2015 – 2018, 2020 e 2021.

A Tabela 27 apresenta as estatísticas descritivas do número de óbitos diários, mostrando que o número de mortes infantis diárias variou entre 0 e 6. O número de mortes diárias ocorridas durante o período neonatal e pós-neonatal variou entre 0 – 5 e 0 – 3 óbitos, respectivamente.

Tabela 27 – Estatísticas descritivas do número de óbitos diários

Faixa Etária	Contagem	Média	Var^a	Desv. pad.^b	Mín^c	Q^d1	Q2	Q3	Máx^e
Infantil	4383	0,99	1,00	1	0	0	1	2	6
Neonatal	4383	0,71	0,71	0,84	0	0	1	1	5
Pós-neonatal	4383	0,28	0,28	0,53	0	0	0	0	3

(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.
Fonte: Elaborado pela autora (2025).

Por outro lado, a Figura 99 ilustra a distribuição temporal e o gráfico de caixa, por mês e ano, do número de óbitos semanais na RMVPLN. Adicionalmente, na Tabela 28 apresenta-se as estatísticas descritivas dos óbitos semanais por faixa etária.

Tabela 28 – Estatísticas descritivas do número de óbitos semanais

Faixa Etária	Contagem	Média	Var^a	Desv. pad.^b	Mín^c	Q^d1	Q2	Q3	Máx^e
Infantil	626	6,95	7,52	2,74	0	5	7	9	19
Neonatal	626	5,00	5,32	2,31	0	3	5	6	15
Pós-neonatal	626	1,94	1,99	1,41	0	1	2	3	9

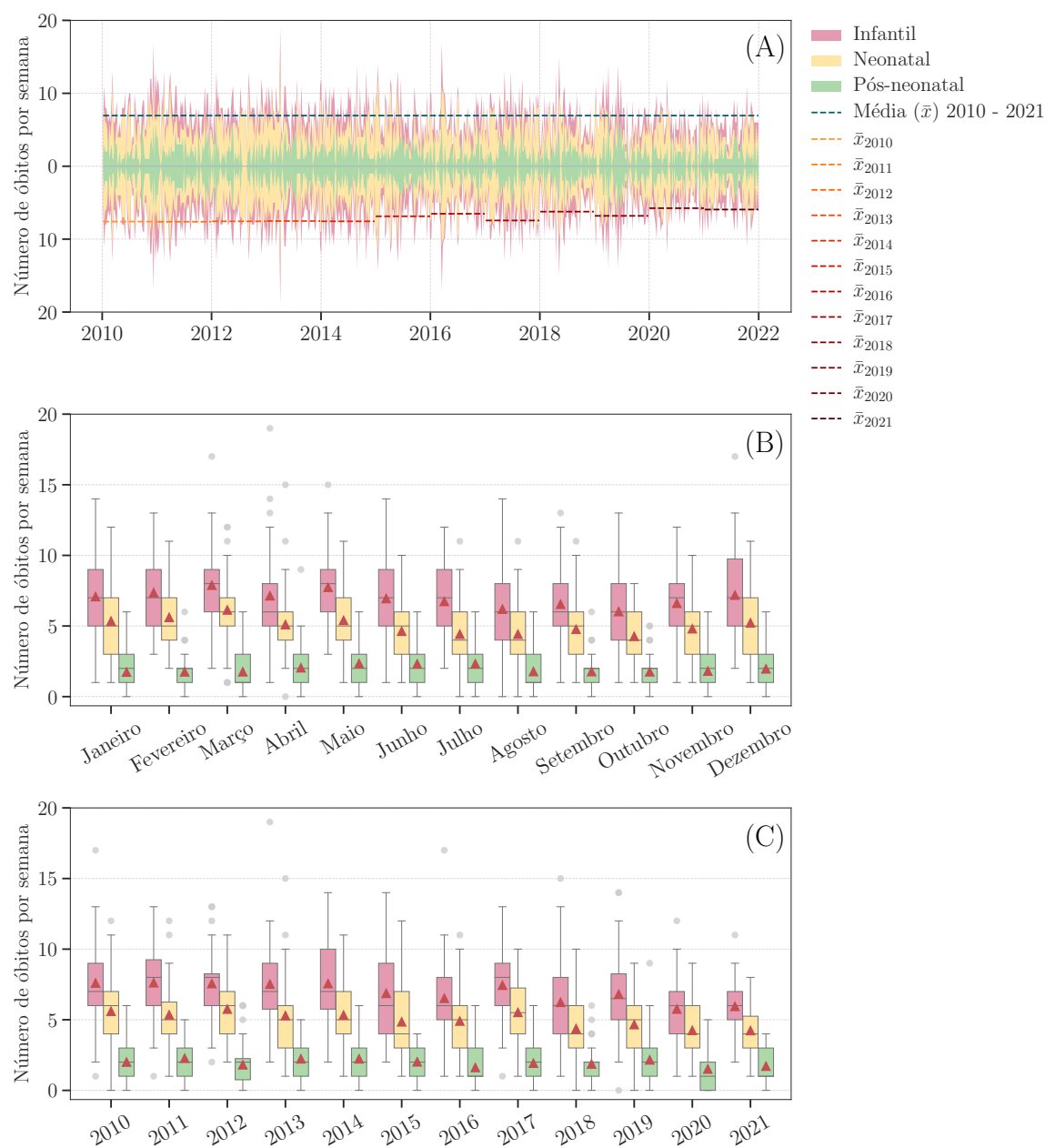
(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.
Fonte: Elaborado pela autora (2025).

Conforme a Tabela 28 e Figura 99 (A), a mortalidade infantil apresentou uma variação entre 0 e 19 óbitos semanais, com valor médio de mortes semanais igual a 7 (± 3). Como pode-se verificar na Figura 99, o número de óbitos infantis semanais apresentou uma variação semelhante entre os diferentes meses (B), além de observar-se uma diminuição na variação de óbitos a longo dos anos (C).

A mortalidade neonatal, Tabela 28 e Figura 99 (A), oscilou entre 0 e 15 óbitos semanais, com aproximadamente 5 (± 2) mortes por semana como valor médio. A variação dos óbitos neonatais semanais apresentaram uma diminuição principalmente entre março, abril, junho, julho e outubro, conforme a Figura 99 (B). Já anualmente (C), os óbitos neonatais apresentaram uma diminuição na variação de óbitos ao longo dos anos.

Na mortalidade pós-neonatal, Tabela 28 e Figura 99 (A), o número de óbitos variou entre 0 e 9, com perto de 2 (± 1) óbitos médios por semana. Na avaliação de mortes semanais por mês (Figura 99), foi observada uma redução entre os meses de fevereiro, setembro e outubro (B), e na avaliação anual, foi observada uma oscilação instável ao longo do tempo (C).

Figura 99 – Distribuição do número de óbitos semanal, e agrupado por mês e ano



Fonte: Elaborado pela autora (2025).

Em relação à avaliação de óbitos quinzenais, na Figura 100 foi apresentada a distribuição temporal, seguida da Tabela 29 que apresenta as estatísticas descritivas por faixa etária.

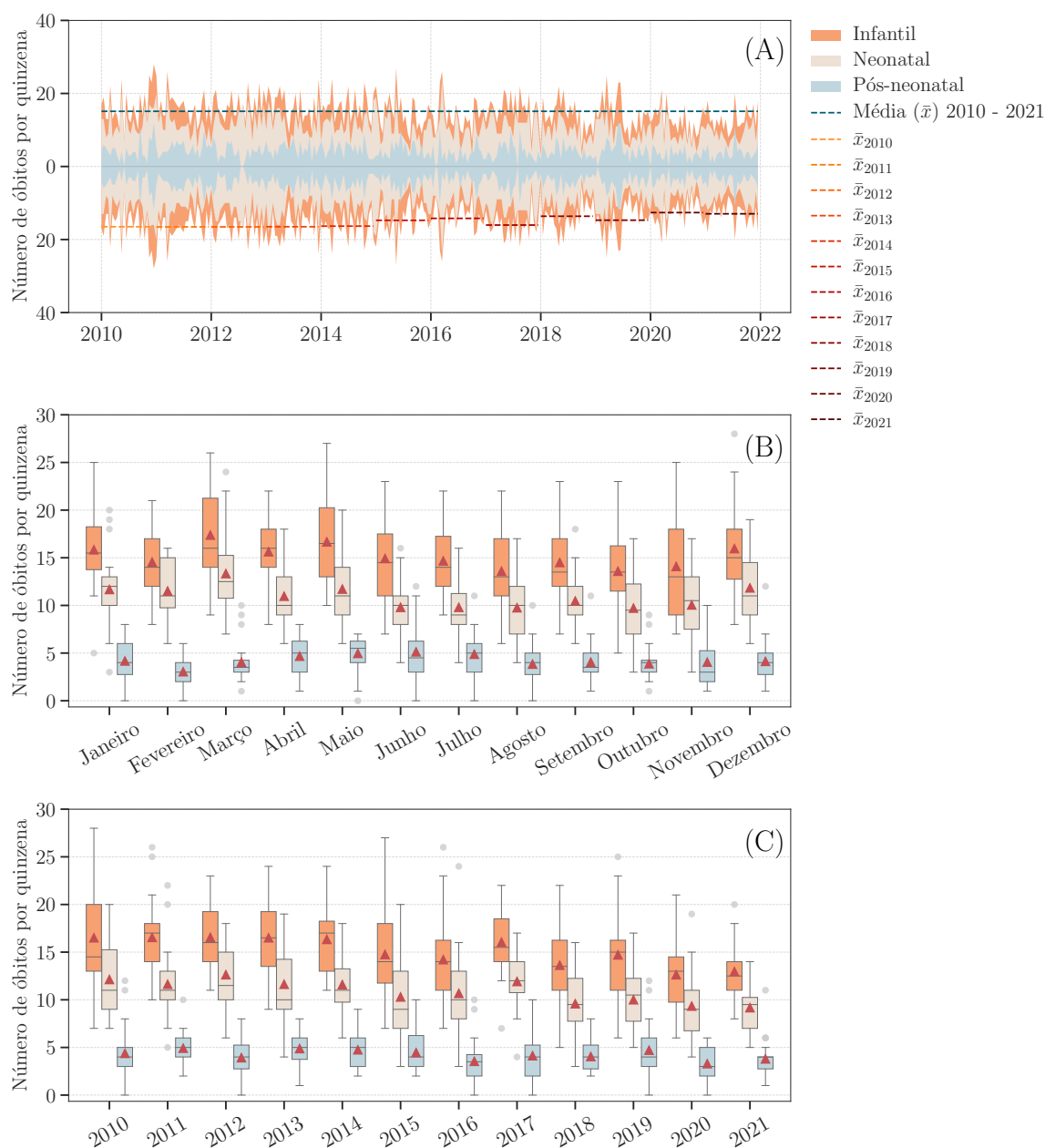
Tabela 29 – Estatísticas descritivas do número de óbitos quinzenais

Faixa Etária	Contagem	Média	Var ^a	Desv. pad. ^b	Mín ^c	Q ^d 1	Q2	Q3	Máx ^e
Infantil	288	15,11	18,96	4,35	5	12	15	18	28
Neonatal	288	10,88	13,05	3,61	3	8	11	13	24
Pós-neonatal	288	4,23	5,11	2,26	0	3	4	6	12

(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.

Fonte: Elaborado pela autora (2025).

Figura 100 – Distribuição do número de óbitos quinzenal, e agrupado por mês e ano



Fonte: Elaborado pela autora (2025).

Na avaliação quinzenal da mortalidade infantil na Figura 100 (A) e Tabela 29, observou-se uma variação quinzenal entre 5 e 28 óbitos, com aproximadamente 15 (± 4) óbitos como valor médio por quinzena. Ao avaliar os óbitos quinzenais agrupados por mês e ano, conforme apresentado na Figura 100, o número de óbitos mensais variou principalmente entre os meses de março, maio e novembro (B). Na análise anual, observou-se maior variação nos anos de 2010 e 2015, além de diminuição no número de óbitos quinzenais a partir de 2019 (C).

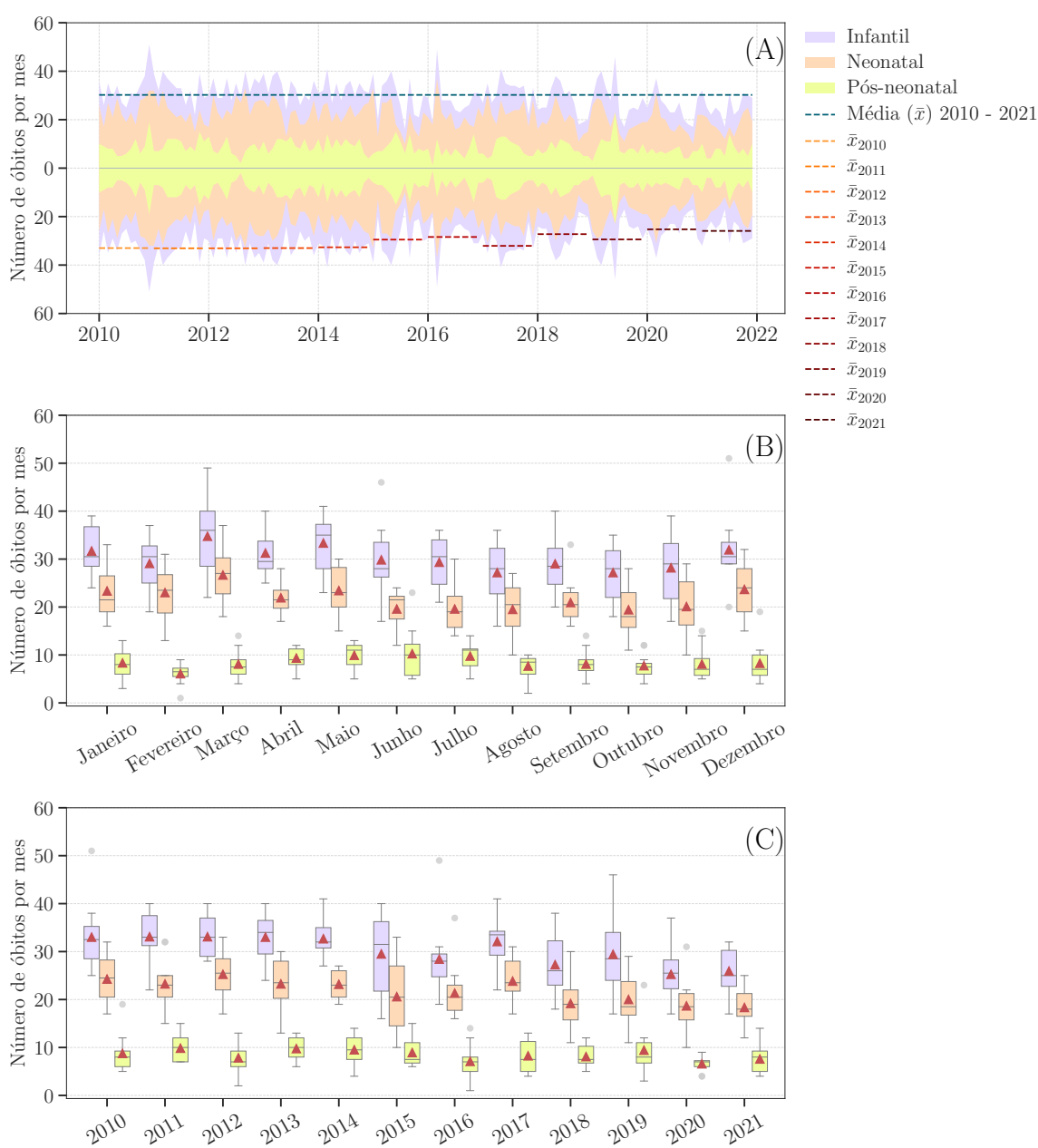
Na mortalidade neonatal da Figura 100 e Tabela 29, o número de óbitos quinzenais variou entre 3 e 24, com valor médio de 11 (± 4) óbitos por quinzena na RMVPLN. Na análise dos óbitos neonatais quinzenais avaliados por mês e ano, observou-se o mesmo comportamento

temporal da análise dos óbitos infantis.

Em relação à mortalidade pós-neonatal, Figura 100 (A) e Tabela 29, o número de óbitos quinzenais oscilou entre 0 – 12, tendo como valor médio 4 (± 2) óbitos quinzenais. Na avaliação mensal, na Figura 100 (B), observou-se comportamento diferente daquele contemplado na faixa etária infantil e neonatal, observando-se maior variação no número de óbitos, principalmente no mês de junho. Anualmente (C), observou-se diminuição da variação dos óbitos quinzenais ao longo dos anos, da mesma forma que foi observado nos óbitos infantis e neonatais.

Do mesmo modo, a Figura 101 apresenta a distribuição temporal de óbitos mensais na RMVPLN.

Figura 101 – Distribuição do número de óbitos mensal, e agrupado por mês e ano



Fonte: Elaborado pela autora (2025).

Adicionalmente, os óbitos infantis, neonatais e pós-neonatais foram analisados por ocorrência total por mês, conforme Tabela 32 que apresenta estatísticas descritivas dos óbitos mensais por faixa etária.

Tabela 30 – Estatísticas descritivas do número de óbitos mensais

Faixa Etária	Contagem	Média	Var^a	Desv. pad.^b	Mín^c	Q^d1	Q2	Q3	Máx^e
Infantil	144	30,22	42,86	6,55	16	26	30	35	51
Neonatal	144	21,76	29,30	5,41	10	18	22	26	37
Pós-neonatal	144	8,47	10,47	3,24	1	6	8	11	23

(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.

Fonte: Elaborado pela autora (2025).

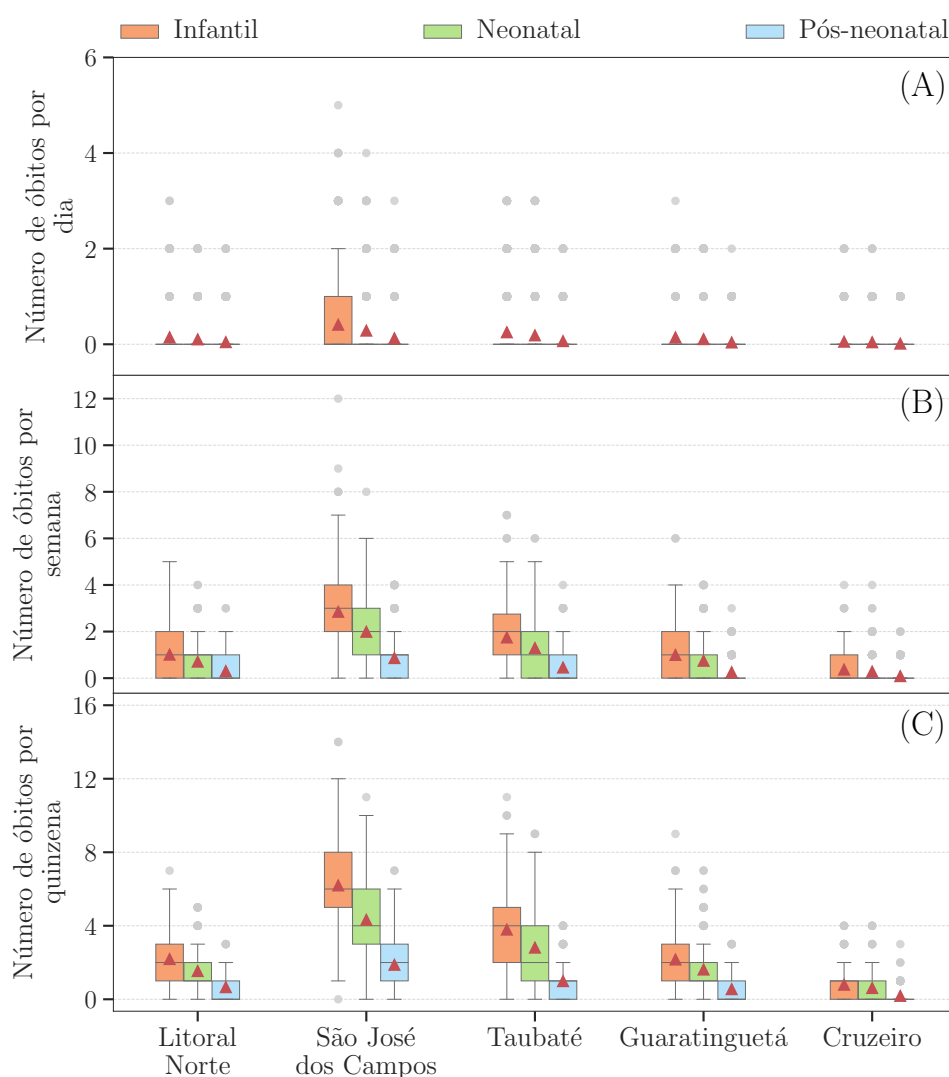
Conforme apresentado na Tabela 32 e Figura 101 (A), a mortalidade infantil mensal variou entre 16 e 51 óbitos, com valor médio de 30 (± 7) óbitos mensais. Ao avaliar os óbitos mensais agrupados por mês, Figura 101 (B), observou-se aumento na variação dos óbitos nos meses de março e maio, além de identificar diminuição na variação dos óbitos em dezembro. Da mesma forma, na análise anual, na Figura 101 (C) foi observada uma maior variação no número de óbitos mensais nos anos de 2015 e 2019.

Nos óbitos neonatais mensais, Tabela 32 e Figura 101 (A), observou-se uma variação mensal entre 10–37 óbitos, com aproximadamente 22 (± 5) óbitos como valor médio mensal. Ao avaliar os óbitos neonatais mensais agrupados por mês e ano de ocorrência, Figura 101 (B) e (C), observou-se variação semelhante ao comportamento dos óbitos infantis entre os meses e ao longo dos anos.

Na mortalidade pós-neonatal, o número de óbitos oscilou entre 1 – 23 óbitos mensais, tendo como valor médio 8 (± 3) óbitos por mês, conforme a Tabela 32 e Figura 101 (A). Na análise de óbitos mensais agrupados por mês de ocorrência, na Figura 101 (B) pode-se observar uma diminuição na variação de óbitos principalmente nos meses de fevereiro, agosto, setembro e outubro. Na avaliação anual apresentada na Figura 101 (C) pode-se observar uma diminuição da variação de óbitos mensais no 2020 e aumento no 2021.

Por outro lado, os óbitos de nascidos vivos também foram avaliados segunda a faixa etária pelas sub-regiões de Litoral Norte, São José dos Campos, Taubaté, Guaratinguetá e Cruzeiro, pertencentes à RMVPLN. A Figura 102 apresenta o gráfico de caixa do número de óbitos por dia (A), semana (B) e quinzena (C).

Figura 102 – Distribuição do número de óbitos por dia, semana e quinzena de cada sub-região por faixa etária, período 2010 – 2021

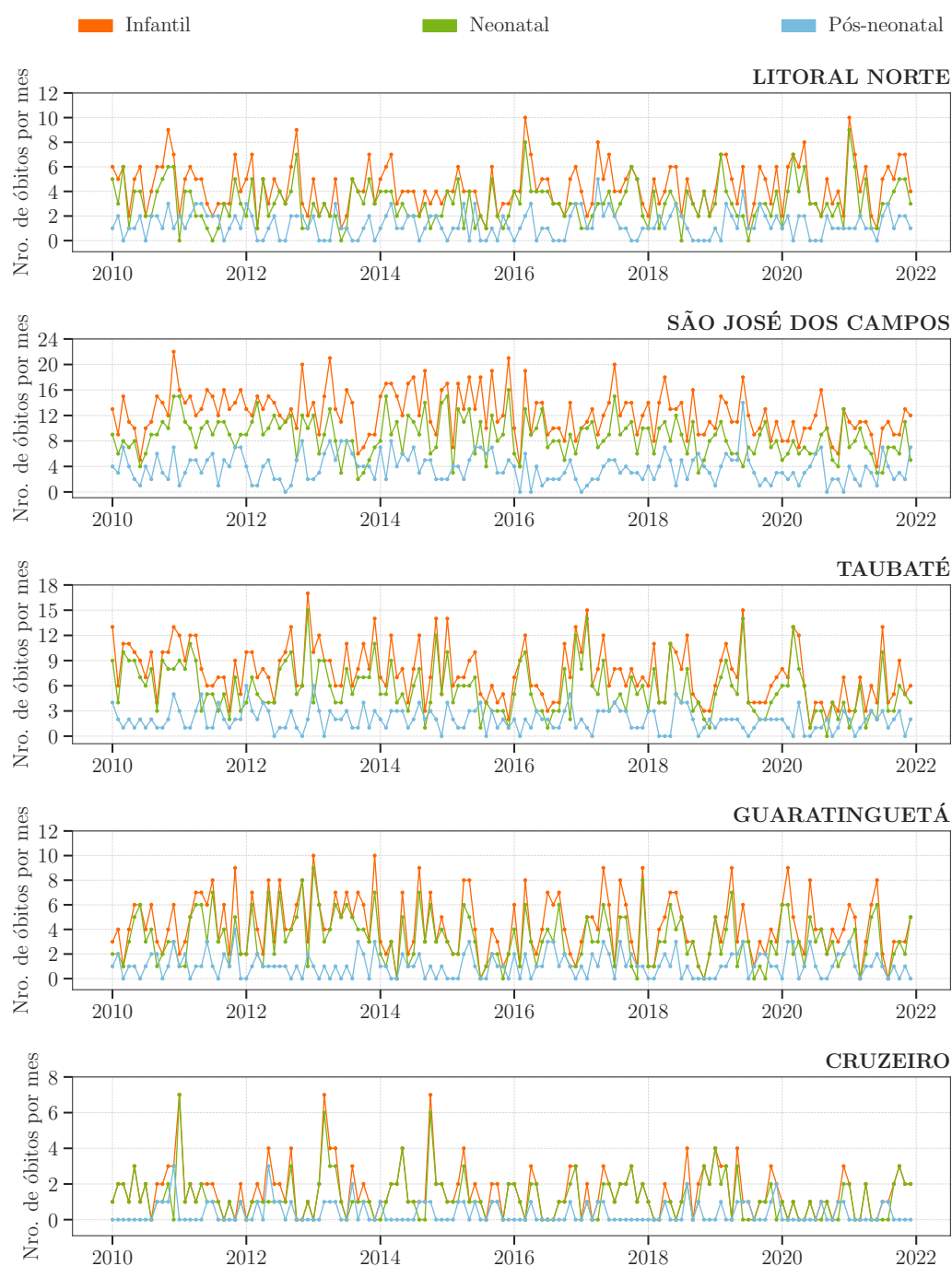


Fonte: Elaborado pela autora (2025).

Da Figura 102, foram observados valores *outliers* (discrepantes) em cada um dos grupos de ocorrência analisados. No número de óbitos por dia (A), observou-se sobreposição dos quartis em cada um das sub-regiões. Ao avaliar os óbitos por semana (B), foi observada sobreposição nos gráficos de caixa, principalmente entre o limite inferior e o primeiro quartil das sub-regiões de Litoral Norte, Guaratinguetá e Cruzeiro. Já na avaliação de óbitos por quinzena (C), observou-se que a sub-região de Cruzeiro apresentou sobreposição do limite inferior e o primeiro quartil, possivelmente por ser a sub-região com menor número nascidos vivos e óbitos.

A Figura 103 ilustra a distribuição temporal do número de óbitos mensal de cada sub-região. Na Tabela 31 apresenta-se as estatísticas descritivas do número de óbitos mensais de cada sub-região da RMVPLN.

Figura 103 – Distribuição do número de óbitos por mês de cada sub-região por faixa etária, período 2010 – 2021



Fonte: Elaborado pela autora (2025).

Como se observa na Figura 103 e na Tabela 31, na sub-região Litoral Norte o número óbitos infantis por mês no período analisado oscilou entre 1 e 10, com valor médio de 4 (± 2) óbitos mensais, na faixa etária neonatal variou entre 0 – 9 e na faixa etária pós-neonatal entre 0 – 5, com valor médio de 3 (± 2) e 1 (± 1) óbitos mensais respectivamente.

Tabela 31 – Estatísticas descritivas do número de óbitos mensais de cada sub-região

Faixa Etária	Contagem	Média	Var ^a	Desv. pad. ^b	Mín ^c	Q ^d 1	Q2	Q3	Máx ^e
Litoral Norte									
Infantil	144	4,37	3,60	1,90	1	3	4	6	10
Neonatal	144	3,06	2,80	1,67	0	2	3	4	9
Pós-neonatal	144	1,31	1,07	1,03	0	1	1	2	5
São José dos Campos									
Infantil	144	12,38	12,52	3,54	4	10	12	14	22
Neonatal	144	8,64	8,99	3,00	2	6	9	11	16
Pós-neonatal	144	3,74	5,00	2,24	0	2	4	5	14
Taubaté									
Infantil	144	7,58	10,69	3,27	1	5	7	10	17
Neonatal	144	5,61	9,09	3,02	0	3	5	8	15
Pós-neonatal	144	1,97	1,77	1,33	0	1	2	3	6
Guaratinguetá									
Infantil	144	4,32	5,86	2,42	0	3	4	6	10
Neonatal	144	3,23	4,19	2,05	0	2	3	5	9
Pós-neonatal	144	1,09	0,98	0,99	0	0	1	2	4
Cruzeiro									
Infantil	144	1,58	1,91	1,38	0	1	1	2	7
Neonatal	144	1,22	1,51	1,23	0	0	1	2	7
Pós-neonatal	144	0,36	0,36	0,60	0	0	0	1	3

(a) Var: Variância, (b) Desv. pad.: Desvio padrão, (c) Mín: Mínimo, (d) Q: Quartil, (e) Máx: Máximo.

Fonte: Elaborado pela autora (2025).

Na sub-região São José dos Campos (Figura 103 e Tabela 31), o número mensal de óbitos infantis variou entre 4 – 22, tendo como valor médio 12 (± 4) óbitos por mês, as mortes neonatais mensais oscilaram entre 2 – 16 com um valor médio de 9 (± 3) óbitos e as mortes pós-neonatais variaram entre 0 – 14 com um valor médio mensal de 4 (± 2).

A sub-região de Taubaté, conforme Figura 103 e Tabela 31, apresentou diferente variação entre o número mensal de óbito para cada faixa etária. O número mensal de óbitos infantis situou-se entre 1 – 17, óbitos neonatais entre 0 – 15 e óbitos pós-neonatais entre 0 – 6. Além disso, o número médio de óbitos em cada faixa etária foi de 4 (± 2), 3 (± 2) e 2 (± 1), respectivamente.

Na sub-região Guaratinguetá (Figura 103 e Tabela 31), o número óbitos infantis por mês no período analisado variou entre 0 e 10, com valor médio de 4 (± 2) óbitos mensais. A faixa etária neonatal variou entre 0 – 9 com valor médio de 3 (± 2), e a faixa etária pós-neonatal variou entre 0 – 4 com valor médio de 1 (± 1) óbitos mensais.

Já na sub-região de Cruzeiro o número óbitos infantis e neonatais mensal variou entre 0 – 7, tendo como valor médio 2 (± 1) e 1 (± 1) óbitos por mês, respectivamente. O número óbitos pós-neonatais oscilou entre 0 – 3, com um valor médio mensal de 0 (± 1), como apresentado na Figura 103 e Tabela 31.

APÊNDICE K – EXPLORAÇÃO DAS CARACTERÍSTICAS MATERNAS, PATERNAS E FÍSICAS DOS NASCIDOS VIVOS E ÓBITOS

A análise das informações coletadas sobre as características dos nascidos vivos da RMVPLN permitiu compreender os aspectos relacionados aos recém-nascidos vivos e óbitos. Além disso, a análise exploratória das variáveis possibilitou identificar a presença de valores atípicos e as relações entre elas, além de auxiliar na identificação das características mais relevantes para uma seleção mais adequada de variáveis para os modelos.

A análise das variáveis foi realizada, levando em consideração a classificação das mesmas como numéricas ou categóricas. Na Tabela 32 apresentam-se as estatísticas descritivas das variáveis numéricas relacionadas às características dos recém-nascidos. As análises foram realizadas de acordo com a rotulagem dos dados, ou seja, as classes-alvo (vivos ou óbitos).

Tabela 32 – Estatísticas descritivas das variáveis numéricas

Variável	Média	Desv. pad.^a	Mín^b	Q^c1	Q2	Q3	Máx^d
Classe vivos							
IDADEMAE	27,21	6,53	8,00	22,00	27,00	32,00	63,00
IDADEPAI	30,51	7,78	12,00	25,00	30,00	35,00	99,00
PESO	3.163,12	521,13	100,00	2.890,00	3.190,00	3.490,00	6.350,00
SEMAGESTAC	38,34	2,01	19,00	38,00	39,00	39,00	45,00
QTDGESTANT	1,24	1,38	0,00	0,00	1,00	2,00	55,00
APGAR1	8,40	1,10	0,00	8,00	9,00	9,00	10,00
APGAR5	9,39	0,73	0,00	9,00	9,00	10,00	10,00
QTDFILVIVO	1,04	1,18	0,00	0,00	1,00	2,00	30,00
QTDFILMORT	0,26	0,58	0,00	0,00	0,00	0,00	15,00
QTDPARTNOR	0,60	1,16	0,00	0,00	0,00	1,00	58,00
QTDPARTCES	0,53	0,82	0,00	0,00	0,00	1,00	80,00
Classe óbitos							
IDADEMAE	27,00	7,22	12,00	21,00	27,00	32,00	58,00
IDADEPAI	30,17	8,23	14,00	24,00	29,00	35,00	67,00
PESO	1.812,73	1.064,25	100,00	815,00	1.665,00	2.760,00	6.175,00
SEMAGESTAC	31,99	6,03	19,00	27,00	33,00	38,00	45,00
QTDGESTANT	1,41	1,60	0,00	0,00	1,00	2,00	12,00
APGAR1	5,27	2,86	0,00	3,00	6,00	8,00	10,00
APGAR5	6,95	2,73	0,00	6,00	8,00	9,00	10,00
QTDFILVIVO	1,09	1,35	0,00	0,00	1,00	2,00	11,00
QTDFILMORT	0,35	0,71	0,00	0,00	0,00	1,00	7,00
QTDPARTNOR	0,73	1,29	0,00	0,00	0,00	1,00	11,00
QTDPARTCES	0,48	0,79	0,00	0,00	0,00	1,00	5,00

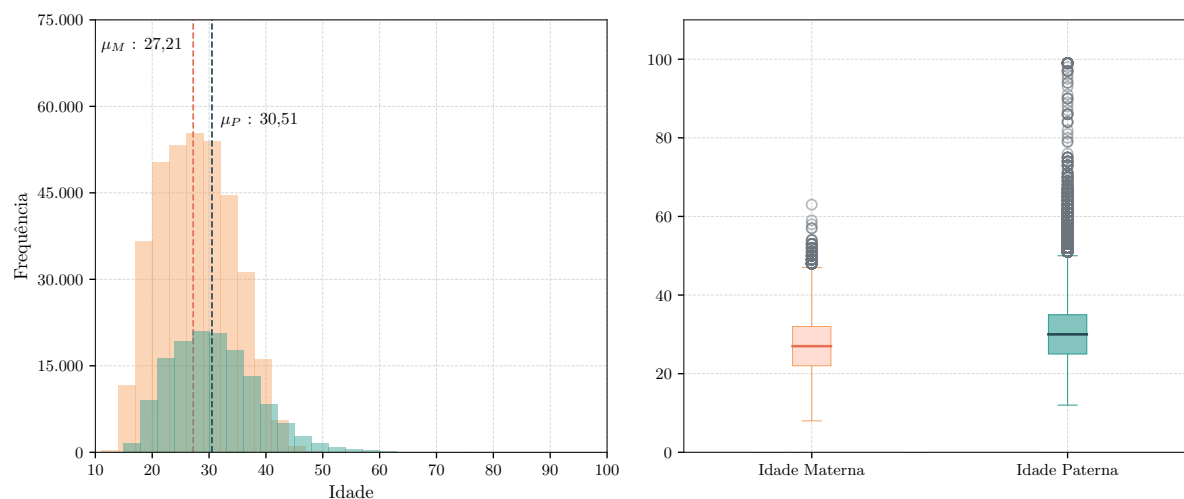
(a) Desv. pad.: Desvio padrão, (b) Mín: Mínimo, (c) Q: Quartil, (d) Máx: Máximo.

Fonte: Elaborado pela autora (2025).

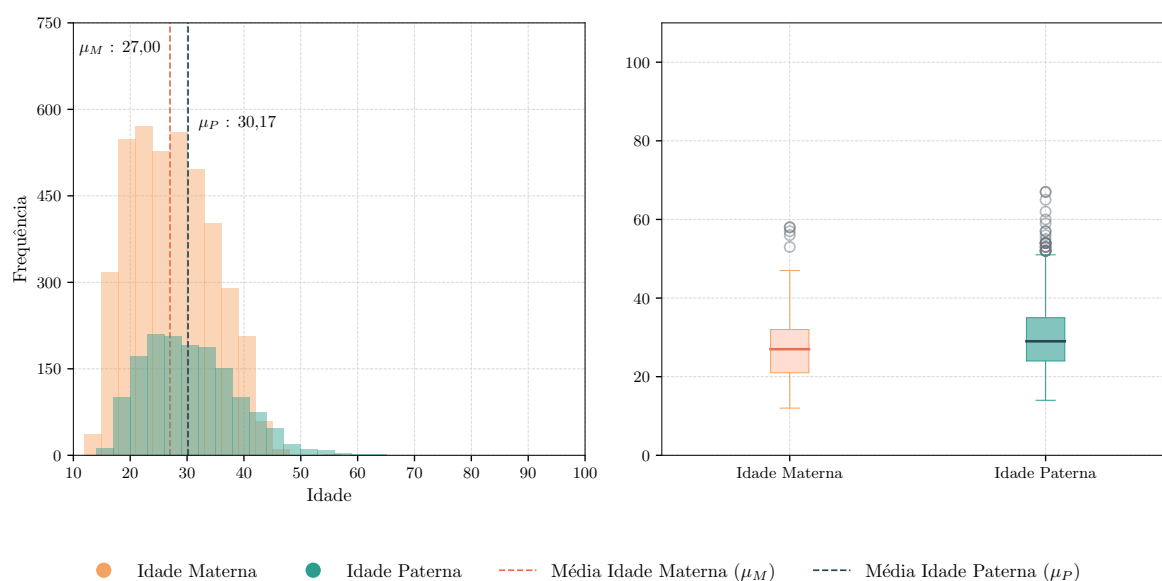
Adicionalmente, para cada variável foi gerado um gráfico de visualização, a fim de facilitar a compreensão do comportamento dos dados.

Figura 104 – Comparação da idade materna e paterna nas classes de recém-nascidos vivos e óbitos

(A) CLASSE VIVOS



(B) CLASSE ÓBITOS



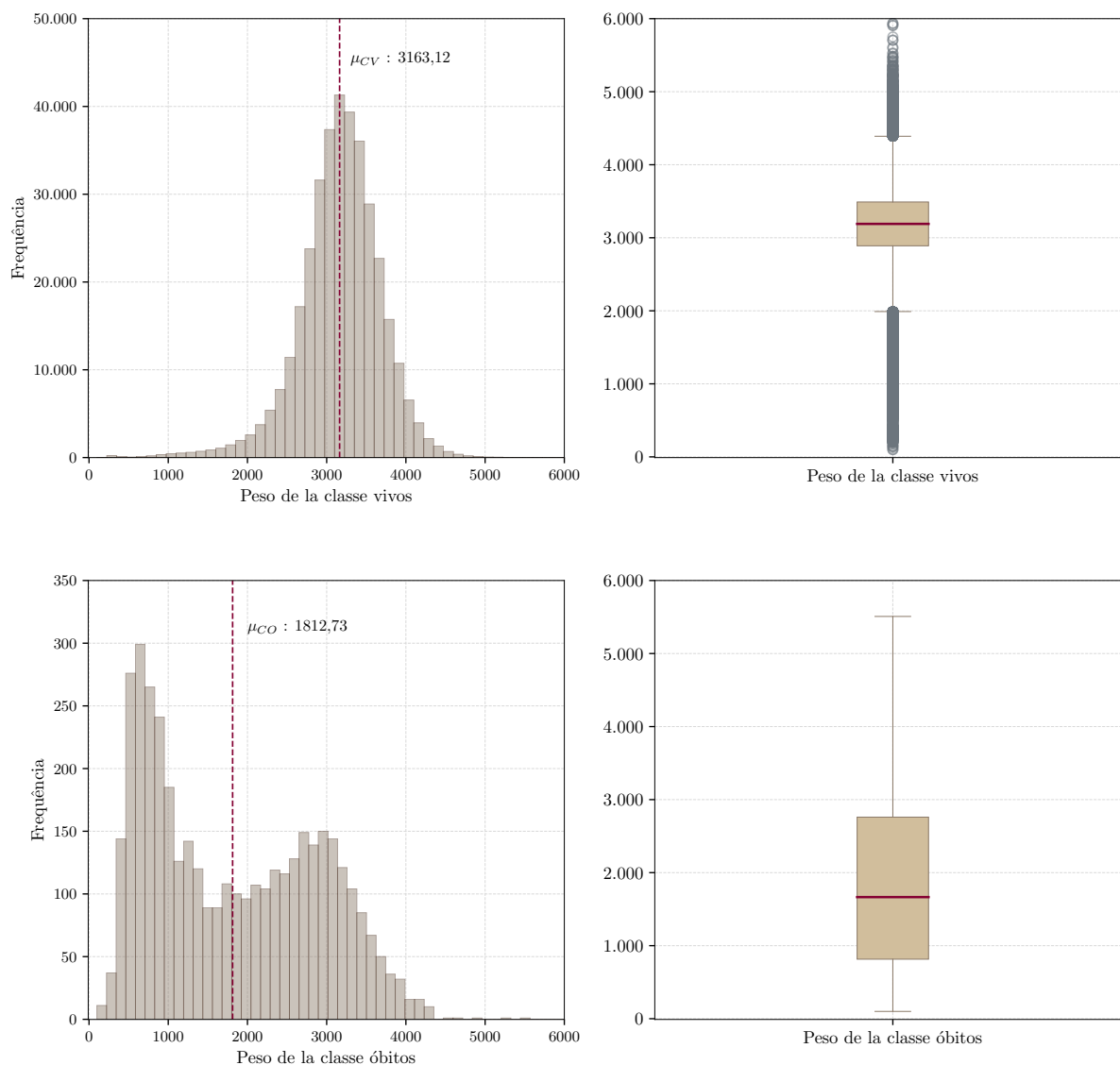
Fonte: Elaborado pela autora (2025).

Como apresentado na Tabela 32 e Figura 104, na RMVPLN a idade média das mães que tiveram filhos vivos foi de 27,21 ($\pm 6,53$) anos, enquanto aquelas mães que seus filhos viraram óbito foi de 27,00 ($\pm 7,22$) anos. Em ambos os casos, foram observadas variações significativas, com idades mínimas de 8 e 12 anos, e máximas de 63 e 58 anos, respectivamente.

Em relação à idade paterna, a média observada foi de 30,51 ($\pm 7,78$) anos para os recém-nascidos vivos e 30,17 ($\pm 8,23$) anos para os óbitos, com um intervalo de idade que variou entre a idade mínima de 12 e 14 anos e idade máxima de 99 e 67 anos, respectivamente. Este último valor da idade paterna, no entanto, pode indicar um possível erro nos dados.

Com respeito ao peso (Tabela 32 e Figura 105), o peso médio dos recém-nascidos vivos foi de 3.163,12 ($\pm 521,13$) gramas, no entanto, a variabilidade é alta, com pesos que vão de 100 a 6.350 gramas, o que aponta para a necessidade de atenção em casos de baixo peso ao nascer.

Figura 105 – Comparação do peso das classes de recém-nascidos vivos e óbitos



CV: Classe vivos, CO: Classe óbitos.

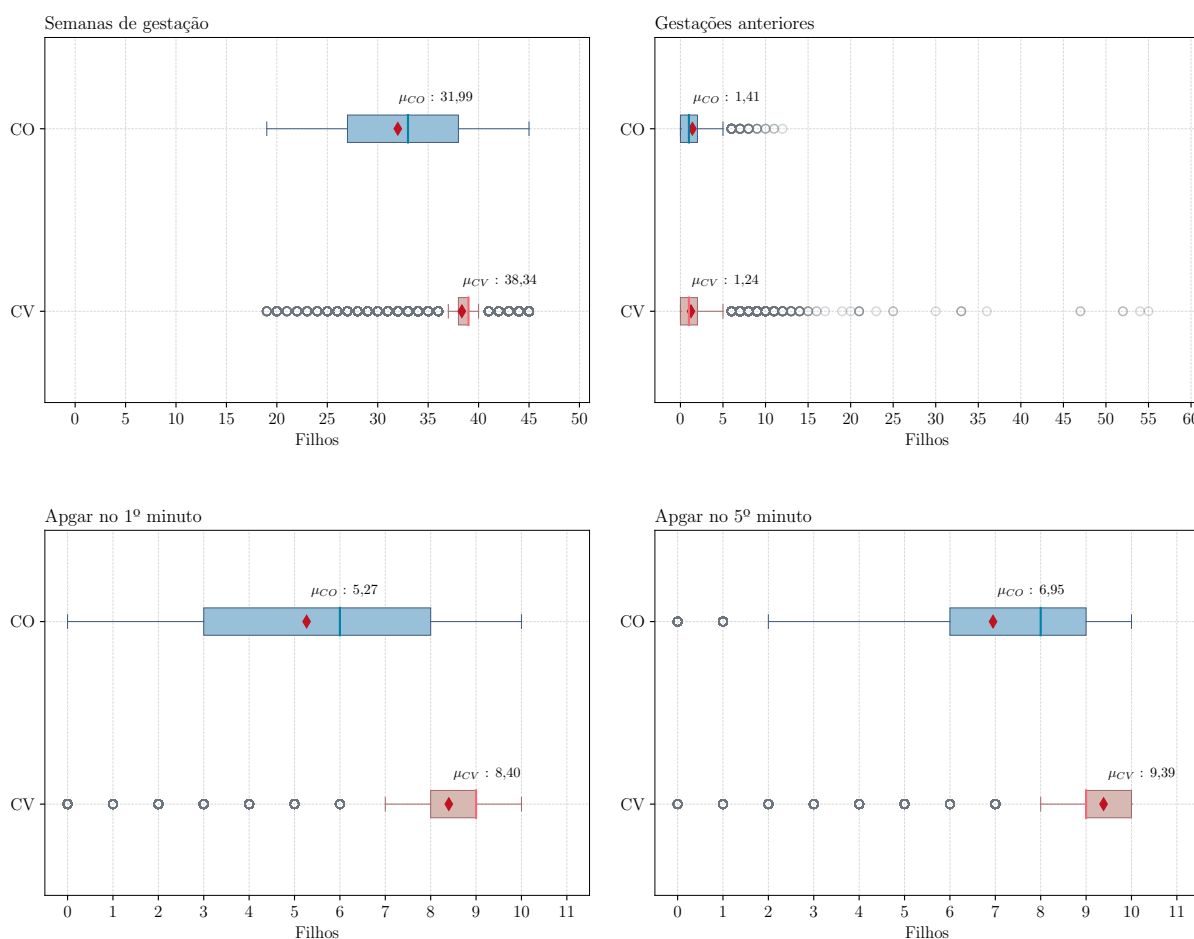
Fonte: Elaborado pela autora (2025).

Por outro lado, o peso médio dos recém-nascidos que vieram a óbito foi de 1.812,73 ($\pm 1.064,25$) gramas. O peso médio dos recém-nascidos que evoluíram para óbito foi significativamente menor, variando de 100 gramas (peso mínimo) a 6.175 gramas (peso máximo) (Figura 105). Essa situação é alarmante, pois indica um alto risco associado ao baixo peso ao nascer com a mortalidade infantil.

A duração da gestação (SEMAGESTAC) é um aspecto relevante. Da Tabela 32 e Figura 106, a duração média da gestação foi de 38,34 ($\pm 2,01$) semanas para os recém-nascidos vivos,

enquanto para os óbitos foi de 31,99 ($\pm 6,03$) semanas. Esses resultados destacam a necessidade de medidas preventivas, dada a relação entre partos prematuros e mortalidade.

Figura 106 – Comparação da duração da gestação, do número de gestações passadas e valores *apgar* das classes de recém-nascidos vivos e óbitos



CV: Classe vivos, CO: Classe óbitos.

Fonte: Elaborado pela autora (2025).

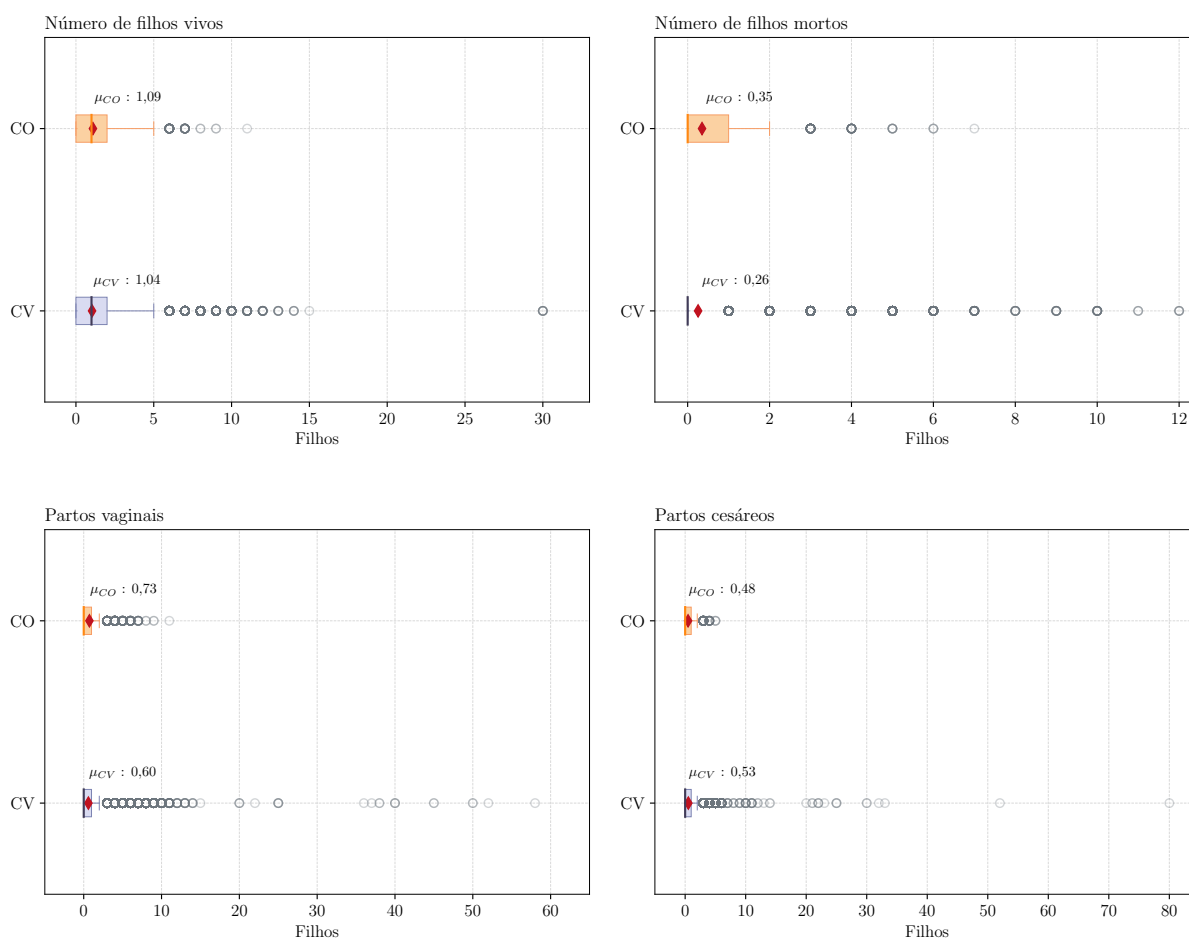
Quanto ao número de gestações passadas QTDGESTANT, foi observado como valor médio 1,24 ($\pm 1,38$) gestações passadas de mães com filhos vivos, e 1,41 ($\pm 1,60$) gestações passadas de mães que perderam filhos (Figura 106). Os valores observados no desvio padrão sugerem que há uma grande variação, com algumas mães tendo muitas gestações anteriores e outras nenhuma.

Em termos das pontuações de *apgar* (APGAR1 e APGAR5), os recém-nascidos vivos obtiveram uma média de 8,40 ($\pm 1,10$) e 9,39 ($\pm 0,73$) no primeiro e quinto minuto, respectivamente. Esses valores refletem uma boa adaptação do recém-nascido à vida extrauterina. Já os recém-nascidos que não sobreviveram mostraram pontuações de *apgar* muito mais baixas, com média de 5,27 ($\pm 2,86$) no primeiro minuto e 6,95 ($\pm 2,73$) no quinto minuto, indicando dificuldades significativas do recém-nascido associados a um maior risco de complicações.

Por sua vez, além das estatísticas apresentadas na Tabela 32, a Figura 107 mostra a vi-

sualização dos dados relacionados ao número de filhos vivos (QTDFILVIVO), filhos mortos (QTDFILMORT), partos vaginais (QTDPARTNOR) e partos cesáreos (QTDPARTCES).

Figura 107 – Comparação do número de filhos vivos, filhos mortos, partos vaginais e partos cesáreos das classes de recém-nascidos vivos e óbitos



CV: Classe vivos, CO: Classe óbitos.

Fonte: Elaborado pela autora (2025).

Assim, conforme ilustrado na Figura 107, na classe de recém-nascidos vivos, a média de filhos vivos foi de 1,04 ($\pm 1,18$), com um intervalo que varia de 0 a 30 filhos. Esse intervalo mostra que muitas mães não tiveram gestações anteriores ou tiveram apenas um filho, além de se observar uma variação significativa com o número máximo de filhos, indicando um possível erro no preenchimento dos dados. No entanto, na classe de óbitos a média de filhos vivos foi de 1,09 ($\pm 1,35$), com uma variação entre 0 e 11 filhos.

Em relação ao número de filhos mortos, a média da classe de recém-nascidos mortos foi de 0,26 ($\pm 0,58$), variando de 0 a 15 filhos como mínimo e máximo. Já na classe de óbitos a média de filhos mortos foi de 0,35 ($\pm 0,71$), com uma variação entre 0 e 7 filhos. Esses resultados demonstram que muitas mães não enfrentaram perdas, mas há uma minoria que lidou com múltiplas perdas, sugerindo a presença de fatores de risco ou condições adversas que podem

afetar a saúde neonatal.

Referente ao tipo de parto dos recém-nascidos vivos, a média de partos normais foi de 0,60 ($\pm 1,16$), apresentando uma variação que vai de 0 até 58 partos. Embora o número máximo de 58 partos indique a possibilidade de dados atípicos, isso pode evidenciar inconsistências ou erros de registro. A média de partos cesáreos foi de 0,53 ($\pm 0,82$), com um máximo de 80 partos, sugerindo a ocorrência de inconsistência no registro.

Em contrapartida, concernente ao tipo de parto dos recém-nascidos óbitos, a média de partos normais foi de 0,73 ($\pm 1,29$) com um máximo de 11, e a média de partos cesáreos foi de 0,48 ($\pm 0,79$) com um máximo de 5 partos. Esses resultados sugerem que, embora a maioria dos partos tenha sido normal, foi observada uma proporção significativa de partos cesáreos, o que reflete a necessidade de um monitoramento cuidadoso das condições maternas e neonatais.

Em adição, na Tabela 33 são apresentadas as estatísticas descritivas das variáveis categóricas relacionadas às características dos recém-nascidos, segregadas de acordo com as classes de vivos e óbitos.

Tabela 33 – Estatísticas descritivas das variáveis categóricas

(continua)

Variável	Categorias	Classe			
		Vivos		Óbitos	
		Freq. ^a	% ^b	Freq.	%
LOCNASC	Hospital	357.023	99,23	3.963	98,39
	Outro Estabelecimento de Saúde	1.665	0,46	30	0,74
	Domicílio	898	0,25	28	0,70
	Outros	172	0,05	6	0,15
	Ignorado	23	0,01	1	0,02
ESTCIVMAE	Solteira	144.942	40,46	1.850	46,10
	Casada	163.722	45,70	1.593	39,70
	Viúva	763	0,21	10	0,25
	Separada judicialmente ou divorciada	7.306	2,04	87	2,17
	União estável	34.060	9,51	390	9,72
	Ignorado	7.426	2,07	83	2,07
CODOCUPMAE	G1 ^c	27.749	36,95	281	37,47
	G2 ^d	15.116	20,13	124	16,53
	G3 ^e	14.106	18,78	149	19,87
	G4 ^f	10.290	13,70	115	15,33
	G5 ^g	3.867	5,15	42	5,60
	G6 ^h	2.859	3,81	31	4,13
	G7 ⁱ	928	1,24	8	1,07
	G8 ^j	188	0,25	0	0,00

(a) Freq.: Frequência; (b) %: Percentagem; (c) G1: Trabalhadores dos serviços, vendedores do comércio em lojas e mercados; (d) G2: Profissionais das ciências e das artes; (e) G3: Trabalhadores de serviços administrativos; (f) G4: Técnicos de nível médio; (g) G5: Membros superiores do poder público, dirigentes de organizações de interesse público e de empresas, gerentes; (h) G6: Trabalhadores da produção de bens e serviços industriais; (i) G7: Trabalhadores agropecuários, florestais e da pesca, G8: Trabalhadores em serviços de reparação e manutenção; (j) G9: Membros das forças armadas, policiais e bombeiros militares.

Fonte: Elaborado pela autora (2025).

Tabela 33 – Estatísticas descritivas das variáveis categóricas

(continua)

Variável	Categorias	Classe			
		Vivos		Óbitos	
		Freq. ^a	% ^b	Freq.	%
STTRABPART	Sim	56.799	20,29	390	12,67
	Não	217.487	77,70	2.607	84,67
	Ignorado	5.618	2,01	82	2,66
STCESPARTO	Sim	117.013	43,73	1.048	35,76
	Não	53.573	20,02	574	19,58
	Não aplica	91.036	34,02	1.226	41,83
	Ignorado	5.963	2,23	83	2,83
TPAPRESENT	Cefálico	265.305	93,95	2.582	83,18
	Pélvica ou podálica	12.757	4,52	411	13,24
	Transversal	533	0,19	29	0,93
	Ignorado	3.796	1,34	82	2,64
TPNASCASSI	Médico	244.261	94,06	2.698	95,95
	Enfermeira/obstetriz	14.575	5,61	86	3,06
	Parteira	95	0,04	3	0,11
	Outros	580	0,22	14	0,50
	Ignorado	185	0,07	11	0,39
GRAVIDEZ	Única	351.576	97,81	3.595	89,49
	Dupla	7.727	2,15	397	9,88
	Tripla ou mais	140	0,04	25	0,62
	Ignorado	11	0,00	0	0,00
PARTO	Vaginal	125.717	34,98	1.746	43,44
	Cesário	233.708	65,02	2.273	56,56
	Ignorado	6	0,00	0	0,00
RACACORMAE	Branca	195.265	70,21	2.086	68,10
	Preta	11.743	4,22	189	6,17
	Amarela	632	0,23	6	0,20
	Parda	70.12	25,21	774	25,27
	Indígena	337	0,12	8	0,26
PARIDADE	Nulípara	89.974	39,31	957	39,24
	Múltipara	138.892	60,69	1.482	60,76
SEXO	Masculino	183.835	51,10	2.196	54,85
	Feminino	175.901	48,89	1.806	45,10
	Ignorado	20	0,01	2	0,05
RACACOR	Branca	261.478	74,16	2.84	72,23
	Preta	12.418	3,52	195	4,96
	Amarela	696	0,20	7	0,18
	Parda	77.598	22,01	880	22,38
	Indígena	399	0,11	10	0,25

(a) Freq.: Frequência; (b) %: Percentagem.

Fonte: Elaborado pela autora (2025).

Tabela 33 – Estatísticas descritivas das variáveis categóricas

		(conclusão)			
Variável	Categorias	Classe			
		Vivos		Óbitos	
		Freq. ^a	% ^b	Freq.	%
IDANOMAL	Sim	2.064	0,58	527	13,19
	Não	354.167	99,07	3.415	85,50
	Ignorado	1.251	0,35	52	1,30
ESCMAE	Nenhuma	553	0,16	9	0,23
	1 – 3 anos	4.582	1,30	66	1,68
	4 – 7 anos	38.539	10,94	584	14,84
	8 – 11 anos	229.847	65,24	2.571	65,34
	12 a mais	77.709	22,06	690	17,53
	Ignorado	1.061	0,30	15	0,38
GESTACAO	Menos de 22 semanas	117	0,03	104	2,66
	22 – 27 semanas	764	0,21	973	24,91
	28 – 31 semanas	2.933	0,82	654	16,74
	32 – 36 semanas	33.604	9,44	904	23,14
	37 – 41 semanas	312.277	87,68	1.227	31,41
	42 ou mais semanas	6.293	1,77	37	0,95
	Ignorado	165	0,05	7	0,18
CONSULTAS	Nenhuma	4.519	1,26	191	4,74
	1 – 3 consultas	11.815	3,29	624	15,50
	4 – 6 consultas	58.946	16,40	1.465	36,39
	7 ou mais consultas	281.117	78,22	1.672	41,53
	Ignorado	3.001	0,84	74	1,84

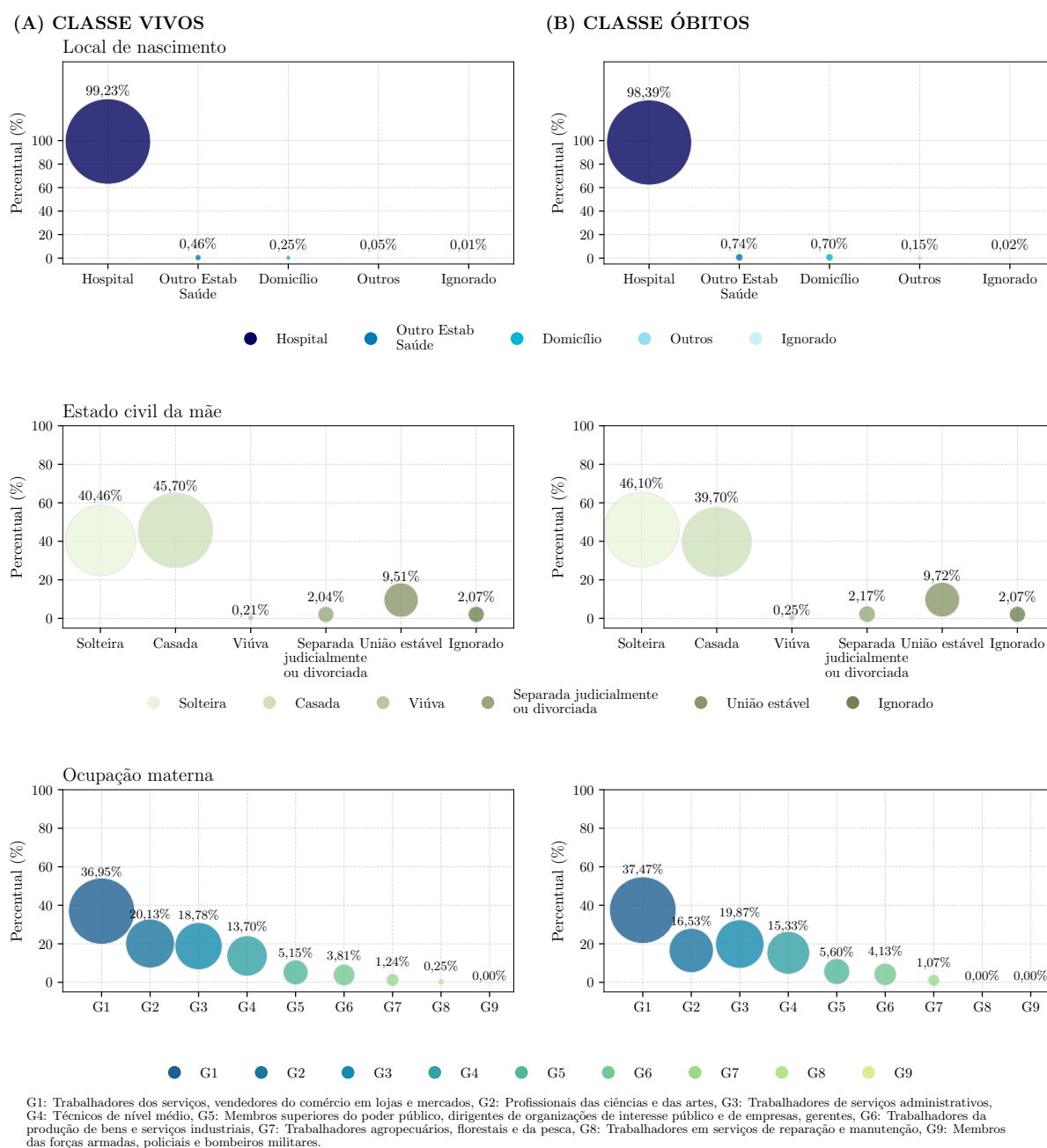
(a) Freq.: Frequência; (b) %: Percentagem.

Fonte: Elaborado pela autora (2025).

Conforme apresentado na Tabela 33 e Figura 108, em relação ao local de nascimento dos recém-nascidos vivos (LOCNASC), a maioria dos nascimentos ocorreram em hospitais (99,23%), indicando um bom acesso aos serviços de saúde. A baixa proporção de nascimentos em domicílios (0,25%) ou em outros estabelecimentos de saúde (0,46%) pode sugerir uma preferência por partos hospitalares. Já nos recém-nascidos que evoluíram para óbito a maior parte dos óbitos também ocorreu em hospitais (98,39%), sendo apenas uma pequena quantidade de falecimentos em outros locais, o que pode indicar a presença de complicações durante o parto ou de atendimento inadequado.

No que diz respeito ao estado civil das mães (ESTCIVMAE), na classe de recém-nascidos vivos observou-se uma alta proporção de mães casadas (45,70%) e solteiras (40,46%), além de uma baixa proporção de viúvas (0,21%) e mães separadas (2,04%). Apesar da significativa quantidade de solteiras, a baixa proporção de viúvas e mães separadas pode indicar uma tendência a formar uniões estáveis durante o período da maternidade.

Figura 108 – Comparação do local de nascimento, estado civil da mãe e ocupação materna das classes de recém-nascidos vivos e óbitos



Fonte: Elaborado pela autora (2025).

Por outro lado, na classe de óbitos (Tabela 33 e Figura 108), as mães solteiras apresentaram uma maior proporção de óbitos (46,10%) em comparação com as casadas (39,70%). Isso pode sugerir que as mães solteiras enfrentam mais riscos, seja por condições sociais desfavoráveis ou pela falta de apoio emocional e/ou financeiro.

Na distribuição das ocupações (CODOCUPMAE), analisada de acordo com as categorias dos grandes grupos de famílias de ocupações da Classificação Brasileira de Ocupações (CBO), observou-se que em ambos os casos, o grupo 1 (G1), relacionado a atividades de trabalhadores

dos serviços, vendedores do comércio em lojas e mercados, apresentou uma maior frequência, com uma proporção de 36,95% de mães na classe de recém-nascidos vivos e 37,47% de mães na classe de recém-nascidos óbitos.

Mães dos grupos G2 (profissionais das ciências e artes), G3 (trabalhadores de serviços administrativos) e G4 (técnicos de nível médio) apresentaram menores taxas de óbitos, sugerindo que a estabilidade financeira e o acesso a melhores condições de saúde podem ser fatores relevantes. Além disso, os grupos com menor frequência de recém-nascidos vivos (G5, G6, G7 e G8) também apresentaram proporções menores de óbitos, sugerindo que aquelas ocupações menos frequentes possuem características que influenciam positivamente os resultados da gestação.

Os resultados da análise da ocupação materna sugerem que, embora o grupo G1 tenha uma alta taxa de nascimentos, a incidência de óbitos é preocupante. Portanto, investigar as condições específicas de trabalho e saúde das mães do grupo G1 pode fornecer insights valiosos para a mitigação do risco de óbito.

Na Figura 109 é apresentada uma comparação das informações relacionadas ao trabalho de parto, incluindo se a cesárea foi realizada antes do início do trabalho de parto, a posição do recém-nascido e a assistência durante o nascimento, nas classes de recém-nascidos vivos e óbitos.

Referente à indução do trabalho de parto (STTRABPART), a maioria dos recém-nascidos vivos ocorreu sem indução (77,70%), com uma maior taxa de óbitos (84,67%) na classe de dos recém-nascidos vieram a óbito. Os resultados sugerem que partos não induzidos podem estar associados a maior risco devido a possíveis complicações não previstas.

Da mesma forma, caso o trabalho de parto tenha sido por cesárea (STCESPARTO), identificou-se que a realização da cesárea antes do trabalho de parto também está associada a uma proporção significativa de óbitos (35,76%), embora também seja comum entre os recém-nascidos vivos (43,73%). Isso sugere que os partos por cesárea podem ser necessários em situações de risco, mas também podem estar associados a complicações. Por outro lado, a percentual notável de "Não aplica", indicou que nem todas as mães puderam ser avaliadas nesse aspecto.

De forma semelhante, a posição do recém-nascido pode ser um fator importante (TPAPRESENT), pois está relacionada à possibilidade de sofrer lesões durante o nascimento. A análise dos dados mostrou que, tanto nas informações relacionadas aos recém-nascidos vivos (93,95%) quanto nos óbitos (83,18%), a maioria dos nascidos vivos apresentou posição cefálica durante o parto, a qual, na maioria das vezes, está associada a melhores desfechos. Já as posições pélvica (13,24%) e transversal (0,93%) apresentaram proporções mais elevadas nas informações relacionadas ao recém-nascidos óbitos, indicando que essas condições de parto apresentam maiores riscos para o nascido vivo.

Em relação à assistência recebida pelo recém-nascido, especificamente às categorias de profissionais que realizaram o parto (TPNASCASSI), observou-se que a maioria dos partos foi conduzida por médicos, representando 94,06% dos recém-nascidos vivos e 95,95% dos óbitos.

Figura 109 – Comparação do trabalho de parto, da cesárea, posição do recém-nascido e atenção do nascimento das classes de recém-nascidos vivos e óbitos



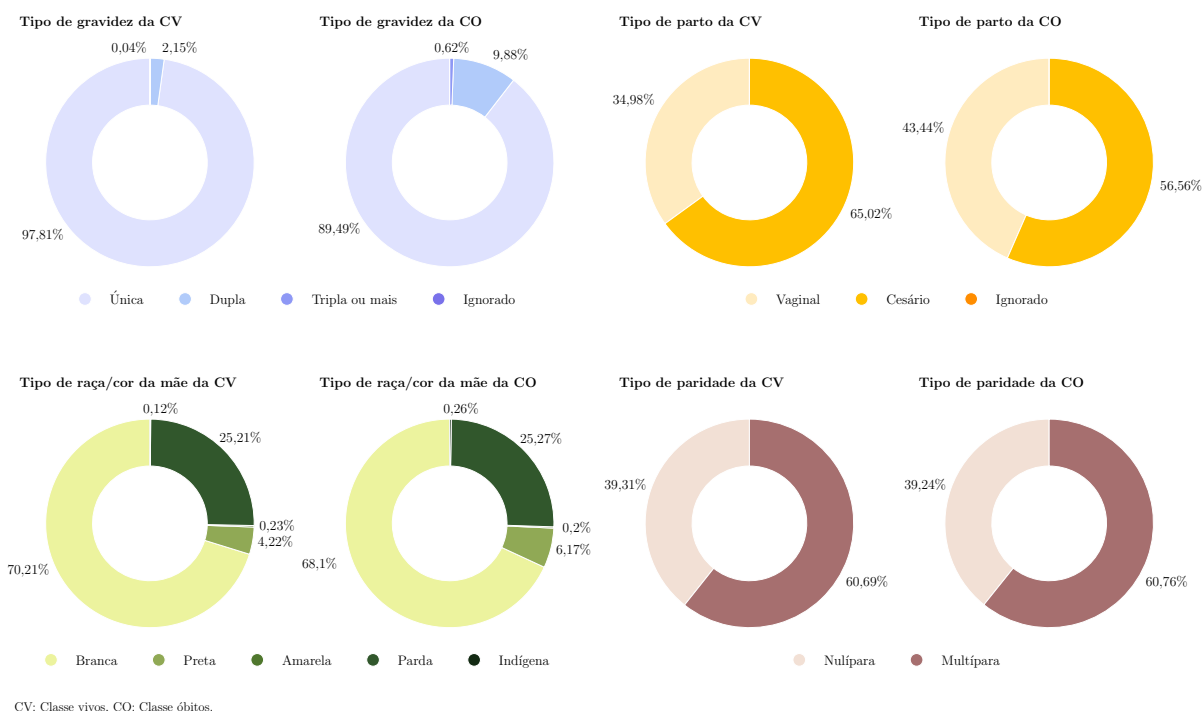
Fonte: Elaborado pela autora (2025).

Em contraste, como ilustrado na Figura 109, a assistência prestada por enfermeiras ou obstetrizas mostrou-se menos frequente, apresentando uma proporção menor entre os recém-nascidos óbitos (3,06%) em comparação aos recém-nascidos vivos (5,61%). Esses dados sugerem

que os partos assistidos por médicos tendem a envolver casos mais complexos, enquanto partos de baixo risco podem receber assistência eficaz de enfermeiras ou obstetrizas. Além disso, a assistência prestada por parteiras e outros profissionais é rara, com poucas ocorrências registradas.

Na Figura 110 apresenta-se uma comparação entre o tipo de gravidez, o tipo de parto, a raça/cor da mãe e a paridade nas classes de recém-nascidos vivos e óbitos.

Figura 110 – Comparação do tipo de gravidez, parto, raça da mãe e a paridade das classes de recém-nascidos vivos e óbitos



Fonte: Elaborado pela autora (2025).

Da Tabela 32 e Figura 110 observou-se que a maioria das gestações (GRAVIDEZ) foi única, com uma proporção de 97,81% dos recém-nascidos vivos e 89,49% dos óbitos. Gestações múltiplas mostraram proporções menores, com 2,15% dos recém-nascidos vivos e 9,88% dos óbitos para gestações duplas, e 0,04% dos recém-nascidos vivos e 0,62% dos óbitos para gestações triplas ou mais. Esses dados indicam que gestações únicas estão associadas a melhores desfechos neonatais, enquanto gestações múltiplas apresentam um risco aumentado de óbitos.

Em relação ao tipo de parto (PARTO), 65,02% dos recém-nascidos vivos foram cesarianos, com 56,56% dos recém-nascidos óbitos. Em contrapartida, 34,98% dos recém-nascidos vivos nasceram por parto vaginal, com 43,44% dos óbitos. Isso sugere que, apesar das cesarianas serem mais frequentes, elas estão ligadas a uma maior proporção de óbitos, possivelmente refletindo que esses partos são realizados em situações de maior risco.

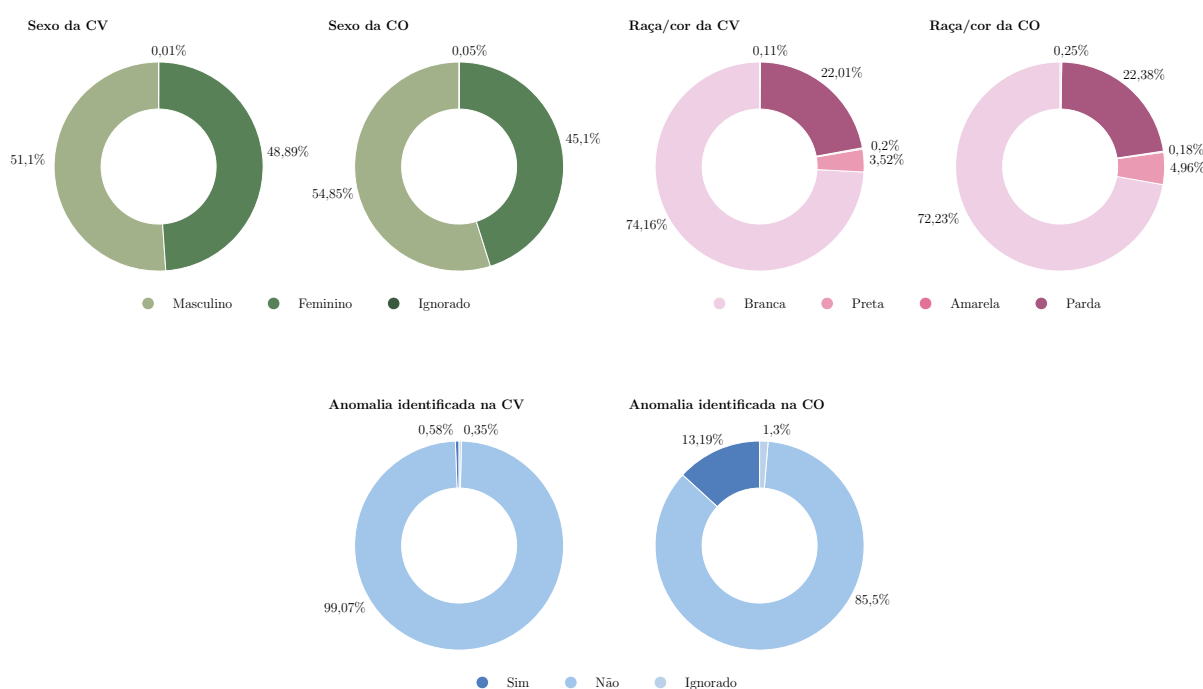
Por outro lado, a maioria dos recém-nascidos vivos foi de mães brancas (RACACORMAE), representando 70,21% dos nascidos vivos. Nos óbitos, a proporção de recém-nascidos de mães brancas alcançou 68,10%. Além disso, 25,21% dos nascimentos que sobreviveram foram de mães pardas, enquanto a proporção de óbitos de mães pardas foi de 25,27%. Mães negras

representaram 4,22% dos recém-nascidos vivos e 6,17% dos óbitos, enquanto mães indígenas e amarelas apresentaram proporções menores.

No que se refere à paridade (PARIDADE), 39,31% dos recém-nascidos vivos foram de mães nulíparas, com 39,24% dos recém-nascidos óbitos nessa categoria. Mães múltíparas representam 60,69% dos recém-nascidos vivos e 60,76% dos óbitos. Os resultados sugerem que a experiência em partos anteriores não exerce um impacto significativo nas proporções de óbitos infantis.

A Figura 111 apresenta que, entre os recém-nascidos vivos, 51,10% foram do sexo masculino e 48,89% do sexo feminino (SEXO). Nos óbitos, a proporção foi de 54,85% para o sexo masculino e 45,10% para o feminino. Esses dados revelam uma ligeira predominância de óbitos entre os recém-nascidos masculinos, sugerindo uma maior vulnerabilidade a complicações.

Figura 111 – Comparação do sexo, raça do recém-nascido e anomalia identificada das classes de recém-nascidos vivos e óbitos



CV: Classe vivos, CO: Classe óbitos.

Fonte: Elaborado pela autora (2025).

A maioria dos recém-nascidos vivos foi classificada como branca (74,16%), enquanto 72,23% dos óbitos também pertencem a essa categoria (RACACOR). Recém-nascidos pardos totalizaram 22,01% dos nascimentos que sobreviveram e 22,38% dos óbitos. Recém-nascidos pretos corresponderam a 3,52% dos vivos e 4,96% dos óbitos. As categorias amarela e indígena apresentaram proporções menores em ambos os grupos. Esses resultados foram semelhantes aos observados na variável de raça/cor da mãe.

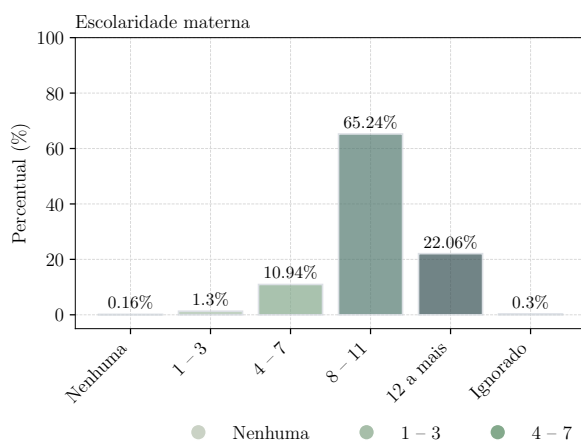
Em relação às anomalias congênitas (IDANOMAL), 0,58% dos recém-nascidos vivos apresentaram anomalias, enquanto essa proporção aumentou significativamente para 13,19% entre os óbitos. Esses valores evidenciam uma forte associação entre a presença de anomalias congênitas e

o aumento da mortalidade, indicando que recém-nascidos com anomalias enfrentaram um risco maior de óbito.

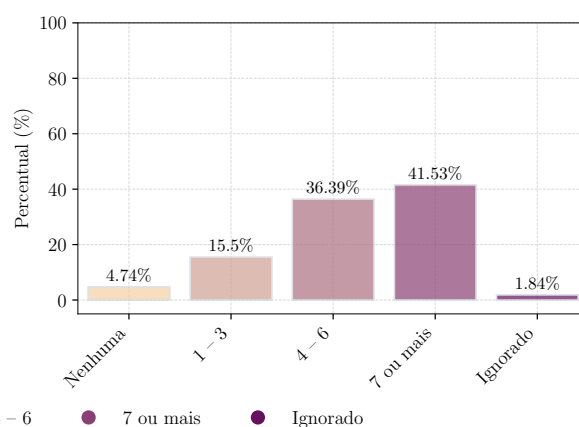
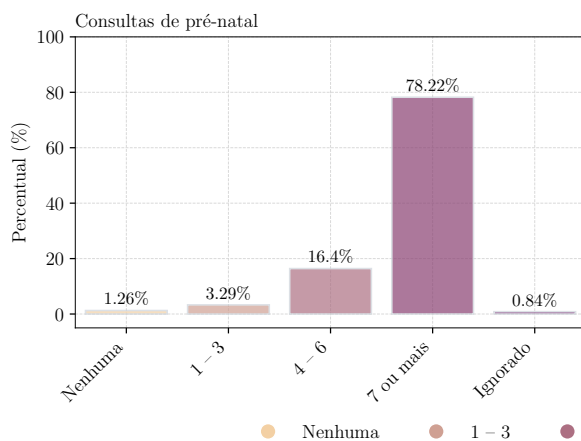
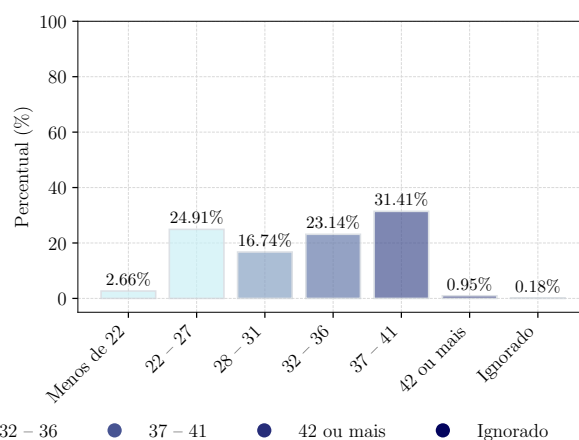
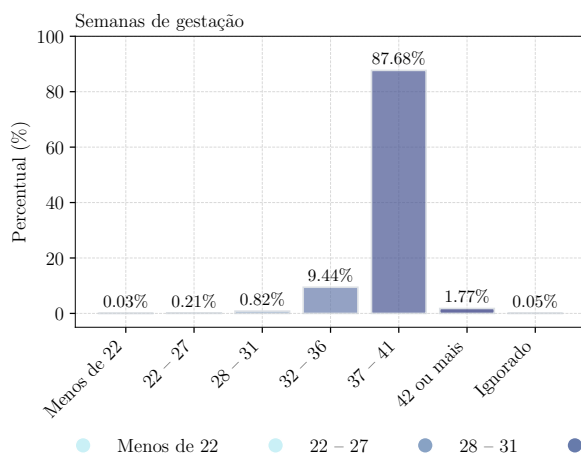
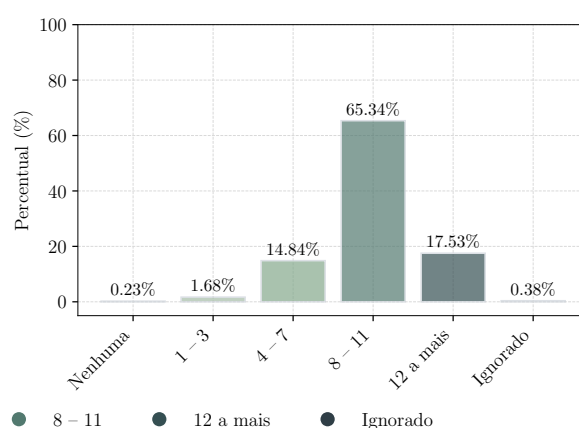
A Figura 112 apresenta a apresenta informações sobre recém-nascidos vivos e óbitos, categorizadas por escolaridade da mãe, idade gestacional e número de consultas de pré-natal.

Figura 112 – Comparação da escolaridade materna, semanas de gestação e consultas de pré-natal das classes de recém-nascidos vivos e óbitos

(A) CLASSE VIVOS



(B) CLASSE ÓBITOS



Fonte: Elaborado pela autora (2025).

Referente à escolaridade materna (ESMAE), uma grande parte dos recém-nascidos vivos (65,24%) teve mães na faixa de 8 a 11 anos de escolaridade, enquanto 65,34% dos óbitos também ocorreram nesse grupo. Essa faixa se destaca como predominante, sugerindo que uma escolaridade mais alta pode estar associada com melhores desfechos neonatais. Por outro lado, as mães sem escolaridade apresentam uma proporção muito baixa tanto em nascimentos vivos (0,16%) quanto em óbitos (0,23%).

Nas semanas de gestação (GESTACAO), 87,68% dos recém-nascidos vivos nasceu em gestação de 37 a 41 semanas, enquanto essa faixa corresponde a apenas 31,41% dos óbitos. As categorias de gestação prematura (menos de 37 semanas) mostraram uma alta proporção de óbitos, especialmente nas faixas de 22 a 27 semanas (24,91%) e 32 a 36 semanas (23,14%). Isso indica que a prematuridade está associada a um aumento significativo no risco de mortalidade neonatal.

Por fim, dado que as consultas pré-natais permitem acompanhar o desenvolvimento do feto e a saúde da gestante, sua análise possibilita observar sua importância na redução da mortalidade. Os dados observados indicaram que grande parte dos recém-nascidos vivos (78,22%) teve mães que realizaram 7 ou mais consultas de pré-natal, com apenas 1,26% dos nascimentos vivos ocorrendo em mães que não realizaram consultas. No entanto, entre os óbitos, 4,74% tiveram mães que não realizaram consultas, e essa proporção aumentou para 41,53% nos casos em que as mães realizaram 7 ou mais consultas. Essa tendência sugere que o acesso e a frequência ao pré-natal podem influenciar positivamente o cuidado e os desfechos neonatais.

As informações apresentadas nas Tabelas 32 e 33 e Figuras 104 - 112 fornecem uma visão ampla da distribuição das variáveis numéricas e categóricas, além de mostrar seu impacto nos resultados de nascimentos vivos e óbitos. Além disso, os resultados apresentados podem ser úteis na identificação de tendências, assim como servir como base para análises mais profundas ou para o planejamento de intervenções em saúde.

Por último, é importante ressaltar que a qualidade dos dados é fundamental, pois a presença de uma grande quantidade de informações com a etiqueta "Ignorado" pode afetar a interpretação dos resultados.

APÊNDICE L – PROCEDIMENTOS COMPUTACIONAIS PARA ANÁLISE ESPACIAL E MODELAGEM COM TÉCNICAS DE MACHINE LEARNING

O código utilizado na Etapa 4, relacionado ao modelamento realizado para a análise espacial e ao desenvolvimento de modelos com técnicas de ML, pode ser acessado por meio do Google Colab através do seguinte link: <https://drive.google.com/drive/folders/1JsqmuN7LYHyhIr1znjg4jWvVH8VEgIJd?usp=sharing>.

- `Path` de `pathlib`: Utilizado para manipulação de caminhos de arquivos e diretórios.
- `pandas`: Utilizado para análise de dados.
- `numpy`: Utilizado para processamento dos dados e cálculo numérico.
- `matplotlib.pyplot`: Utilizado para visualização gráfica.
- `seaborn`: Utilizado para visualização gráfica.
- `gridspec` de `matplotlib`: Usado para controlar o layout e a disposição de múltiplos gráficos dentro de uma única figura.
- `ListedColormap` de `matplotlib.colors`: Utilizada para criar mapas de cores personalizados.
- `ScaleBar` de `matplotlib_scalebar.scalebar`: Usado para adicionar barras de escala a gráficos criados com `matplotlib`.
- `matplotlib.patches`: Utilizado para desenhar formas geométricas nos gráficos.
- `Rectangle` de `matplotlib.patches`: Permite desenhar retângulos nos gráficos de `matplotlib`.
- `geopandas`: Utilizado para lidar com dados geoespaciais. O pacote é uma extensão do `pandas`.
- `pygeoda`: Utilizado para análise espacial e estatísticas espaciais. O pacote baseia-se em `GeoDa`.
- `esda`: Utilizado para análise espacial exploratória de dados.
- `libpysal`: Utilizado para análise espacial, com ferramentas para calcular pesos espaciais, análises de autocorrelação espacial, entre outras técnicas.
- `Queen` de `libpysal.weights.contiguity`: Usada para criar uma matriz de pesos espaciais baseada na contiguidade "rainha".

- `weights de libpysal`: Utilizado para trabalhar com matrizes de pesos espaciais, que são essenciais para análises de dependência espacial e autocorrelação.
- `Moran, Moran_Local, Moran_BV e Moran_Local_B de esda`: Utilizado para calcular índices de autocorrelação espacial global e a nível local.

ANÁLISE ESPACIAL

<https://drive.google.com/file/d/1SWnqFl2FUcHhHU7uMbjqfGoJGEmBWXdU/view?usp=sharing>

MODELO DE PREVISÃO

<https://drive.google.com/file/d/18ahsvTtkplKC2tw3qoKSe0WTqKgrfkQd/view?usp=sharing>

MODELO DE PREDIÇÃO

<https://drive.google.com/drive/folders/1JsquN7LYHyhIr1znjg4jWvVH8VEgIJd?usp=sharing>

APÊNDICE M – MAPA MORAN BIVARIADO ENTRE A MORTALIDADE INFANTIL, NEONATAL E PÓS-NEONATAL COM AS VARIÁVEIS SELECIONADAS

Os resultados do mapa Moran e gráfico de dispersão da análise local bivariada entre a mortalidade infantil, neonatal e pós-neonatal com as variáveis da Tabela 4 são apresentados nas Seções D1, D2 e D3.

Os mapas foram elaborados de acordo com a explicação apresentada em 2.4.2, sendo considerado como variável 1 as taxas de mortalidade e como variável 2 as variáveis da Tabela 4 (ver Figura 23). Além disso, foram gerados as figuras para cada faixa etária, sendo classificada cada variável conforme descritas a seguir:

- Caraterísticas dos pais:

- | | |
|-----------------------------------|-------------------------------------|
| (a) Mães com menos de 20 anos (%) | (g) Mães com parto cesáreo (%) |
| (b) Mães com mais de 34 anos (%) | (h) Mães solteiras (%) |
| (c) Gravidez dupla (%) | (i) Mães com baixa escolaridade (%) |
| (d) Gravidez tripla ou mais (%) | (j) Pré-natal inadequado (%) |
| (e) Mães com filhos morto (%) | (k) Pais com menos de 20 anos (%) |
| (f) Mães com parto normal (%) | (l) Pais com mais de 40 anos (%) |

- Caraterísticas do recém-nascido:

- | | |
|-----------------------------------|--|
| (a) Nascidos vivos prematuros (%) | (d) <i>Apgar</i> < 7 no 1 o minuto (%) |
| (b) Baixo peso ao nascer (%) | (e) <i>Apgar</i> < 7 no 5 o minuto (%) |
| (c) Anomalia congênita (%) | |

- Atenção à saúde:

- | | |
|--------------------------------|---------------------|
| (a) Taxa de leitos pediátricos | (b) Taxa de médicos |
|--------------------------------|---------------------|

- Caraterísticas do município:

- | | |
|----------|---------------------------|
| (a) IDH | (d) PIB |
| (b) IVS | (e) PIB <i>per capita</i> |
| (c) IPRS | |

O *link* para visualizar as imagens do Apêndice M é: https://drive.google.com/file/d/1_kKu72ibSZkMlqjQSCaPKa2T29cqp3j0/view?usp=sharing.

ANEXO A – INDICADORES SOCIAIS DA REGIÃO METROPOLITANA DO VALE DO PARAÍBA E LITORAL NORTE

Tabela 34 – Indicadores sociais da região metropolitana do Vale do Paraíba e Litoral Norte

Município	População	IPRS^a	IVS^b	IDH^c	PIB^d	PIB <i>per capita</i>
Aparecida	35.007	1	0,19	0,76	1.144.056	32.100
Arapeí	2.493	4	0,24	0,68	34.001	13.755
Areias	3.696	5	0,29	0,70	46.311	12.107
Bananal	10.223	4	0,26	0,73	181.838	17.146
Caçapava	84.752	2	0,18	0,79	4.041.591	44.591
Cachoeira Paulista	30.091	5	0,18	0,76	618.062	19.298
Campos Do Jordão	47.789	2	0,26	0,75	1.318.467	26.426
Canas	4.385	5	0,30	0,70	107.670	21.551
Caraguatatuba	100.840	2	0,20	0,76	4.034.042	35.186
Cruzeiro	77.039	5	0,20	0,79	2.380.399	29.883
Cunha	21.866	5	0,32	0,68	264.898	12.216
Guaratinguetá	112.072	2	0,18	0,80	6.488.615	55.100
Igaratá	8.831	4	0,24	0,71	209.303	22.593
Ilhabela	28.196	1	0,22	0,76	14.967.867	454.012
Jacareí	211.214	1	0,22	0,78	13.361.648	59.030
Jambeiro	5.349	2	0,18	0,76	254.628	41.484
Lagoinha	4.841	3	0,31	0,69	61.903	12.846
Lavrinhas	6.590	4	0,30	0,73	103.394	14.544
Lorena	82.537	5	0,21	0,77	2.839.020	32.921
Monteiro Lobato	4.120	4	0,33	0,71	65.120	14.696
Natividade Da Serra	6.678	4	0,34	0,66	73.592	10.994
Paraibuna	17.388	4	0,24	0,72	296.200	16.324
Pindamonhangaba	146.995	3	0,22	0,77	8.471.160	52.275
Piquete	14.107	4	0,27	0,76	165.598	12.102
Potim	19.397	5	0,25	0,70	235.286	11.439
Queluz	11.309	4	0,30	0,72	338.412	27.023
Redenção Da Serra	3.873	4	0,36	0,66	55.337	14.414
Roseira	9.599	4	0,20	0,74	299.662	28.431
Santa Branca	13.763	4	0,27	0,74	262.467	18.565
Santo Antônio Do Pinhal	6.486	3	0,23	0,71	119.547	18.080
São Bento Do Sapucaí	10.468	3	0,30	0,72	213.669	20.311
São José Do Barreiro	4.077	5	0,28	0,68	50.659	12.450
São José Dos Campos	629.921	1	0,24	0,81	43.457.370	61.829
São Luiz Do Paraitinga	10.397	4	0,28	0,70	170.041	16.119
São Sebastião	73.942	1	0,22	0,77	3.575.340	41.650
Silveiras	5.792	4	0,27	0,68	76.118	12.353
Taubaté	278.686	2	0,19	0,80	18.357.172	60.267
Tremembé	40.984	3	0,21	0,79	734.913	16.145
Ubatuba	78.801	2	0,24	0,75	2.382.689	27.111

(a) IPRS: Índice paulista de responsabilidade social, (b) IVS: Índice de vulnerabilidade social, (c) IDH: Índice de desenvolvimento humano, (d) PIB: Produto interno bruto.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2010), Instituto de Pesquisa Econômica Aplicada (2010), Fundação SEADE (2018b), Fundação SEADE (2018a).

ANEXO B – DECLARAÇÃO DE NASCIDO VIVO

Figura 113 – Declaração de Nascido Vivo

 República Federativa do Brasil Ministério da Saúde 1ª VIA - SECRETARIA DE SAÚDE		Declaração de Nascido Vivo	
I	Cartório	1 Cartório	2 Registro
		3 Data	
II	Local da Ocorrência	4 Município	5 UF
	6 Local da Ocorrência	7 Estabelecimento	8 Código
	1 Hospital 2 Outros Estab. Saúde 3 Domicílio 4 Outros 9 Ignorado		
	8 Endereço da ocorrência, se fora do estab. ou da resid. da mãe (Rua, praça, avenida, etc)	9 CEP	
III	Mãe	10 Bairro/Distrito	11 Município de ocorrência
	12 Nome da Mãe	13 Código	14 Cartão SUS
	15 Idade (anos)	16 Estado Civil	17 Escolaridade
	18 Ocupação habitual e ramo de atividade	19 N° de filhos tidos em gestações anteriores	
IV	Gestação e Parto	20 Residência da mãe	21 CEP
	22 Logradouro	23 Bairro/Distrito	24 Município
	25 Duração da gestação (em semanas)	26 Tipo de gravidez	27 Tipo de parto
	28 Número de consultas de pré-natal		
V	Recém-Nascido	29 Nascimento	30 Sexo
	31 Raça/cor	32 Peso ao nascer	33 Índice de Apgar
	34 Detectada alguma malformação congênita e/ou anomalia cromossômica?		
	35 Polegar direito da mãe	36 Pé direito da criança	
VI	Identificação		
VII	Preench.	37 Responsável pelo preenchimento	38 Função
		39 Identidade	40 Órgão Emissor
ATENÇÃO: ESTE DOCUMENTO NÃO SUBSTITUI A CERTIDÃO DE NASCIMENTO O Registro de Nascimento é obrigatório por lei. Para registrar esta criança, o pai ou responsável deverá levar este documento ao cartório de registro civil.			

Versão 12/08 - 1ª Impressão 12/2008

Fonte: Brasil (2009).

ANEXO C – DECLARAÇÃO DE ÓBITO

Figura 114 – Declaração de Óbito

República Federativa do Brasil Ministério da Saúde 1ª VIA - SECRETARIA DE SAÚDE		Declaração de Óbito	
I	Cartório	☐ Cartório Código ☐ Registro Data	
		☐ Município 5. UF ☐ Cemitério	
II	Identificação	☐ Tipo de Óbito ☐ Óbito 9. Cartão SUS 10. Naturalidade	
		☐ Fetal ☐ Não Fetal 11. Nome do falecido	
		12. Nome do pai 13. Nome da mãe	
		14. Data de Nascimento 15. Idade 16. Sexo 17. Rapa/cor	
III	Residência	☐ Estado civil ☐ Escolaridade 19. Ocupação habitual e ramo de atividade	
		☐ Solteiro ☐ Casado ☐ Viúvo ☐ Separado judicialmente ☐ Divorçado ☐ Ignorado	
IV	Ocorrência	☐ Local de ocorrência do óbito ☐ Estabelecimento 27. CEP	
		☐ Hospital ☐ Outros estab. saúde ☐ Domicílio ☐ Via pública ☐ Outros ☐ Ignorado	
V	Fatal ou menor que 1 ano	☐ Endereço da ocorrência, se fora do estabelecimento ou da residência 28. CEP	
		☐ Bairro/Distrito 29. Município de ocorrência 30. UF	
VI	Condições e causa da morte	PREENCHIMENTO EXCLUSIVO PARA ÓBITOS FETAIS E DE MENORES DE 1 ANO	
		INFORMAÇÕES SOBRE A MÃE	
		☐ Idade ☐ Escolaridade 35. Ocupação habitual e ramo de atividade da mãe 36. Número de filhos tidos	
		☐ Menos de 22 ☐ De 22 a 27 ☐ De 28 a 31 ☐ De 32 a 36 ☐ De 37 a 41 ☐ De 42 e mais ☐ Ignorado	
VII	Médico	☐ Duração da gestação (Em semanas) ☐ Tipo de Gravidez ☐ Tipo de parto 40. Morte em relação ao parto	
		☐ Única ☐ Dupla ☐ Tripla e mais ☐ Ignorada	
		☐ Vaginal ☐ Cesáreo ☐ Ignorado	
		☐ Antes ☐ Durante ☐ Depois ☐ Ignorado	
VIII	Causas externas	☐ Peso ao nascer ☐ Num. da Declar. de Nascidos Vivos	
		☐ Gramas	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
IX	Localidade	ÓBITOS EM MULHERES	
		☐ A morte ocorreu durante a gravidez, parto ou aborto? ☐ A morte ocorreu durante o puerpério?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
X	Assistência Médica	☐ Recebeu assist. médica durante a doença que ocasionou a morte?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XI	Assistência Médica	☐ Exame complementar? ☐ Cirurgia? ☐ Necropsia?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XIII	Assistência Médica	☐ Meio de contato (Telefone, fax, e-mail etc.) ☐ Data do atestado ☐ Assinatura	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XIV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XVI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XVII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XVIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XIX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXIV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXVI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXVII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXVIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXIX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXIV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXVI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXVII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXVIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XXXIX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XL	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLIV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLVI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLVII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLVIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
XLIX	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
CL	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
CLI	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
CLII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
CLIII	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
		☐ Sim ☐ Não ☐ Ignorado	
CLIV	Assistência Médica	☐ Nome do médico ☐ CRM ☐ O médico que assina atendeu ao falecido?	
		☐ Sim ☐ Não ☐ Ignorado	

ANEXO D – DESCRIÇÃO DAS VARIÁVEIS DO CONJUNTO DE DADOS DO SISTEMA DE INFORMAÇÕES SOBRE NASCIDOS VIVOS

Quadro 10 – Descrição das variáveis do conjunto de dados do SINASC

(continua)

Variável	Tipo	Descrição
LOCNASC	Categórico	Local de nascimento. 1: Hospital; 2: Outro Estabelecimento de Saúde; 3: Domicílio; 3: Outros; 9: Ignorado.
IDADEMAE	Numérico	Idade materna
ESTCIVMAE	Categórico	Estado civil da mãe: 1: Solteira; 2: Casada; 3: Viúva; 4: Separada judicialmente ou divorciada; 5: União estável; 9: Ignorado.
ESCMAE	Categórico	Escolaridade materna: 1: Nenhuma; 2: 1 – 3 anos; 3: 4 – 7 anos; 4: 8 – 11 anos ; 5: 12 a mais; 9: Ignorado.
CODOCUPMAE	Categórico	Código da ocupação materna, conforme o Código Brasileiro de Ocupações.
QTDFILVIVO	Numérico	Número de filhos vivos (9: ignorado).
QTDFILMORT	Numérico	Número de filhos mortos (perdas fetais ou abortos, 9: ignorado).
CODMUNRES	Categórico	Código IBGE ^a do município de residência.
GESTACAO	Categórico	Semanas de gestação da mãe. 1: Menos de 22 semanas; 2: 22 – 27 semanas; 3: 28 – 31 semanas; 4: 32 – 36 semanas; 5: 37 – 41 semanas; 6: 42 ou mais semanas; 9: Ignorado.

(a) IBGE: Instituto Brasileiro de Geografia e Estatística.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Quadro 10 – Descrição das variáveis do conjunto de dados do SINASC

(continua)

Variável	Tipo	Descrição
GRAVIDEZ	Categórico	Tipo de gravidez: 1: Única; 2: Dupla; 3: Tripla ou mais; 9: Ignorado.
PARTO	Categórico	Tipo de parto: 1: Vaginal; 2: Cesário; 9: Ignorado.
CONSULTAS	Categórico	Número de consultas de pré-natal: 1: Nenhuma; 2: 1 – 3 consultas; 3: 4 – 6 consultas; 4: 7 ou mais consultas; 9: Ignorado.
DTNASC	Numérico	Data de nascimento do recém-nascido: <i>dd mm aaaa</i>
HORANASC	Numérico	Hora do nascimento: <i>hh mm</i>
SEXO	Categórico	Sexo do recém-nascido: 1: Masculino; 2: Feminino; 9: Ignorado.
APGAR1	Numérico	<i>Apgar</i> do recém-nascido no 1º minuto, de 0 – 10
APGAR5	Numérico	<i>Apgar</i> do recém-nascido no 5º minuto, de 0 – 10
RACACOR	Categórico	Raça/cor do recém-nascido: 1: Branca; 2: Preta; 3: Amarela; 4: Parda; 5: Indígena.
PESO	Numérico	Peso do recém-nascido em gramas.
PARIDADE	Categórico	Define se é a primeira gravidez ou não. 0: Nulípara; 1: Multípara;
IDANOMAL	Categórico	Anomalia identificada no recém-nascido: 1: Sim; 2: Não; 9: Ignorado.
CODANOMAL	Categórico	Código CID-10 ^b da anomalia.
QTDGESTANT	Numérico	Número de gestações passadas.
QTDPARTNOR	Numérico	Número de partos vaginais passados.

(b) CID-10: Classificação Internacional de Doenças.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Quadro 10 – Descrição das variáveis do conjunto de dados do SINASC

(conclusão)

Variável	Tipo	Descrição
RACACORMAE	Categórico	Raça/cor da mãe: 1: Branca; 2: Preta; 3: Amarela; 4: Parda; 5: Indígena.
IDADEPAI	Numérico	Idade do pai do recém-nascido
DTULTMENST	Numérico	Data da última menstruação: <i>dd mm aaaa</i>
SEMAGESTAC	Numérico	Número de semanas de gestação.
CONSPRENAT	Numérico	Número de consultas de pré-natal.
QTDPARTCES	Numérico	Número de partos cesáreos passados.
MESPRENAT	Categórico	Mês de gestação da mãe em que iniciou o pré-natal.
TPAPRESENT	Categórico	Posição ou apresentação do recém-nascido: 1: Cefálico; 2: Pélvica ou podálica; 3: Transversa; 9: Ignorado.
STTRABPART	Categórico	Trabalho de parto induzido: 1: Sim; 2: Não; 3: Não aplica; 9: Ignorado.
STCESPARTO	Categórico	Cesárea ocorreu antes de iniciar o trabalho de parto: 1: Sim; 2: Não; 3: Não aplica; 9: Ignorado.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

ANEXO E – DESCRIÇÃO DAS VARIÁVEIS DO CONJUNTO DE DADOS DO SISTEMA DE INFORMAÇÕES SOBRE MORTALIDADE

Quadro 11 – Descrição das variáveis do conjunto de dados do SIM

(continua)

Variável	Tipo	Descrição
DTOBITO	Numérico	Data do óbito: <i>dd mm aaaa</i> .
HORAOBITO	Numérico	Hora do óbito: <i>hh : mm</i> (padrão 24 horas).
DTNASC	Numérico	Data de nascimento do óbito: <i>dd mm aaaa</i> .
IDADE	Numérico	Idade do óbito, com três dígitos referente a dois subcampos, consistindo: Primeiro subcampo (1 dígito): é a unidade da idade, 1 – minuto, 2 – hora, 3 – mês, 4 – ano, 5 – idade maior que 100 anos. Segundo subcampo (2 dígitos): é quantidade de unidades, 01 a 59 minutos – idade menor a 1 hora, 01 a 23 – idade entre 1 e 23 horas, 01 a 29 – idade entre 1 e 29 dias, 01 a 11 – idade entre 1 e 11 meses completos, 00 a 99 – idade em anos. 9: Ignorado.
SEXO	Catégorico	Sexo do óbito: 1: Masculino; 2: Feminino; 9: Ignorado.
CODMUNRES	Catégorico	Código IBGE ^a do município de residência.
LOCOCOR	Catégorico	Local de ocorrência do óbito 1: Hospital; 2: Outros estabelecimentos de saúde; 3: Domicílio; 4: Via pública; 5: Outros; 6: Aldeia indígena; 9: Ignorado.
CODMUNOCOR	Catégorico	Código do município em que ocorreu o óbito.
IDADEMAE	Numérico	Idade materna
ESMAE	Catégorico	Escolaridade materna: 1: Nenhuma; 2: 1 – 3 anos; 3: 4 – 7 anos; 4: 8 – 11 anos ; 5: 12 a mais; 9: Ignorado.
OCUPMAE	Catégorico	Código da ocupação materna, conforme o Código Brasileiro de Ocupações.

(a) IBGE: Instituto Brasileiro de Geografia e Estatística.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Quadro 35 – Descrição das variáveis do conjunto de dados do SIM

(continua)

Variável	Tipo	Descrição
QTDFILVIVO	Numérico	Número de filhos vivos (9: ignorado).
QTDFILMORT	Numérico	Número de filhos mortos (perdas fetais ou abortos, 9: ignorado).
GRAVIDEZ	Categórico	Tipo de gravidez: 1: Única; 2: Dupla; 3: Tripla ou mais; 9: Ignorado.
GESTACAO	Categórico	Semanas de gestação da mãe. 1: Menos de 22 semanas; 2: 22 – 27 semanas; 3: 28 – 31 semanas; 4: 32 – 36 semanas; 5: 37 – 41 semanas; 6: 42 ou mais semanas; 9: Ignorado.
PARTO	Categórico	Tipo de parto: 1: Vaginal; 2: Cesário; 9: Ignorado.
PESO	Numérico	Peso do óbito em gramas.
OBITOPUERP	Categórico	Óbito no puerpério: 1: Até 42 dias após o parto; 2: De 43 dias a 1 ano; 3: Não; 9: Ignorado.
CAUSABAS	Categórico	Causa básica do óbito, de acordo com os códigos CID-10 ^b .
CIRCOBITO	Categórico	Tipo de circunstâncias em que se deu o óbito: 1: Acidente; 2: Suicídio; 3: Homicídio; 4: Outros; 9: Ignorado.
TPOBITOCOR	Numérico	Sem descrição.

(b) CID-10: Classificação Internacional de Doenças.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

Quadro 35 – Descrição das variáveis do conjunto de dados do SIM

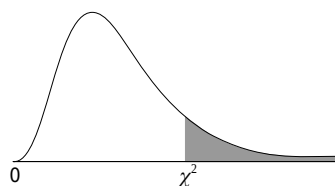
(conclusão)

Variável	Tipo	Descrição
TPMORTEOCO	Categórico	Quando a morte ocorreu. 1: Gravidez; 2: Parto; 3: No aborto; 4: 42 dias após o parto; 5: 43 dias a 1 ano após o parto; 8: Não ocorreu nestes períodos; 9: Ignorado.
MORTEPARTO	Numérico	Sem descrição.
CIRURGIA	Categórico	Indica se houve cirurgia. 1: Sim; 2: Não; 9: Ignorado.
LINHAA	String	Linha A do atestado, conforme a CID-10.
LINHAB	String	Linha B do atestado, conforme a CID-10.
LINHAC	String	Linha C do atestado, conforme a CID-10.
LINHAD	String	Linha D do atestado, conforme a CID-10.
LINHAI I	String	Linha II do atestado, conforme a CID-10.
ESMAE2010	Numérico	Escolaridade 2010. 0: Sem escolaridade;; 1: Fundamental I (1a a 4a série); 2: Fundamental II (5a a 8a série); 3: Médio (antigo 2do Grau); 4: Superior incompleto; 5: Superior completo; 9: Ignorado.
CODMUNNATU	Categórico	Código do município de naturalidade.
CODESTAB	Categórico	Código do estabelecimento.
ASSISTMED	Categórico	Indica se houve assistência medica. 1: Com assistência; 2: Sem assistência; 9: Ignorado.

(b) CID-10: Classificação Internacional de Doenças.

Fonte: Elaborado pela autora (2025), com dados de DATASUS (2023).

ANEXO F – TABELA DA DISTRIBUIÇÃO QUI-QUADRADO (χ^2)



Área à direita do valor crítico

Tabela 35 – Tabela da da distribuição Qui-quadrado (χ^2)

df	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$	$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Fonte: Elaborado pela autora (2025).

DADOS CURRICULARES

IDENTIFICACIÓN	
	ELEN YANINA AGUIRRE RODRÍGUEZ 17 de agosto de 1994
Nacionalidad	Peruana
Nombre en citas bibliográficas	Aguirre-Rodríguez, Elen Yanina; Aguirre-Rodríguez, E. Y.; Aguirre Rodríguez, Elen Yanina; Aguirre Rodríguez, E. Y.; Rodríguez, Elen Yanina Aguirre; Rodríguez, E. Y. A.
Currículo Lattes	http://lattes.cnpq.br/5084452020999999
ORCID	https://orcid.org/0000-0002-3829-4118
Otro identificador	https://www.scopus.com/authid/detail.uri?authorId=58853704600 https://ctivitaie.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=320722 https://scholar.google.com/citations?user=AXUby2cAAAAJ&hl=es
FORMACIÓN ACADÉMICA	
2011 / 2015	Curso de Grado en Ingeniería de Sistemas, Universidad Nacional de Trujillo
2018 / 2020	Maestría en Ingeniería de Producción, Universidad Estatal Paulista “Júlio de Mesquita Filho”
PRODUCCIÓN BIBLIOGRÁFICA EM PERIÓDICOS	
AGUIRRE-RODRÍGUEZ, E. Y.; GAMBOA, A. A. R.; RODRÍGUEZ, E. C. A.; SANTOS-FERNÁNDEZ, J. P.; NASCIMENTO, L. F. C.; DA SILVA, A. F.; MARINS, F. A. S. Rice leaf disease detection using the Stretched Neighborhood Effect Color to Grayscale method and Machine Learning. Scientia Agropecuaria , v.16, p.123-136, 2025.	
AGUIRRE-RODRÍGUEZ, E. Y.; RODRÍGUEZ, E. C. A.; DA SILVA, A. F.; RIZOL, P. M. S. R.; MIRANDA, R. C.; MARINS, F. A. S. A decision-making framework with machine learning for transport outsourcing based on cost prediction: an application in a multinational automotive company. International Journal of Information Technology , v.16, p.1495-1503, 2024.	
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; DA SILVA, A. F.; RIZOL, P. M. S. R.; MARINS, F. A. S.. Analysis of machine learning integration into supply chain management. International Journal of Logistics Systems and Management (Print), v.47, p.327-355, 2024.	
SILVA, A. J.; RODRÍGUEZ, E. Y. A.; TRAMARICO, C. L.; MARINS, F. A. S.. Revisão de processos na logística interna em uma empresa do segmento automotivo. Revista Produção Online , v.23, p.4968, 2024.	
RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; MARINS, F. A. S.; DA SILVA, A. F.; NASCIMENTO, L. F. C.. Spatial patterns of prematurity and its determinants in the metropolitan region of São Paulo, Brazil, 2010-2019. Revista Brasileira de Epidemiologia , v.27, p.e240008, 2024.	
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; MARINS, F. A. S.; SILVA, A. F.; NASCIMENTO, L. F. C.. Spatial patterns of mortality in low birth weight infants at term and its determinants in the State of São Paulo, Brazil. Revista Brasileira de Epidemiologia , v.26, p.1-10, 2023.	

RODRÍGUEZ, E. Y. A.; GAMBOA, A. A. R.; RODRÍGUEZ, E. C. A.; SILVA, A. F.; RIZOL, Paloma M. S. R.; MARINS, F. A. S.. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and machine learning algorithms for electricity production forecasting. **IEEE Latin America Transactions**, v.20, p.2288-2294, 2022.

AGUIRRE-RODRÍGUEZ, E. Y.; SANTOS-FERNÁNDEZ, J. P.; RODRÍGUEZ, E. C. A.; MARINS, F. A. S.. Spatial analysis of crime and identification of clusters in the province of Trujillo. **Revista Ciencia y Tecnología**, v.17, p.53-61, 2021.

DEFALQUE, C. M.; MARINS, F. A. S.; DA SILVA, A. F.; RODRÍGUEZ, E. Y. A.. A review of waste paper recycling networks focusing on quantitative methods and sustainability. **Journal of Material Cycles and Waste Management**, v.01, p.1-22, 2020.

PRODUCCIÓN BIBLIOGRÁFICA EM EVENTOS

LUZ, C. A. N.; MARINS, F. A. S.; RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A. Análise Bibliométrica do Impacto das Tecnologias Blockchain, Big Data e Cloud Computing no Gerenciamento da Cadeia de Suprimentos. In: Simpósio de Excelência em Gestão e Tecnologia, 2024, Resende - RJ. **Anais do XXI SEGeT**, v.21, 2024.

RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; GAMBOA, A. A. R.; XIMENES, E. D.; QUEIROZ, A. R.; SILVA, A. F.; MARINS, F. Bibliometric analysis on the optimization of offshore renewable energy portfolios using Mixture Design and Monte Carlo Simulation. In: LVI Simpósio Brasileiro de Pesquisa Operacional, 2024, Fortaleza - CE. **Anais do LVI Simpósio Brasileiro de Pesquisa Operacional**, 2024.

AGUIRRE-RODRIGUEZ, E. Y.; RODRIGUEZ, E. C. A.; SANTOS-FERNÁNDEZ, J. P.; NASCIMENTO, L. F. C.; SILVA, A. F. S.; MARINS, F.. Identificação de doenças do arroz usando aprendizado de máquina. In: Anais do Simpósio Brasileiro de Pesquisa Operacional, 2023. **Anais do LVI Simpósio Brasileiro de Pesquisa Operacional**, 2023.

RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. A data-driven approach for neonatal mortality rate forecasting. In: 5th International Conference on Informatics & Data-Driven Medicine, 2022, Lyon. **Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine**. CEUR Workshop Proceedings, v.3302, p.86-98, 2022.

RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; SANTOS-FERNÁNDEZ, J. P.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, Fernando Augusto Silva. Inteligencia artificial para detección y diagnóstico de enfermedades en cultivos de arroz. In: Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología, 2022, Estado de México. **Memorias del Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología**, v.4, p.10-18, 2022.

RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Machine learning techniques to predict overweight or obesity. In: 4th International Conference on Informatics and Data-Driven Medicine, 2021, Valencia. **Proceedings of the 4th International Conference on Informatics & Data-Driven Medicine**. CEUR Workshop Proceedings, v.3038, p.190-204, 2021.

RODRÍGUEZ, E. Y. A.; GAMBOA, A. A. R.; RODRÍGUEZ, E. C. A.; RIZOL, P. M. S. R.; SILVA, A. F.; MARINS, F. A. S. Machine Learning Techniques and Neuro-fuzzy applied to the electricity production. In: 13th Latin-American Congress on Electricity Generation and Transmission - CLAGTEE, 2019, Santiago de Chile. Proceedings of the 13th Latin-American Congress on Electricity Generation and Transmission - CLAGTEE , 2019.
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; SILVA, A. F.; MARINS, F. A. S. Previsão do custo de transporte por meio do método de árvore de decisão de aprendizagem de máquina. In: XXVI SIMPEP - Simpósio de Engenharia de Produção, 2019, Bauru. Simpósio de Engenharia de Produção: Desafios da Engenharia de Produção no Contexto da Indústria 4.0 , 2019.
MARINS, F. A. S.; RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; CAVA, D. E. Q. S.; NASCIMENTO, L. F. C.; SILVA, A.; QUEIROZ, A. R. Data envelopment analysis with machine learning in healthcare efficiency: a prematurity study. In: 33rd European Conference on Operational Research (EURO XXXIII), 2024, Copenhagen, Denmark. EURO 2024 Conference Handbook & Abstracts , 2024.
CAVA, D. E. Q. S.; MARINS, Fernando; RODRÍGUEZ, E. Y. A.; SILVA, A.; RODRÍGUEZ, E. C. A. Spatial analysis techniques for diagnosis situational of low birth weight at term in the state of São Paulo. In: 33rd European Conference on Operational Research (EURO XXXIII), 2024, Copenhagen, Denmark. EURO 2024 Conference Handbook & Abstracts , 2024.
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Análises espacial do baixo peso ao nascer no Vale do Paraíba Paulista. In: 11º Congresso Brasileiro de Epidemiologia, 2021, Fortaleza. Anais do 11º Congresso Brasileiro de Epidemiologia , v.2, 2021.
RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Análise envoltória de dados com aprendizado de máquina na predição da eficiência do sistema de saúde com base na prematuridade. In: VII Workshop da Pós-graduação em Engenharia - UNESP, 2023, Guaratinguetá. Anais do VII Workshop da Pós-graduação em Engenharia - UNESP , 2023.
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Random Forest para diagnóstico de doenças em folhas de arroz. In: VII Workshop da Pós-graduação em Engenharia - UNESP, 2023, Guaratinguetá. Anais do VII Workshop da Pós-graduação em Engenharia - UNESP , 2023.
RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, Fernando Augusto Silva. Métodos de ensemble com árvores de decisão para predição da idade gestacional. In: VI Workshop da Pós-Graduação em Engenharia Mecânica e de Produção, 2022, Guaratinguetá. Anais do VI Workshop da Pós-Graduação em Engenharia Mecânica e de Produção , v.1, p.340-345, 2022.
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Support vector machine aplicado à previsão de taxas de mortalidade neonatal. In: VI Workshop da Pós-Graduação em Engenharia Mecânica e de Produção, 2022, Guaratinguetá. Anais do VI Workshop da Pós-Graduação em Engenharia Mecânica e de Produção , v.1, p.361-366, 2022.
RODRÍGUEZ, E. Y. A.; NASCIMENTO, L. F. C.; DA SILVA, A. F.; MARINS, F. A. S. Análise espacial de áreas de risco de mortalidade em nascidos vivos com baixo peso ao nascer, São Paulo, 2010-2018. In: V Workshop da Pós-Graduação em Engenharia Mecânica e de Produção, 2021, Guaratinguetá. Anais do V Workshop da Pós-Graduação em Engenharia Mecânica e de Produção , v.5, 2021.

RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; NASCIMENTO, L. F. C.; SILVA, A. F.; MARINS, F. A. S. Mapeando o uso do aprendizado de máquina nos dados de baixo peso ao nascer. In: IV Workshop da Pós-Graduação em Engenharia Mecânica e Produção UNESP, 2020, Guaratinguetá. Anais do IV Workshop da Pós-Graduação em Engenharia Mecânica e Produção UNESP , 2020.	
RODRÍGUEZ, E. Y. A.; RODRÍGUEZ, E. C. A.; SILVA, A. F.; MARINS, F. A. S. Sistema de inferência neuro-fuzzy adaptativo para prever a produção de eletricidade de uma usina de ciclo combinado. In: III Workshop dos Programas de Pós-Graduação em Engenharia Mecânica e em Engenharia de Produção, 2019, Guaratinguetá - SP. Anais do 3º Workshop de Engenharia Mecânica e de Produção: energia, materiais, projetos, gestão e otimização , 2019.	
RODRÍGUEZ, E. Y. A.; SILVA, A. F.; LUCHE, J. R. D.; MARINS, F. A. S. Aplicação de aprendizado de máquina para previsão de custos logísticos de transporte. In: II Workshop da Pós-Graduação em Engenharia Mecânica e Engenharia de Produção da UNESP, 2018, Guaratinguetá. Anais do II Workshop da Pós-Graduação em Engenharia Mecânica e Engenharia de Produção da UNESP , 2018.	
PARTICIPACIÓN EN BANCAS	
MARINS, F.; RODRÍGUEZ, E. C. A.; RODRÍGUEZ, E. Y. A.. Participação em banca de Cesar Augusto Nobrega da Luz. Participação em banca de César Augusto Nóbrega da Luz. Análise Bibliométrica do Impacto das Tecnologias Blockchain, Big Data e Cloud Computing no Gerenciamento da Cadeia de Suprimentos , 2024. (Engenharia de Produção) Universidade Estadual Paulista Júlio de Mesquita Filho, Guaratinguetá.	
MARINS, F.; RODRÍGUEZ, E. C. A.; RODRIGUEZ, E. Y. A.. Participação em banca de Diego Eduardo Quagliato Scarelli Cava. Participação em banca de Diego Eduardo Quagliato Scarelli Cava. Técnicas de análise espacial para a exploração do baixo peso ao nascer no estado de São Paulo , 2024. (Engenharia de Produção) Universidade Estadual Paulista Júlio de Mesquita Filho, Guaratinguetá.	
REGISTRO DE PROGRAMA DE COMPUTADOR	
RODRÍGUEZ, E. Y. A.; SILVA, A. F.; MARINS, F. A. S. Sistema de apoio à decisão com aprendizagem de máquina para avaliação de custos na terceirização de transporte. Registro de programa de computador . Instituto Nacional da Propriedade Industrial (INPI), BR512020002828-4, 14 dez. 2020. Concessão em 22 dez. 2020. Instituições financiadoras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Finalidade: ferramenta de apoio à decisão.	
PRÊMIOS E DISTINÇÕES	
2023	Primeiro lugar na categoria Doutorado , no VII Workshop de Pós-Graduação em Engenharia Mecânica e Produção - Área de Projetos, Gestão e Otimização, Pró-reitoria de Pós-Graduação da UNESP .
2022	Primeiro lugar na categoria Doutorado , no VI Workshop de Pós-Graduação em Engenharia Mecânica e Produção - Área de Projetos, Gestão e Otimização, Pró-reitoria de Pós-Graduação da UNESP .
2020	Primeiro lugar na categoria Doutorado , no IV Workshop de Pós-Graduação em Engenharia Mecânica e Produção - Área de Projetos, Gestão e Otimização, Pró-reitoria de Pós-Graduação da UNESP .

2018	Primeiro lugar na categoria Mestrado , no 2º Workshop da Pós-Graduação em Engenharia Mecânica e Engenharia de Produção - FEG UNESP.
-------------	--