



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Faculdade de Engenharia e Ciências de Guaratinguetá

ELIAS CARLOS AGUIRRE RODRÍGUEZ

**Aplicação de técnicas de aprendizado de máquina para a identificação de mães com
risco de parto prematuro na cidade de São Paulo**

Guaratinguetá - SP
2023

Elias Carlos Aguirre Rodríguez

**Aplicação de técnicas de aprendizado de máquina para a identificação de mães com
risco de parto prematuro na cidade de São Paulo**

Dissertação apresentada à Faculdade de Engenharia e Ciências do Campus de Guaratinguetá, Universidade Estadual Paulista, para obtenção do título de Mestre em Engenharia, na área de Produção.

Orientador:

Prof. Dr. Fernando Augusto Silva Marins

Co-orientadores:

Prof. Dr. Luiz Fernando Costa Nascimento

Prof. Dr. Aneirson Francisco da Silva

Guaratinguetá - SP
2023

A284a	Aguirre Rodriguez, Elias Carlos Aplicação de técnicas de aprendizado de máquina para a identificação de mães com risco de parto prematuro na cidade de São Paulo / Elias Carlos Aguirre Rodriguez – Guaratinguetá, 2024. 71 : il. Bibliografia: f. 65-71 Dissertação – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2024. Orientador: Prof. Dr. Fernando Augusto Silva Marins Coorientadores: Prof. Dr. Luiz Fernando Costa Nascimento e Prof. Dr. Aneirson Francisco da Silva 1. Prematuros. 2. Aprendizado do computador. 3. Processo decisório. I. Título. CDU 614(043)
-------	--

Luciana Máximo
Bibliotecária CRB-8/3595

IMPACTO POTENCIAL DESTA PESQUISA

Este estudo apoia a tomada de decisões na área de saúde e fornece subsídios para gestores municipais, regionais e nacionais implementarem medidas e avaliarem políticas de saúde pública que possam controlar e reduzir nascimentos prematuros, além de diminuir a mortalidade neonatal, visando alcançar as metas do ODS 3 sobre Saúde e Bem-Estar até 2030.

POTENTIAL IMPACT OF THIS RESEARCH

This study supports healthcare decision-making and provides insights for municipal, regional, and national managers to implement measures and evaluate public health policies that can control and reduce premature births, as well as decrease neonatal mortality, aiming to achieve the targets of SDG 3 on Health and Well-being by 2030.

ELIAS CARLOS AGUIRRE RODRÍGUEZ

**ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
“MESTRE EM ENGENHARIA”**

**PROGRAMA: ENGENHARIA
CURSO: MESTRADO**

APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO

 Documento assinado digitalmente
MANOEL CLEBER DE SAMPAIO ALVES
Data: 14/11/2023 09:31:16-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Manoel Cleber de Sampaio Alves
Coordenador

BANCA EXAMINADORA:

 Documento assinado digitalmente
FERNANDO AUGUSTO SILVA MARINS
Data: 09/11/2023 11:14:30-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. FERNANDO AUGUSTO SILVA MARINS
Orientador - UNESP

 Documento assinado digitalmente
ALEXANDRE DIAS PORTO CHIAVEGATTO FILHO
Data: 09/11/2023 17:54:07-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. ALEXANDRE DIAS PORTO CHIAVEGATTO FILHO
USP

 Documento assinado digitalmente
MARCELA APARECIDA GUERREIRO MACHADO DE
Data: 10/11/2023 11:47:39-0300
Verifique em <https://validar.iti.gov.br>

**Prof^a. Dr^a. MARCELA APARECIDA GUERREIRO MACHADO DE
FREITAS**
UNESP

DADOS CURRICULARES

ELIAS CARLOS AGUIRRE RODRÍGUEZ

NASCIMENTO 20 de fevereiro de 1997 - Chepén / Perú

FILIAÇÃO Elias Aguirre Tejada
Edita Rodríguez Ninaquispe

2015 / 2019 Curso de Graduação em Estatística
Bacharelado em Ciências Estatísticas
Universidad Nacional de Trujillo - Perú

Dedicado aos meus pais Elias e Edita,
meus irmãos Richard, Cristina, Elen e Christian,
meus sobrinhos Mattias, Jassiel, Betsabé e Caleb, e
meus avós Juan e Sofia (*in memoriam*)

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, pelo dom da vida e sabedoria.

Aos meus pais Elias e Edita, pelo amor e apoio incondicional ao longo da minha vida e da minha formação profissional, e também por me ensinar que apesar das dificuldades que possam se apresentar, sempre tenho que lutar para cumprir os meus sonhos.

Aos meus irmãos, pela compreensão, motivação e conselhos, os quais me ajudam a melhorar cada dia como pessoa e profissional.

Ao meu orientador, o Prof. Dr. Fernando Augusto Silva Marins, pela oportunidade de trabalhar neste período do mestrado, pela confiança e paciência para comigo, e também pelo apoio, motivação e ensinamento constante, que permitiu culminar com o desenvolvimento deste trabalho.

Ao meu coorientador, o Prof. Dr. Luiz Fernando Costa Nascimento, pela confiança e paciência para comigo, além do apoio, ensinamento, e compartilhamento de conhecimentos, que ajudaram muito no desenvolvimento deste trabalho.

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

À Profa. Dra. Marcela Aparecida Guerreiro Machado e ao Prof. Dr. Alexandre Dias Porto Chiavegatto Filho, pela participação e contribuição como membros da banca examinadora.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES¹), pelo apoio financeiro para o desenvolvimento deste trabalho.

¹ <http://www.capes.gov.br>

Este trabalho contou com o apoio da seguinte entidade:

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

“Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer.”
(Albert Einstein)

RESUMO

O parto prematuro ocorre antes da 37ª semana de gestação, apresentando um risco elevado de complicações físicas e neuropsicomotoras nos recém-nascidos, sendo considerado a principal causa de mortalidade neonatal. Em consonância com um dos Objetivos de Desenvolvimento Sustentável (ODS) voltados à Saúde e Bem-Estar, que visa reduzir a mortalidade neonatal até 2030, a aplicação de ferramentas tecnológicas, como o Aprendizado de Máquina, pode oferecer suporte na assistência médica e na tomada de decisões. Dessa forma, esta pesquisa tem como objetivo aplicar técnicas de aprendizado de máquina para identificar mães com risco de parto prematuro, utilizando dados do município de São Paulo no período de 2015 a 2019. Este trabalho pode ser classificado como uma pesquisa aplicada com objetivos empírico-normativos, sua abordagem é quantitativa, utilizando o método de pesquisa de modelagem e simulação. Foram utilizados dados do Departamento de Informática do Sistema Único de Saúde – DATASUS, especificamente registros da declaração de nascidos vivos. A análise baseou-se na abordagem de regressão para estimar a duração da gestação (idade gestacional), considerando informações maternas e da gestação atual ou anterior e utilizando métodos como Regressão Linear, Árvore de Decisão, Floresta Aleatória, *Gradient Boosting*, e *Extreme Gradient Boosting*. Os modelos foram validados internamente com validação cruzada de *10-Fold* e foi utilizada técnica *Random Search* para encontrar a melhor combinação de hiperparâmetros. Como resultado, o modelo de *Gradient Boosting* destacou-se como o melhor, apresentando bons resultados nas diferentes métricas de desempenho, com valores de EMA = 1,186, EQM = 2,833, REQM = 1,683, e EPAM = 3,182. Concluindo, o modelo de regressão demonstrou bom desempenho, evidenciando que as técnicas de aprendizado de máquina são ferramentas eficazes na área da saúde. Elas permitem que os especialistas tomem decisões oportunas em relação à identificação de nascimentos prematuros, contribuindo para a implementação de medidas de controle e, conseqüentemente, para a redução das taxas de mortalidade neonatal, em conformidade com os ODS.

PALAVRAS-CHAVE: Aprendizado de Máquina; Nascimento prematuro; Assistência Perinatal; Saúde materno-infantil; Tomada de Decisões.

ABSTRACT

The premature birth occurs before the 37th week of gestation, presenting a high risk of physical and neuropsychomotor complications in newborns, being considered the main cause of neonatal mortality. In line with one of the Sustainable Development Goals (SDGs) related to Health and Well-being, aiming to reduce neonatal mortality by 2030, the application of technological tools, such as Machine Learning, can provide support in medical care and decision-making. Thus, this research aims to apply machine learning techniques to identify mothers at risk of premature birth, using data from the municipality of São Paulo from 2015 to 2019. This work can be classified as applied research with empirical-normative objectives, its approach is quantitative, using the research method of modeling and simulation. Data from the Department of Informatics of the Unified Health System - DATASUS were used, specifically records of live births. The analysis was based on the regression approach to estimate the duration of gestation (gestational age), considering maternal information and current or previous pregnancy and using methods such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting. The models were internally validated with 10-Fold cross-validation, and the Random Search technique was used to find the best combination of hyperparameters. As a result, the Gradient Boosting model stood out as the best, showing good results in different performance metrics, with values of $MAE = 1.186$, $MSE = 2.833$, $RMSE = 1.683$, and $MAPE = 3.182$. In conclusion, the regression model demonstrated good performance, highlighting that machine learning techniques are effective tools in the field of health. They allow specialists to make timely decisions regarding the identification of premature births, contributing to the implementation of control measures and, consequently, the reduction of neonatal mortality rates, in accordance with the SDGs.

KEYWORDS: Machine Learning; Premature Birth; Perinatal Care; Maternal and Child Health; Decision Making.

LISTA DE FIGURAS

Figura 1 – Prevalência e taxa de mortalidade infantil (TMI) em recém-nascidos prematuros.....	19
Figura 2 – Número de publicações e citações por ano.....	23
Figura 3 – Número de publicações por país do autor correspondente.....	23
Figura 4 – Número de citações por país do autor correspondente.....	24
Figura 5 – Rede de colaboração internacional.....	25
Figura 6 – Gráfico de densidade das palavras-chave.....	26
Figura 7 – Número de publicações e citações por ano das palavras-chave (“ <i>Gestational Age</i> ” OR “ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ”OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ” AND “ <i>Health Records</i> ” OR “ <i>Medical Records</i> ”	26
Figura 8 – Número de publicações e citações por ano das palavras-chave (“ <i>Gestational Age</i> ” OR “ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ”OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ” AND “ <i>Decision making</i> ”	27
Figura 9 – Classificação das técnicas de AM.....	31
Figura 10 – Procedimento do Aprendizado Supervisionado.....	32
Figura 11 – Regressão.....	32
Figura 12 – Classificação.....	33
Figura 13 – Procedimento do Aprendizado Não Supervisionado.....	38
Figura 14 – <i>Clustering</i>	39
Figura 15 – Redução de Dimensionalidade.....	39
Figura 16 – Procedimento do Aprendizado por Reforço.....	40
Figura 17 – Procedimento do Aprendizado Profundo.....	40
Figura 18 – Classificação da Pesquisa Científica.....	41
Figura 19 – Mapa da localização geográfica do município de São Paulo.....	42
Figura 20 – Fluxo Metodológico.....	44
Figura 21 – Porcentagem de dados ausentes por variável.....	47
Figura 22 – Variáveis categóricas ordinais.....	50
Figura 23 – Variável nominal com mais de duas categorias.....	50
Figura 24 – Variáveis nominais de duas categorias.....	51
Figura 25 – Correlações entre variáveis de entrada quantitativas.....	53
Figura 26 – Correlações entre variáveis de entrada, excluindo QTDFILVIVO e	

QTDFILVIVO.....	54
Figura 27 – Valores das métricas de desempenho dos modelos de regressão.....	55
Figura 28 – Importância das variáveis no modelo de Gradient Boosting selecionado para regressão.....	56
Figura 29 – Interface de usuário desenvolvida para estimar a idade gestacional.....	57
Figura 30 – Estimação da idade gestacional - Caso 1.....	57
Figura 31 – Estimação da idade gestacional - Caso 2.....	58

LISTA DE QUADROS

Quadro 1 – Estudos relacionados com a abordagem proposta.....	28
---	----

LISTA DE TABELAS

Tabela 1 – Publicações em SCOPUS e PubMed.....	21
Tabela 2 – Informações principais do conjunto final de publicações.....	22
Tabela 3 – Descrição do conjunto de variáveis sobre as informações maternas e da gestação atual ou anterior.....	43
Tabela 4 – Tipo de variável e valores atribuídos.....	48
Tabela 5 – Estatísticas descritivas das variáveis numéricas.....	49
Tabela 6 – Características sobre os nascidos vivos pré-termo e à termo (n = 794.696)....	51
Tabela 7 – Otimização de hiperparâmetros dos modelos de regressão.....	55
Tabela 8 – Valores das métricas de desempenho dos modelos de regressão.....	55

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
ANS	Aprendizado Não Supervisionado
AP	Aprendizado Profundo
AR	Aprendizado por Reforço
AS	Aprendizado Supervisionado
BD	<i>Big Data</i>
DATASUS	Departamento de Informática do Sistema Único de Saúde
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ODM	Objetivos de Desenvolvimento do Milênio
ODS	Objetivos de Desenvolvimento Sustentável
OMS	Organização Mundial da Saúde
SINASC	Sistema de Informações Sobre Nascidos Vivos
UNESP	Universidade Estadual Paulista

SUMÁRIO

1	INTRODUÇÃO.....	18
1.1	CONTEXTUALIZAÇÃO.....	18
1.2	QUESTÕES DE PESQUISA.....	20
1.3	OBJETIVOS.....	20
1.3.1	Objetivo Geral.....	20
1.3.2	Objetivos Específicos.....	21
1.4	JUSTIFICATIVA E CONTRIBUIÇÃO CIENTÍFICA.....	21
1.4.1	Aprendizado de máquina na predição da idade gestacional para a identificação de recém-nascidos prematuros.....	27
1.5	DELIMITAÇÃO DA PESQUISA.....	29
1.6	ESTRUTURA DA PESQUISA.....	29
2	FUNDAMENTAÇÃO TEÓRICA.....	30
2.1	PREMATURIDADE.....	30
2.2	APRENDIZADO DE MÁQUINA.....	30
2.2.1	Aprendizado Supervisionado.....	31
2.2.1.1	Algumas técnicas de AM.....	33
2.2.1.2	Validação cruzada e métricas de desempenho.....	36
2.2.2	Aprendizado Não Supervisionado.....	38
2.2.3	Aprendizado por Reforço.....	39
2.2.4	Aprendizado Profundo.....	40
3	MATERIAIS E MÉTODOS.....	41
3.1	MATERIAIS.....	41
3.2	MÉTODOS E PROCEDIMENTOS.....	44
3.2.1	Pré-processamento e experimentação.....	45
3.2.1.1	Limpeza de dados.....	45
3.2.1.2	Transformação e escalonamento de dados.....	46
3.3	ASPECTOS ÉTICOS.....	46
4	RESULTADOS.....	47
4.1	ANÁLISE EXPLORATÓRIA DE DADOS E SELEÇÃO DE VARIÁVEIS.....	47
4.2	PRE-PROCESSAMENTO, EXPERIMENTAÇÃO E AVALIAÇÃO.....	53
4.2.1	Abordagem de regressão para estimar a idade gestacional.....	54

4.3	IMPLEMENTAÇÃO DO MODELO SELECIONADO.....	56
5	DISCUSSÕES.....	59
6	CONCLUSÕES.....	62
6.1	SUGESTÕES PARA FUTURAS PESQUISAS.....	63
	REFERÊNCIAS.....	65

1 INTRODUÇÃO

Neste capítulo, abordaremos a contextualização do estudo, as questões de pesquisa, os objetivos propostos, a justificativa e a contribuição científica do trabalho, além da delimitação do escopo da pesquisa. A seguir, apresentaremos a estrutura detalhada deste estudo.

1.1 CONTEXTUALIZAÇÃO

O parto prematuro ocorre antes da 37ª semana de gestação e está associado a um maior risco de complicações físicas e neuropsicomotoras no recém-nascido (RODRÍGUEZ *et al.*, 2024). Além disso, é uma das principais causas de morte em crianças menores de 5 anos (BERGER *et al.*, 2016; ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2022) e, basicamente, essa condição está fortemente relacionada com a mortalidade neonatal (ADRIANO *et al.*, 2022; RODRÍGUEZ *et al.*, 2024).

Alguns dos fatores de risco mais comuns associados ao parto prematuro estão relacionados às informações maternas, como condições de saúde, hábitos alimentares e estilos de vida, além de dados sobre gestações atuais ou anteriores (LEONE, 2015). De acordo com a Organização Mundial da Saúde - OMS (2022), estima-se que cada ano mais de 1 em cada 10 recém-nascidos nasça antes de completar 37 semanas de gestação, e, aproximadamente 1,1 milhão de recém-nascidos morrem devido a complicações associadas ao nascimento prematuro, tais como síndromes de problemas respiratórios, doenças pulmonares crônicas, lesões nos intestinos, doenças cardiovasculares, problemas auditivos e visuais, além de lesões neurológicas (BEHRMAN; BUTLER, 2007).

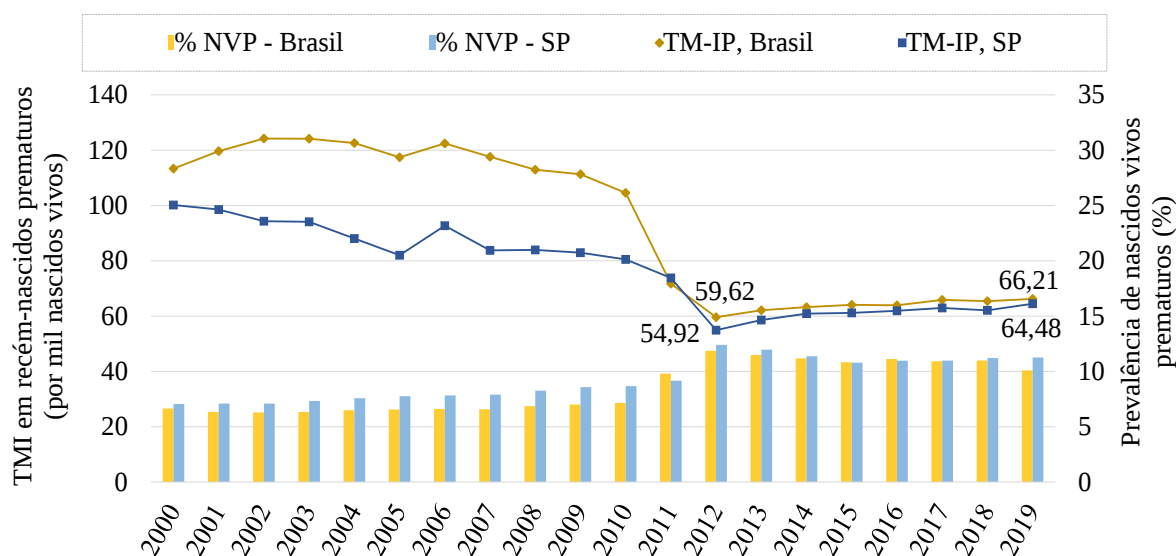
Dada a relação significativa entre prematuridade e mortalidade, a redução da mortalidade infantil tornou-se uma prioridade global nos últimos anos. Para isso, diversos líderes mundiais se comprometeram a reduzir a mortalidade infantil por meio do estabelecimento de objetivos e metas. Nos Objetivos de Desenvolvimento do Milênio (ODM), os estados se comprometeram a reduzir em dois terços a mortalidade de menores de 5 anos até 2015 (UNITED NATIONS, 2015a). Apesar dos esforços globais, a taxa de mortalidade infantil diminuiu em mais da metade, passando de 90 para 43 mortes por 1.000 nascidos vivos entre 1990 e 2015 (UNITED NATIONS, 2015a). No entanto, os resultados obtidos não foram suficientes para atingir a meta estabelecida nos ODM.

Após os ODM, os Objetivos de Desenvolvimento Sustentável (ODS), implementados em 2016, assumiram o compromisso de alcançar 17 objetivos e 169 metas até 2030. Uma das metas do objetivo 3 dos ODS, sobre Saúde e Bem-Estar, está focado em reduzir a mortalidade neonatal para pelo menos 12 por 1.000 nascidos vivos e a mortalidade de crianças menores de 5 anos para pelo menos 25 por 1.000 nascidos vivos (UNITED NATIONS, 2015b).

Nesse contexto, de acordo com a OMS (2022), o Brasil ocupa a 10ª posição entre os países

com o maior número de nascimentos prematuros no mundo, a 2ª posição no continente americano, depois de Estados Unidos, e a 1ª posição na América Latina. Como mostrado na Figura 1, as taxas de mortalidade infantil em recém-nascidos prematuros do Brasil e do estado de São Paulo tiveram um decréscimo até 2012, e de 2013 em diante tem aumentado. Da mesma forma, a prevalências de recém-nascidos vivos antes das 37ª semanas de gestação teve um aumento de 2012 em diante, sendo um aumento de 21% para a prevalência no Brasil e 35% no estado de São Paulo.

Figura 1 – Prevalência e taxa de mortalidade infantil (TMI) em recém-nascidos prematuros



Fonte: DATASUS, 2022

Além disso, em conformidade com a Rede Nacional Primeira Infância (2019), o parto prematuro no Brasil gera um custo anual aproximado de 8 bilhões de reais e, depois do parto, os recém-nascidos permanecem em média 51 dias internados em cuidados intensivos, o que gera um custo de mais de 15 bilhões de reais por ano. Portanto, é de extrema importância aprimorar a assistência hospitalar para mães e recém-nascidos prematuros.

Nessa sequência, algumas ferramentas tecnológicas como o Aprendizado de Máquina (AM) foram integradas à área da saúde nos últimos anos, para oferecer suporte na assistência médica e na tomada de decisões (VOLLMER *et al.*, 2020). Essas ferramentas surgiram graças ao avanço tecnológico nos processos da área de saúde, permitindo o armazenamento digital de uma enorme quantidade de dados para a extração posterior de informações valiosas (DOUPE; FAGHMOUS; BASU, 2019). Dados brutos, por si só, não geram informações relevantes, e seu processamento pode aprimorar a qualidade e a eficiência da prestação de cuidados de saúde e da tomada de decisão, além do fato de que os dados podem ser analisados para ampliar a capacidade de gerar novos conhecimentos, bem como previsões (OBERMEYER; EMANUEL, 2016).

Deste modo, o AM pode ajudar a melhorar a capacidade para tomar decisões dos profissionais com base no estabelecimento de previsões (OBERMEYER; EMANUEL, 2016), sendo um dos principais ramos da Inteligência Artificial (IA) aplicado para a detecção automatizada de

padrões significativos em dados, com o objetivo de construir modelos preditivos (ZHOU *et al.*, 2017).

Na literatura, alguns estudos integraram o Aprendizado de Máquina com dados de prematuridade para estimar o risco de parto prematuro. Por exemplo, analisou-se a saúde materna de gestantes que vivem em bairros violentos e sua relação com o parto prematuro (GOIN *et al.*, 2020)). Bodnar *et al.* (2020) analisaram o consumo alimentar de frutas e hortaliças em mães gestantes em relação ao consumo energético total e seus resultados quanto ao risco de parto prematuro. Em outras pesquisas tentaram-se analisar os fatores de risco para a incidência de parto prematuro e testaram diferentes algoritmos de AM (LEE; AHN, 2019; GAO *et al.*, 2019; KHATIBI; KHEYRIKOOCHAKSARAYEE; SEPEHRI, 2019; RITTENHOUSE *et al.*, 2019; LEE *et al.*, 2020; HAMILTON *et al.*, 2020).

No contexto brasileiro, de acordo com Santos (2018), a aplicação de técnicas de AM na área de saúde ainda é incipiente, evidenciando uma oportunidade de estudo para a utilização e aplicação desses métodos e algoritmos nos diversos campos da saúde.

Neste cenário, é proposto o desenvolvimento de um modelo preditivo baseado em técnicas de AM, para a identificação de mães com risco de parto prematuro, tomando como objeto de estudo os dados da cidade de São Paulo. Assim, o estudo proposto pode ajudar na tomada de decisões, e até o próprio governo para implementar e avaliar novas políticas de saúde pública, para o uso de ferramentas inteligentes como suporte para ações preventivas.

1.2 QUESTÕES DE PESQUISA

- Como desenvolver um modelo preditivo para a tomada de decisão com Aprendizado de Máquina para a identificação de mães com risco de parto prematuro?
- Quais dados são relevantes para este estudo e como devem ser analisados e tratados?
- Quais variáveis de entrada devem ser consideradas no modelo preditivo?
- Quais algoritmos de Aprendizado de Máquina Supervisionado são adequados para o desenvolvimento do modelo preditivo?
- Qual seria o impacto da integração de um sistema de tomada de decisão na avaliação do risco de uma mãe ter parto prematuro?
- Como validar os resultados obtidos por meio desta abordagem?

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Aplicar técnicas de Aprendizado de Máquina para a identificação de mães com risco de parto prematuro, tomando como objeto de estudo os dados da cidade de São Paulo.

1.3.2 Objetivos Específicos

- Identificar e selecionar as variáveis mais relevantes para o modelo.
- Selecionar um algoritmo de Aprendizado de Máquina Supervisionado para o desenvolvimento do modelo preditivo.
- Aplicar e analisar o impacto do modelo criado na tomada de decisão para avaliação do risco de uma mãe ter parto prematuro.
- Desenvolver uma interface de usuário.
- Validar os resultados obtidos da abordagem.

1.4 JUSTIFICATIVA E CONTRIBUIÇÃO CIENTÍFICA

O AM tem sido incorporado em diferentes áreas de estudo, proporcionando vantagens significativas, seja para a tomada de decisão, predição, classificação, entre outros. Na área de saúde, o AM representa uma ferramenta valiosa para especialistas, sendo útil tanto na tomada de decisão quanto no cuidado dos pacientes (CHAR; SHAH; MAGNUS, 2018; NGIAM; KHOR, 2019).

Para destacar a contribuição desta pesquisa, foi realizado um estudo bibliométrico descritivo retrospectivo dos artigos e revisões publicadas nos bancos de dados de literatura como SCOPUS e PubMed; enfatizando-se que os dados foram extraídos até Agosto de 2023. Na SCOPUS, a busca baseou-se em consulta de “Resumo, Título e Palavra-chave”, enquanto no PubMed, utilizou-se “Título e Resumo”, conforme detalhado na Tabela 1. As palavras-chave utilizadas foram “*Gestational Age*”, “*Premature Birth*”, “*Preterm Birth*”, “*Prematurity*”, “*Machine Learning*”, “*Health Records*”, “*Medical Records*”, e “*Decision Making*”.

Tabela 1 – Publicações em SCOPUS e PubMed

PALAVRAS-CHAVE	SCOPUS	PUBMED
“ <i>Gestational Age</i> ” AND “ <i>Machine Learning</i> ”	461	198
(“ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ” OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ”	434	200
(“ <i>Gestational Age</i> ” OR “ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ” OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ”	724	353
(“ <i>Gestational Age</i> ” OR “ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ” OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ” AND (“ <i>Health Records</i> ” OR “ <i>Medical Records</i> ”)	63	18
(“ <i>Gestational Age</i> ” OR “ <i>Premature Birth</i> ” OR “ <i>Preterm Birth</i> ” OR “ <i>Prematurity</i> ”) AND “ <i>Machine Learning</i> ” AND “ <i>Decision Making</i> ”	39	17

Fonte: Base de dados de SCOPUS e PubMed (2023)

Na Tabela 1, pode-se observar uma quantidade semelhante de artigos publicados utilizando o termo “*Gestational Age*” e os termos “*Premature Birth*” OR “*Preterm Birth*” OR “*Prematurity*”, adicionando “*Machine Learning*”, totalizando 461 e 434 artigos publicados em SCOPUS, respectivamente, e 198 e 200 artigos publicados no PubMed, respectivamente.

Considerando a combinação de palavras-chave “*Gestational Age*” OR “*Premature Birth*” OR “*Preterm Birth*” OR “*Prematurity*”, adicionando o termo “*Machine Learning*”, foram encontrados 724 artigos publicados na SCOPUS e 353 no PubMed. Esses artigos foram considerados na análise bibliométrica, ujeitos a uma revisão para identificar documentos duplicados. Após o processo de depuração, obteve-se um conjunto final de dados contendo 757 itens únicos.

Tabela 2 – Informações principais do conjunto final de publicações

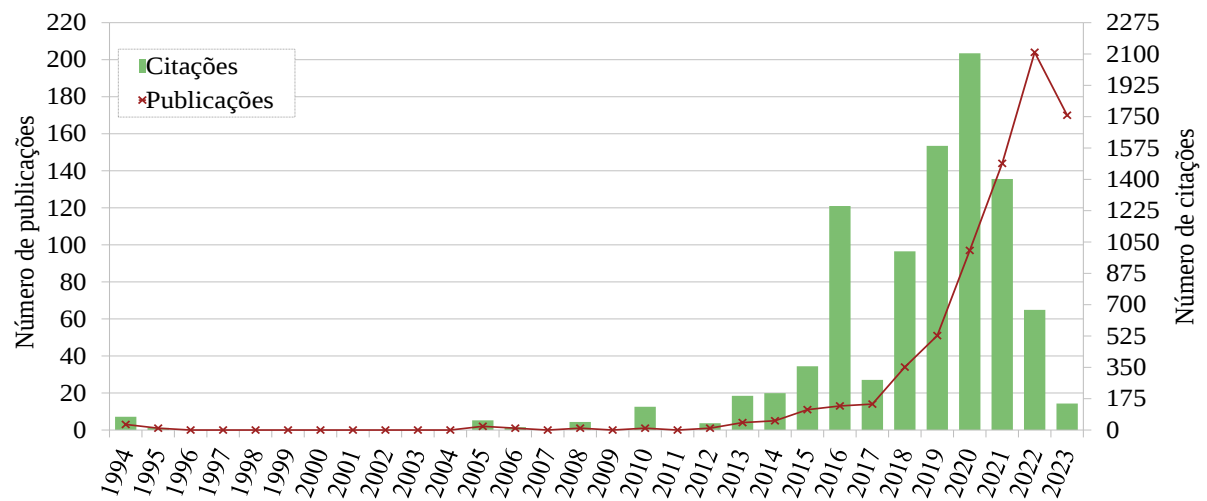
DESCRIÇÃO	RESULTADO
Período	1994 - 2023
Fontes (Revistas)	381
Total de publicações	757
Publicações de autoria única	9
Publicações de autoria múltipla	748
Média de citações por publicação	12,63
Palavras-chave do autor	1789
Autores	4636
Autores de publicações de autoria única	9
Autores de publicações de autoria múltipla	4627
Autores por publicação	8,80
Índice de colaboração	6,19

Fonte: Base de dados de SCOPUS e PubMed (2023)

A Tabela 2 resume as principais informações das publicações, revelando que, no período de 1994 a 2023, foram identificadas 757 publicações, entre artigos e revisões, relacionadas à idade gestacional, nascimentos prematuros, Aprendizado de Máquina (AM), dados de saúde e tomada de decisões. Essas publicações estão distribuídas em 318 revistas distintas e contam com a contribuição de 4.636 autores no total. A média de citações por publicação é de aproximadamente 13, totalizando 1.789 palavras-chave. Além disso, nota-se que o índice de colaboração é de 6,19, representando a média de autores por publicação de autoria múltipla.

Observa-se na Figura 2 série temporal da distribuição dos documentos publicados ao longo dos anos e a tendência do impacto dessas publicações. Da mesma forma, percebe-se um aumento no número de publicações nos últimos 5 anos, representando 87,98% dos documentos analisados, com uma taxa de crescimento percentual anual de 14,94%. Similarmente, o número de citações

Figura 2 – Número de publicações e citações por ano

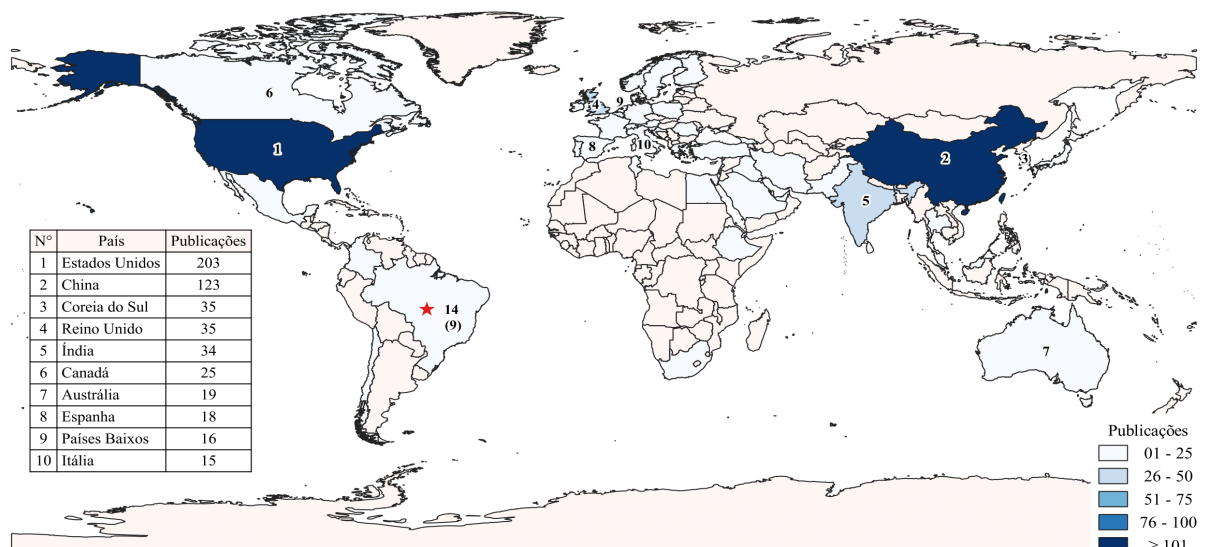


Fonte: Base de dados de SCOPUS e PubMed (2023)

tem aumentado nos últimos anos, acumulando um total de 9563 citações, sendo 61,82% destas referentes a documentos publicados nos últimos cinco anos. Por outro lado, espera-se que o número de documentos publicados quanto o número de citações continuem a crescer nos próximos anos, dado o cenário de crescimento constante na publicação de documentos nesta área.

Sob uma perspectiva diferente, analisamos as publicações por país, uma vez que essa análise ajuda a identificar o interesse de cada país em um determinado campo de estudo. A Figura 3 apresenta um mapa de calor global com o número de documentos publicados por país do autor correspondente. Identificou-se que os documentos do conjunto de dados foram publicados por 50 países.

Figura 3 – Número de publicações por país do autor correspondente



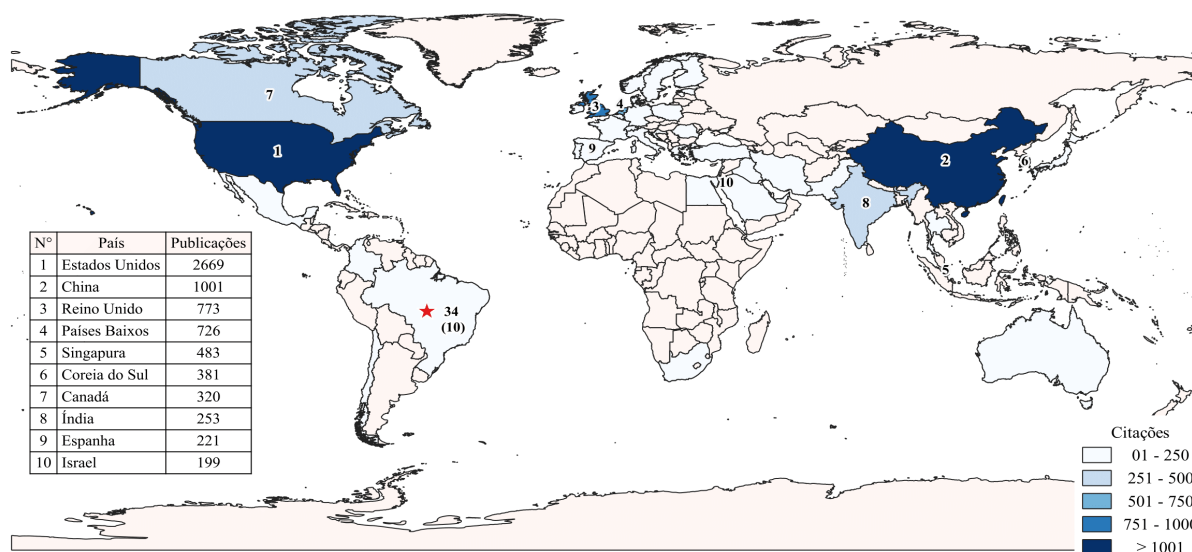
Fonte: Base de dados de SCOPUS e PubMed (2023)

Dessa análise, observamos que Estados Unidos (203) é o país com maior número de documentos publicados, seguido pela China e Coreia do Sul, com 123 e 35 publicações, respectivamente. Destaca-se ainda que 36,30% dos documentos publicados são provenientes de países do continente Americano, 36,15% do continente Asiático, 23,61% do continente Europeu, e apenas 2,77% e 1,17% restantes pertencem aos continentes da Oceania e África, respectivamente.

No entanto, é relevante destacar que o Brasil está na 14ª posição, contribuindo com apenas nove publicações em autoria. Essas publicações abordam diferentes estudos, tais como a identificação de diferentes fenótipos em recém-nascidos que apresentam peso extremamente baixo mediante técnicas de ANS (MATSUSHITA; KREBS; CARVALHO, 2022), o uso de ANS para estratificar o risco de prematuridade no Brasil por meio de dados socioeconômicos (LOPES; BARBOSA; FERNANDES, 2022), o uso de AS para compreender o papel da exposição ao contato pele a pele e a duração mínima na determinação da amamentação exclusiva em casos de alta hospitalar de recém-nascidos prematuros com peso de até 1.800 gramas (GOUDARD *et al.*, 2022), e a validação de um dispositivo adotando o modelo fotobiológico de maturidade da pele em recém-nascidos prematuros em função dos dados clínicos para a detecção da idade gestacional com AM. (REIS *et al.*, 2022).

Do mesmo modo, quando os países foram avaliados pelo número de citações (Figura 4), destaca-se que Estados Unidos (2669) lidera também nesse aspecto, recebendo o maior número de citações. A China (1001) fica em segundo lugar em número de citações, seguida pelo Reino Unido, com 773 citações, evidenciando a forte influência dessas publicações no avanço de novas pesquisas.

Figura 4 – Número de citações por país do autor correspondente



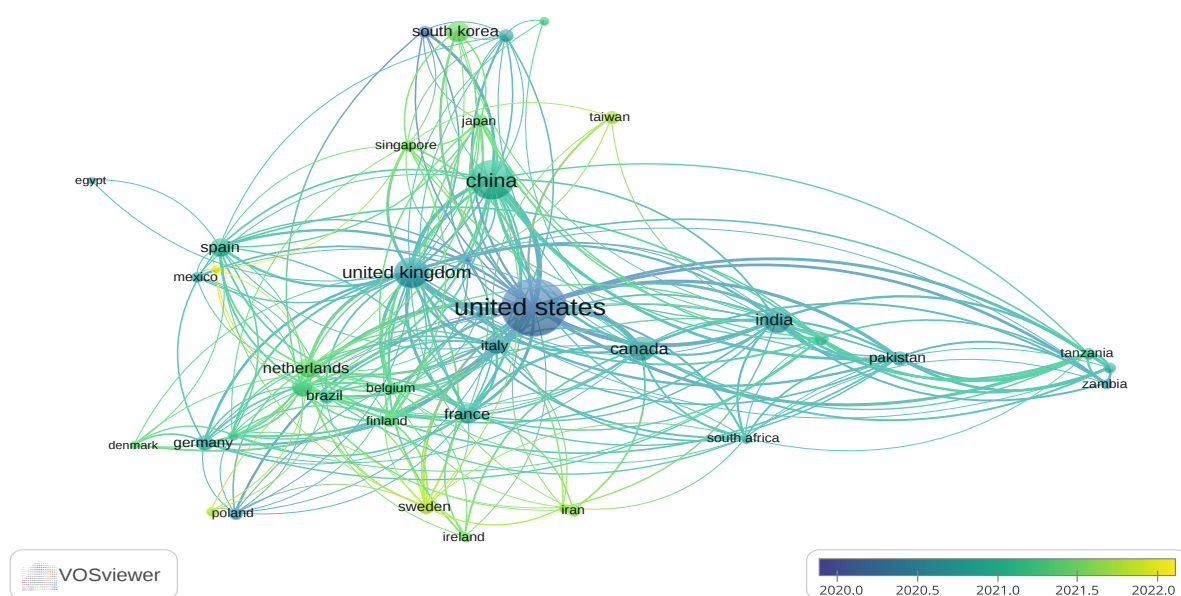
Fonte: Base de dados de SCOPUS e PubMed (2023)

Assim, o baixo número de publicações sugere uma oportunidade de pesquisa para utilizar as técnicas de AM com os dados disponibilizados pelo Ministério de Saúde, com a finalidade de desenvolver ferramentas que auxiliem especialistas na tomada de decisões. Nesse contexto, este

trabalho propõe o desenvolvimento de um modelo preditivo para identificação de mães com risco de parto prematuro. Sob uma perspectiva mais ampla e conforme apontado por Santos (2018), no Brasil, há poucas pesquisas sobre a aplicação de AM na área da saúde, o que representa uma significativa oportunidade para o desenvolvimento de modelos inteligentes.

Além disso, foi identificada uma rede colaborativa internacional envolvendo 68 países por meio da avaliação da produção científica. Utilizando o VOSviewer, a Figura 5 representa a rede de colaboração internacional, composta por 40 países que apareceram em cinco ou mais documentos publicados. De acordo com Eck e Waltman (2023), o tamanho do círculo indica o número de documentos nos quais o país aparece, enquanto a cor representa o ano médio de publicação (no período 2020-2022). Dessa forma, observamos que os Estados Unidos e a China lideram não apenas em número de documentos, mas também em colaborações internacionais, seguidos pelo Reino Unido.

Figura 5 – Rede de colaboração internacional

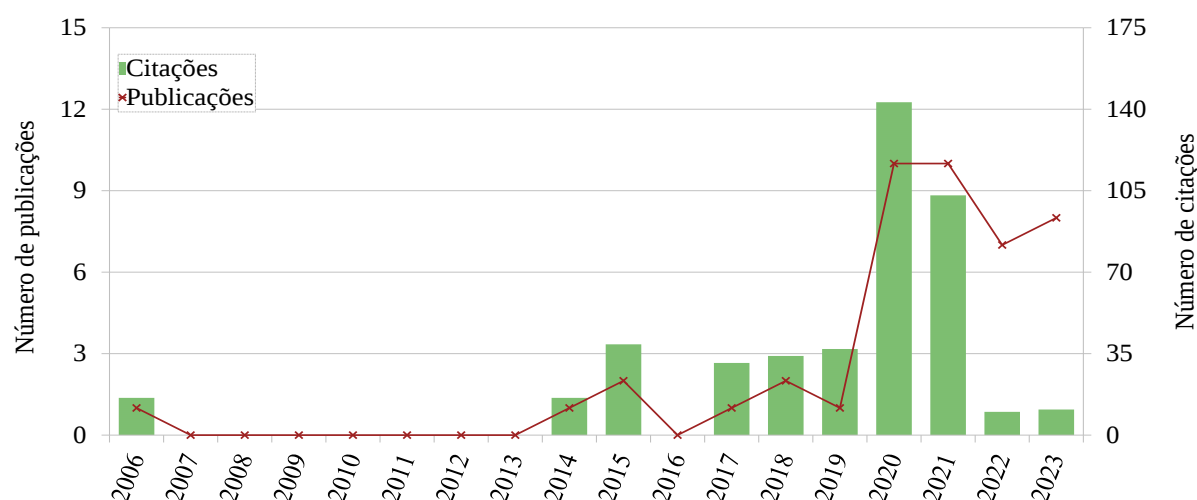


Fonte: Base de dados de SCOPUS e PubMed (2023)

Adicionalmente, foi realizada uma análise das palavras-chave, por meio do gráfico de densidade, onde foram mapeadas as palavras-chave mais utilizadas pelos autores das publicações, como mostrado na Figura 6. A cor no gráfico indica a densidade dos itens, variando de azul, verde a amarelo, sendo o amarelo indicativo de uma frequência maior no uso da palavra-chave (ECK; WALTMAN, 2023). Notavelmente, as palavras-chave "Aprendizado de Máquina", "Aprendizado Profundo" e "Inteligência Artificial" foram as mais frequentemente utilizadas, assim como a relevância do termo "Nascimento prematuro" e "Idade gestacional".

Também, utilizando a sequência de palavras-chave (“*Gestational Age*” OR “*Premature Birth*” OR “*Preterm Birth*” OR “*Prematurity*”) AND “*Machine Learning*” e adicionando os termos “*Health Records*” OR “*Medical Records*”, obteve-se um total de 65 documentos únicos

Figura 8 – Número de publicações e citações por ano das palavras-chave (“*Gestational Age*” OR “*Premature Birth*” OR “*Preterm Birth*” OR “*Prematurity*”) AND “*Machine Learning*” AND “*Decision making*”



Fonte: Base de dados de SCOPUS e PubMed (2023)

período de 2006 a 2023, com um crescimento percentual anual de 13,01%, conforme ilustrado na Figura 8.

Por último, é relevante abordar a visibilidade e o impacto do desenvolvimento desta pesquisa em âmbitos acadêmicos, sociais, regionais e nacionais. Nota-se que no Brasil há uma escassez de pesquisas sobre a aplicação de técnicas de AM na área da Saúde. De acordo com as bases de dados SCOPUS e PubMed, foram encontradas apenas nove publicações pertencentes ao Brasil, predominantemente nos últimos dois anos, com a autoria principal, embora estas não apresentem uma forte relação com a abordagem proposta. Em contraste, este trabalho seguirá sua metodologia utilizando dados da cidade de São Paulo.

1.4.1 Aprendizado de máquina na predição da idade gestacional para a identificação de recém-nascidos prematuros

Considerando a idade gestacional como uma ferramenta crucial na avaliação da saúde materna e neonatal (NAIDU; FREDLUND, 2023), a inclusão desta em modelos preditivos contribui para uma abordagem mais precisa e personalizada no cuidado pré-natal, permitindo intervenções oportunas e melhorando os resultados tanto para as gestantes quanto para os recém-nascidos.

Portanto, a descrição de alguns artigos relacionados com a abordagem proposta, incluindo Autor, Título, Data, Métodos de Aprendizado de Máquina e Modelo Final selecionado, está apresentada no Quadro 1.

Quadro 1 – Estudos relacionados com a abordagem proposta

Autor	Título	Dados	Modelo final
Reis <i>et al.</i> (2022)	<i>Newborn Skin Maturity Medical Device Validation for Gestational Age Prediction: Clinical Trial</i>	Grupo de estudo composto por 702 gestantes que deram à luz 781 recém-nascidos, com dados da ultrassonografia do primeiro trimestre, uma segunda ultrassonografia comparadora e os dados referentes ao último período menstrual dos relatórios pré-natais.	o modelo Extreme Gradient Boosting com EMA = 1,688 semanas
Sazawal <i>et al.</i> (2022)	<i>Machine learning prediction of gestational age from metabolic screening markers resistant to ambient temperature transportation: Facilitating use of this technology in low resource settings of South Asia and East Africa</i>	Um subconjunto de 330 pares de amostras de manchas de sangue secas por punção de calcanhar, em gelo seco e em temperatura ambiente.	O modelo de regressão de Random Forest com REQM=1,24 e 1,28 / EMA=1,09 e 1,12, para amostras em gelo seco e em temperatura ambiente, respectivamente.
Shen <i>et al.</i> (2022)	<i>Attention-guided deep learning for gestational age prediction using fetal brain MRI</i>	Um conjunto de dados de 741 imagens de ressonância magnética ponderadas em T2.	O modelo de Redes Neurais Convolucionais (RNC) com $R^2=0,945$, $r=0,970$ e EMA=6,7 (dias).
Rocha <i>et al.</i> (2021)	<i>Data-driven risk stratification for preterm birth in Brazil: a population-based study to develop of a machine learning risk assessment approach</i>	Dados de 3.876.666 mães com nascimentos vivos em 3.929 municípios brasileiros, incluindo variáveis sobre procedimentos de cuidados pré-natais e infraestrutura física em unidades de atenção primária, características da saúde da mulher e do parto.	O modelo Extreme Gradient Boosting com REQM=2,094.
Sazawal <i>et al.</i> (2021)	<i>Machine learning guided postnatal gestational age assessment using newborn screening metabolomic data in South Asia and sub-Saharan Africa</i>	Dados sobre ultrassonografia na triagem metabólica de 1.318 recém-nascidos na Ásia Meridional e na África Subsaariana.	O modelo Regressor de Floresta Aleatória com EMA=0,74 e REQM=1,02.

Fonte: Próprio autor (2023)

1.5 DELIMITAÇÃO DA PESQUISA

Entre as limitações desta pesquisa, destaca-se que os modelos foram desenvolvidos exclusivamente com algoritmos de Regressão de Aprendizado de Máquina, portanto, não houve exploração ou aplicação de outros algoritmos pertencentes à mineração de dados ou Inteligência Artificial. As variáveis que são consideradas são aquelas disponibilizadas pelo Sistema de Informações Sobre Nascidos Vivos (SINASC). O estudo se baseou nos dados da cidade de São Paulo para a identificação do risco de prematuridade, logo, os resultados não podem ser generalizados para outras cidades ou diferentes complicações de parto. Adicionalmente, a análise abrangeu um período de 5 anos (2015-2019) devido à ausência de informações em algumas das variáveis estudadas nos anos anteriores, como nos anos de 2010 e 2011, apresentando porcentagens de 99,99% e 86,03%, respectivamente, de instâncias com dados faltantes.

1.6 ESTRUTURA DA PESQUISA

Este trabalho está estruturado da seguinte forma: O Capítulo 2 apresenta a fundamentação teórica, envolvendo a definição dos temas relevantes para o trabalho. No Capítulo 3 descreve-se a área geográfica em análise, o conjunto de dados e a metodologia utilizada no desenvolvimento do trabalho. O Capítulo 4 apresenta os resultados e a discussão. Finalmente, nos Capítulos 5 e 6, são abordadas as conclusões, sugestões e discussões, seguidas das referências bibliográficas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 PREMATURIDADE

A prematuridade é definida como os recém-nascidos vivos com uma idade gestacional inferior a 37 semanas. Na maioria dos casos, os recém-nascidos podem apresentar a condição de nascer pré-termo por causa de um trabalho de parto espontâneo ou devido a uma indicação médica de planejar uma indução do trabalho de parto ou um parto cesáreo precoce (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2022).

Assim, o nascimento prematuro resulta de um relacionamento complexo entre múltiplos determinantes ou fatores, tais como biológicos, maternos, sociais, econômicos e de serviços de saúde (TASCÓN *et al.*, 2016; MALIK; BAARDA, 2020; ROCHA *et al.*, 2022; RODRÍGUEZ *et al.*, 2024).

- Determinantes biológicos: considera características como histórico de parto prematuro, gravidez múltipla, comorbidades maternas (hipertensão ou diabetes) ou outros fatores relacionados ao sofrimento fetal.
- Determinantes maternos: incluem a idade da mãe (jovem ou avançada), raça/etnia, nível de educação ou estado conjugal.
- Determinantes sociais: relacionados a indicadores de desigualdade.
- Determinantes econômicos: incluem a ocupação ou renda familiar.
- Determinantes de serviços de saúde: envolvem a presença de profissionais de saúde, disponibilidade de insumos e qualidade na atenção pré-natal.

A sobrevivência infantil, aliada à redução dos casos de nascimentos pré-termo, demanda uma colaboração abrangente entre diversos órgãos, incluindo governos e a comunidade, para a formulação de políticas sociais e de saúde. Essas políticas visam possibilitar a criação de programas abrangentes, envolvendo planejamento familiar, nutrição, cuidado no consumo de substâncias psicoativas, além de contemplar melhorias nos serviços de atendimento em saúde (TASCÓN *et al.*, 2016; RODRÍGUEZ *et al.*, 2024).

2.2 APRENDIZADO DE MÁQUINA

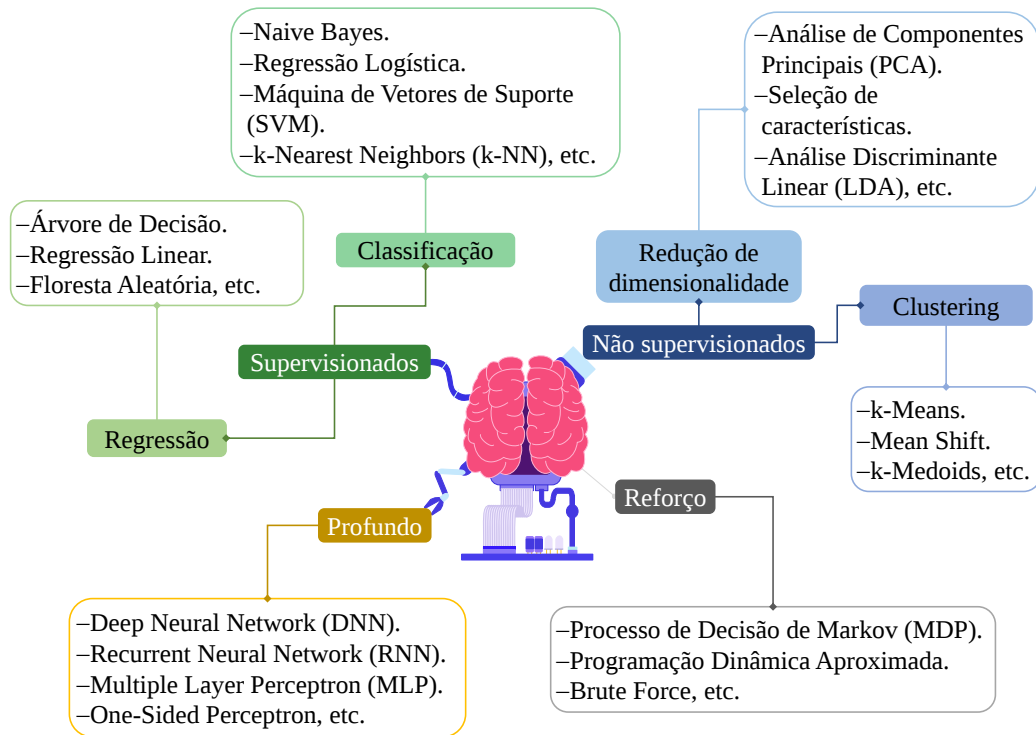
A principal tarefa do AM é realizar inferências a partir de uma quantidade específica de dados amostrais, focando em como os computadores aprendem por meio desses dados (DEO, 2015).

O AM é um subcampo da Inteligência Artificial (IA), onde a inteligência humana é simulada, permitindo que os computadores adquiram conhecimento por meio de dados históricos e

algoritmos para aplicação futura (PORTUGAL; ALENCAR; COWAN, 2018). Assim, utilizando métodos computacionais, o AM possibilita a identificação de padrões complexos em grandes conjuntos de dados, aprendendo um modelo por meio da aquisição de conhecimento (JORDAN; MITCHELL, 2015; OBERMEYER; EMANUEL, 2016; DOUPE; FAGHMOUS; BASU, 2019). O AM utiliza técnicas estatísticas para construir modelos matemáticos (DEO, 2015).

Além disso, como ilustrado na Figura 9, os modelos de AM podem ser desenvolvidos por meio de diferentes técnicas, classificadas como técnicas supervisionadas (Aprendizado Supervisionado - AS), não supervisionadas (Aprendizado Não Supervisionado - ANS), por reforço (Aprendizado por Reforço - AR) e aprendizado profundo (Aprendizado Profundo - AP).

Figura 9 – Classificação das técnicas de AM



Fonte: Adaptação de Deo (2015) e Jordan e Mitchell (2015)

2.2.1 Aprendizado Supervisionado

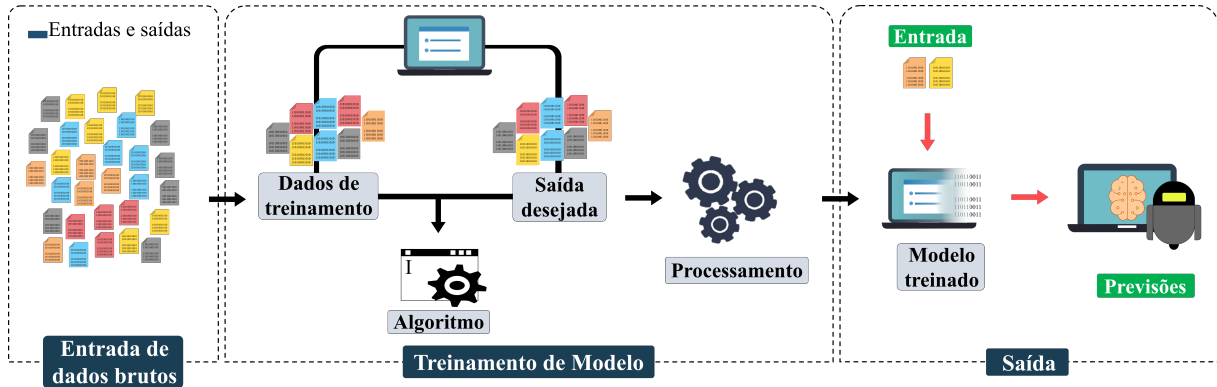
No caso do AS (também conhecido como "aprendizado com instrutor"), os algoritmos pertencentes a esta técnica requerem dados rotulados, ou seja $X \times Y$, sendo o conjunto de dados de entrada representado por um vetor de variáveis $X = x_1, x_2, \dots, x_d \in \mathbb{R}^d$, d a dimensão do vetor, x_d cada uma das variáveis de entrada, e o Y representa o vetor de variáveis de saída (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; SHAI; SHAI, 2014; FLATH; STEIN, 2018).

O principal objetivo do AS é aprender uma função $x \rightarrow f(x)$, para prever uma saída e baseado em um conjunto de variáveis aleatória de entrada x , conforme a expressão 2.1.

$$y = f(x, \beta) \quad (2.1)$$

sendo y o valor de saída calculado pela função $f(x)$ e os parâmetros β , que foram gerados a partir do treinamento (MARSLAND, 2015). De acordo com RODRÍGUEZ (2020), o processo de análise do AS pode se ilustrar conforme a Figura 10.

Figura 10 – Procedimento do Aprendizado Supervisionado



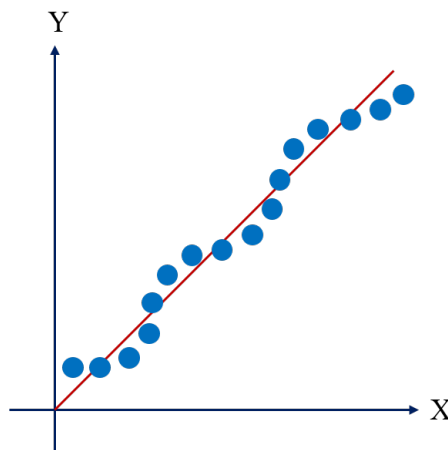
Fonte: RODRÍGUEZ (2020)

Assim, os algoritmos de AS visam deduzir no processo de treinamento uma função de predição tanto para saídas com valores contínuos ou categóricos (FLATH; STEIN, 2018), com métodos de Regressão ou Classificação respectivamente.

- **Métodos de Regressão:** A tarefa principal dos métodos de Regressão é prever saídas com valores contínuos (variável quantitativa). O objetivo dos métodos de regressão consiste em determinar a relação entre uma variável de saída aleatória Y e um vetor de variáveis de entrada X (Figura 11), buscando estimar a função de regressão dada pela Equação 2.2 (IZBICKI; SANTOS, 2020).

$$r(X) := \mathbb{E}[Y/X = x] \quad (2.2)$$

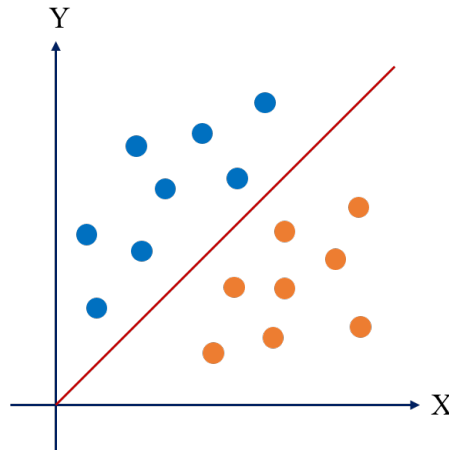
Figura 11 – Regressão



Fonte: Próprio autor (2023)

- **Métodos de Classificação:** No contrário dos métodos de regressão, os métodos de classificação têm a tarefa de prever saídas com valores categóricos (variável qualitativa). O objetivo é prever a probabilidade de uma instância pertencer a cada uma das categorias da variável alvo (JAMES *et al.*, 2021), assim a variável de saída pode ser binária (Figura 12) ou multiclasse.

Figura 12 – Classificação



Fonte: Próprio autor (2023)

2.2.1.1 Algumas técnicas de AM

AM para estudos de estimação e diagnóstico na área da saúde, tem a tarefa de prever respostas contínuas ou classificar um evento em categorias binárias ou multiclasse (MARSLAND, 2015; ALPAYDIN, 2020). Assim, algumas das técnicas ou modelos de regressão e classificação são descritos a seguir:

– Regressão Linear

Regressão linear multivariada utilizada para prever respostas contínuas, além de estudar a relação linear entre os recursos de entrada e o recurso de saída de um conjunto de dados (MARSLAND, 2015; SANTOS, 2018), de acordo com a função de regressão expressada nas covariáveis X e os coeficientes dos parâmetros β .

$$F(X) = \beta^T X = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon \quad (2.3)$$

$$F(X) = \beta_0 + \sum_{j=1}^p \beta_j X_j \quad (2.4)$$

Este tipo de modelo utiliza com maior frequência o método dos mínimos quadrados para estimar os coeficientes β (IZBICKI; SANTOS, 2020). O processo do modelo de regressão linear é calcular e minimizar a distância (erro) entre uma linha ajustada e cada um dos pontos dos dados (RODRÍGUEZ, 2020), e em outras palavras, basou-se na estimação dos

parâmetros β que permitam minimizar função de perda do modelo (Soma de Quadrados dos Resíduos - SQR) (SANTOS, 2018).

$$SQR(\beta) = \sum_{i=1}^n (Y_i - F(X_i))^2 \quad (2.5)$$

– Regressão Logística

Modelo linear utilizado como método de classificação, por meio da predição da probabilidade de ocorrência de um determinado evento, de acordo com as classes da variável alvo, por meio da utilização da função *logit* (JAIN; CHATTERJEE, 2020).

$$\text{logit} = \log \left(\frac{p}{1-p} \right) = \log \left(\frac{\text{Probabilidade de ocorrência}}{\text{Probabilidade do evento não ocorrer}} \right) \quad (2.6)$$

Um destaque muito importante a fazer, é que a estimação da probabilidade de ocorrência de um evento deve estar definida no intervalo $[0 ; 1]$ (SANTOS, 2018). Desta forma, quando a variável alvo Y é binária, a função de regressão logística é expressa de acordo com a Equação 2.7 (IZBICKI; SANTOS, 2020).

$$\mathbb{P}(Y = 1/X) = \frac{e^{\beta_0 + \sum_{i=1}^d \beta_i X_i}}{1 + \beta_0 + \sum_{i=1}^d \beta_i X_i} \quad (2.7)$$

– Árvore de Decisão

Método utilizado para problemas de regressão ou classificação, sendo muito popular por apresentar uma estrutura simples e muito eficiente.

Numa árvore de decisão, a região local é determinada em uma sequência de divisões recorrentes para um número reduzido de etapas, além de ser composta por nós internos e folhas terminais, e dado um *input*, um teste é aplicado e uma das ramificações é escolhida, como um processo repetitivo que começa na raiz até que um nó folha atinja o valor que constitui a saída (ALPAYDIN, 2020).

Para modelos de regressão, a predição da variável de saída Y , com vetor X de variáveis de entrada é determinada pela média das observações da variável resposta, segundo o expressado na Equação 2.8.

$$F(X) = \text{média}\{Y_i : X_i \in \mathbb{R}^d\} \quad (2.8)$$

Caso contrário, em modelos de classificação, a predição da variável de saída Y , é dada pela moda das observações, de acordo com as classes e expressado na Equação 2.9 (IZBICKI; SANTOS, 2020; JAMES *et al.*, 2021).

$$F(X) = \text{moda}\{Y_i : X_i \in \mathbb{R}^d\} \quad (2.9)$$

Por outro lado, para melhorar desempenho dos modelos, foram desenvolvidos outros métodos que integram mais do que uma árvore. Alguns destes métodos, também conhecidos como métodos de *ensemble*, são a Floresta Aleatória, *Bagging*, e *Gradient Boosting* (JAMES *et al.*, 2021).

– Floresta Aleatória:

Considerada uma extensão do método *bagging*, a floresta aleatória é composta por um conjunto de árvores de decisão, gerado a partir de uma seleção aleatória de variáveis em cada nó, como forma de determinar a divisão (JAIN; SHUKLA; WADHVANI, 2018)

No modelo de floresta aleatória, a predição da b -ésima árvore é denotada por $\hat{F}_{rf}(X)$ de acordo com as classes de alguma variável de saída (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; SANTOS, 2018). O método de floresta aleatória é determinado pela média (Equação 2.10) ou pelo voto majoritário (2.11), seja para métodos de regressão ou classificação respectivamente.

$$\hat{F}_{rf}(X) = \text{média}\{F_b(X)\}_1^B \quad (2.10)$$

$$\hat{F}_{rf}(X) = \text{voto majoritário}\{F_b(X)\}_1^B \quad (2.11)$$

Apontando que para métodos de classificação, a classe mais indicada pelas B árvores agregadas no modelo vai determinar a predição final.

– *Gradient Boosting*:

Método que combina uma série de classificadores sequenciais, apresentando um peso diferente e utilizando o erro calculado na melhora do valor de predição, tornando-se um método de regressão ou classificação robusto (MARSLAND, 2015).

Neste método, o estimador $\hat{F}_{gb}(X)$ é construído incrementalmente, com a atribuição inicial de $\hat{F}_{gb}(X) \equiv 0$, mas o estimador possui alto viés e variância baixa que vai melhorando conforme seja atualizado o valor de F , ao adicionar uma função que permita perver os resíduos $r_i = Y_i - \hat{F}_{gb}(X_i)$ (IZBICKI; SANTOS, 2020; JAMES *et al.*, 2021).

– *Extreme Gradient Boosting*:

Algoritmo utilizado para problemas de regressão ou classificação. *Extreme Gradient Boosting* também é considerado como um método de *ensemble* de árvores de decisão, desenvolvido para ser altamente flexível, minimizando a função de perda com a criação de uma expansão aditiva (CHEN; GUESTRIN, 2016).

Este algoritmo é considerado uma versão melhorada do *Gradient Boosting*, utilizando uma busca dividida de padrões no conjunto de dados, tendo maior eficiência e controle no risco de sobreajuste (RODRÍGUEZ *et al.*, 2021).

2.2.1.2 Validação cruzada e métricas de desempenho

A avaliação dos modelos de AM pode ser realizada por meio de validação cruzada. A Equação 2.12 mostra a validação cruzada de *k-Fold* (ALPAYDIN, 2020).

$$\begin{aligned} V_1 &= X_1 & T_1 &= X_2 \cup X_3 \cup \dots \cup X_k \\ V_2 &= X_2 & T_2 &= X_1 \cup X_3 \cup \dots \cup X_k \\ &\vdots & &\vdots \\ V_k &= X_k & T_k &= X_1 \cup X_2 \cup \dots \cup X_{k-1} \end{aligned} \quad (2.12)$$

O conjunto de dados X é dividido aleatoriamente em k subconjuntos (*k-Fold*) de igual tamanho (X_i , onde $i = 1, 2, \dots, k$). Além disso, é usado para o treinamento do modelo e os dados restantes são usados para a validação do modelo, assumindo que ambos os conjuntos de dados (validação e treinamento) seguem a mesma distribuição (BERRAR, 2019; ALPAYDIN, 2020).

A avaliação do desempenho dos modelos testados é baseada na comparação de diferentes métricas. De acordo com Chicco, Warrens e Jurman (2021), as métricas de erro mais utilizadas para avaliar o desempenho dos algoritmos de regressão são os seguintes:

- Erro Médio Absoluto (EMA): Formulado como a diferença média absoluta entre os valores de saída reais e previstos, conforme mostrado na Equação 2.13. Como exemplo, se o resultado de EMA for 3, significa que a predição do modelo pode estar errando em média ± 3 pontos com relação ao valor real.

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - \hat{Y}_i| \quad (2.13)$$

- Erro Quadrático Médio (EQM): Definido como a diferença média ao quadrado entre os valores de saída reais e previstos, de acordo com a Equação 2.14. Esta métrica pode apresentar problemas de interpretabilidade pela elevação ao quadrado.

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (2.14)$$

- Raiz do Erro Quadrático Médio (REQM): Trata-se da raiz quadrada simples do EQM, conforme mostrado na Equação 2.15. No caso de ter um resultado de REQM = 5, significa

que a predição do modelo pode estar errando em ± 5 pontos com relação ao valor real.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2} \quad (2.15)$$

- Erro Percentual Absoluto Médio (EPAM): Também conhecido como o desvio percentual absoluto médio, de acordo com a Equação 2.16. Com um valor de EPAM = 25%, significa que o nosso modelo realizou predições com um 25% de diferença média entre o valor predito e o real.

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i - \hat{Y}_i|}{Y_i} * 100 \quad (2.16)$$

Onde para cada uma das métricas, N representa o número de dados, Y_i os valores de saída atuais ou reais, e $\hat{Y}_i$ indica os valores de saída preditos.

Por outro lado, para algoritmos de classificação (MARSLAND, 2015), as métricas de desempenho utilizando a matriz de confusão são apresentadas a seguir:

		Previsto		Total
		Positivo	Negativo	
Atual	Positivo	VP	FP	VP + FP
	Negativo	FN	VN	FN + VN
Total		VP + FN	FP + VN	

- Acurácia: Trata-se da soma do total de verdadeiros positivos e verdadeiros negativos divididos pelo número total de resultados (Equação 2.17). Um valor de acurácia de 0,70 indica que um 70% do total de valores preditos foram classificados corretamente entre as classes.

$$Acurácia = \frac{n^o VP + n^o VN}{n^o VP + n^o VN + n^o FP + n^o FN} \quad (2.17)$$

- Precisão: É a razão entre o número de verdadeiros positivos e o número total de valores positivos (Equação 2.18). Como exemplo, um valor de precisão de 0,60 indica que 60% dos recém-nascidos prematuros realmente eram prematuros.

$$Precisão = \frac{n^o VP}{n^o VP + n^o FP} \quad (2.18)$$

- Revocação: Relação do número de verdadeiros positivos para as previsões totais classificadas como positivas (Equação 2.19). Considerando um valor de revocação de 0,65 indica que o modelo conseguiu acertar um 65% dos recém-nascidos prematuros.

$$Revocação = \frac{n^o VP}{n^o VP + n^o FN} \quad (2.19)$$

- *F1-Score*: Medida que combina a precisão e a sensibilidade (revocação) da previsão do modelo (Equação 2.20).

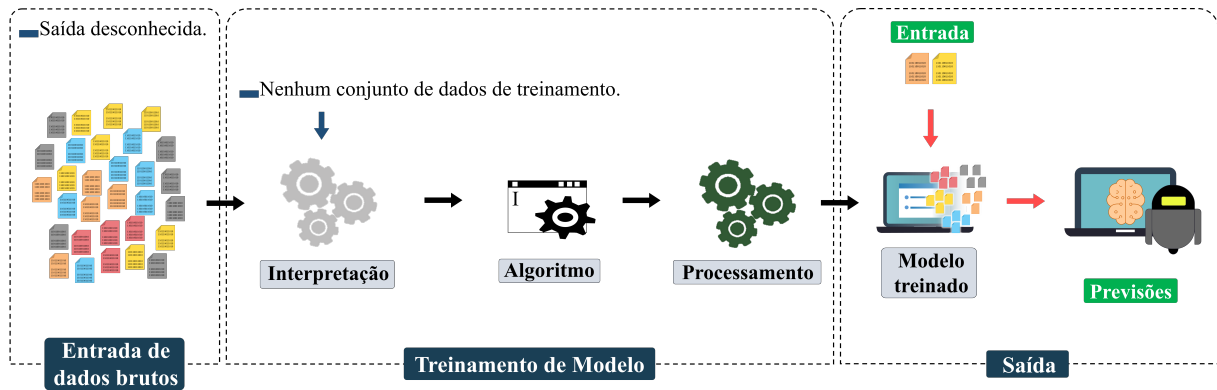
$$F1-Score = 2 * \frac{Precisão * Revocação}{Precisão + Revocação} \quad (2.20)$$

Outra métrica avaliada nos algoritmos de classificação é a Área Sob a Curva Característica de Operação do Receptor (ROC-AUC), utilizada para medir o desempenho geral dos modelos (MARSLAND, 2015; SANTOS, 2018), com o intuito de ordenar os recém-nascidos com maior risco, para aqueles com um risco baixo.

2.2.2 Aprendizado Não Supervisionado

Caso contrário, os algoritmos que pertencem ao ANS (também conhecido como "aprendizado sem instrutor") não requerem dados rotulados e, nessa situação, os algoritmos de ANS recebem apenas os dados com variáveis de entrada, expressado pelo vetor $X = x_1, x_2, \dots, x_d \in \mathbb{R}^d$ de dimensão d , com a finalidade de extrair conhecimento (MÜLLER; GUIDO, 2016; FLATH; STEIN, 2018). RODRÍGUEZ (2020) ilustra o procedimento de análise do ANS por meio da Figura 13.

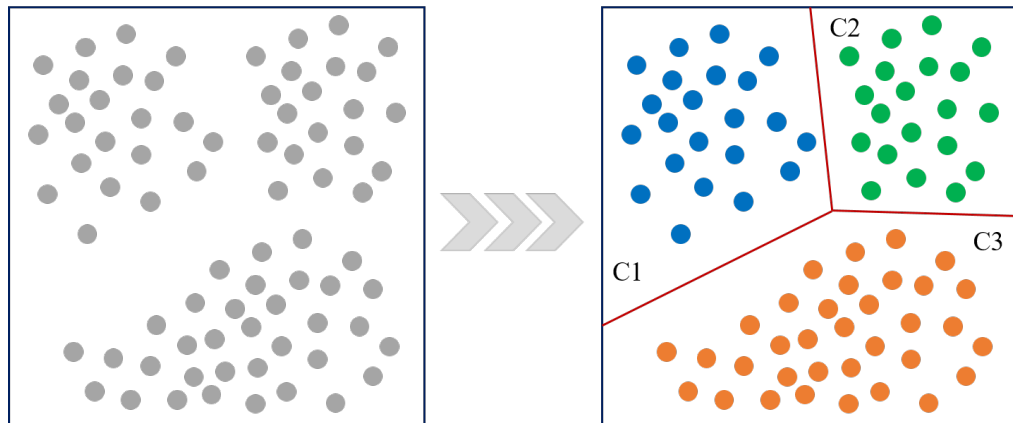
Figura 13 – Procedimento do Aprendizado Não Supervisionado



Fonte: RODRÍGUEZ (2020)

O objetivo dos métodos de ANS é encontrar regularidades nas variáveis de entrada, analisando a estrutura, identificando possíveis padrões que ocorram com maior frequência do que outros ((ALPAYDIN, 2020)). Desta forma, os métodos mais conhecidos são o *Clustering* e de Redução de Dimensionalidade.

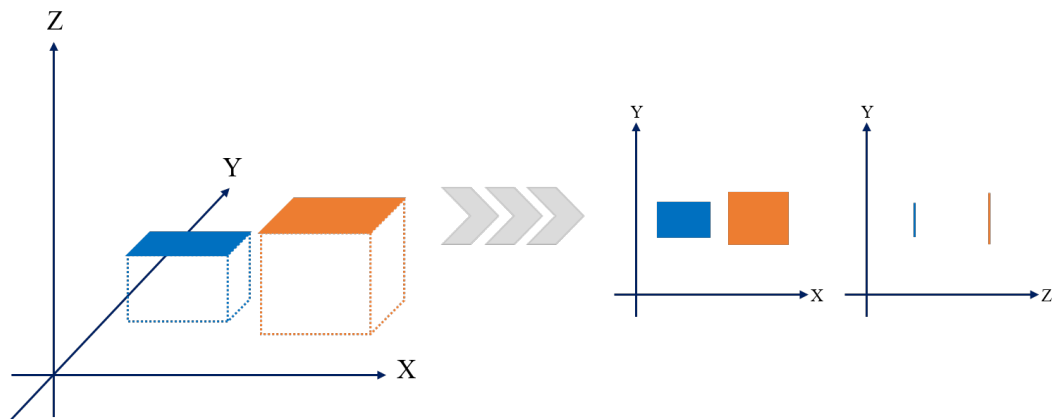
- Métodos de *Clustering*: A análise de *Clusters* refere-se à identificação de agrupamentos no conjunto de dados (Figura 14), com o objetivo de encontrar grupos distintos com observações muito semelhantes dentro de cada grupo e diferentes com observações de outros grupos (IZBICKI; SANTOS, 2020; JAMES *et al.*, 2021).

Figura 14 – *Clustering*

Fonte: Próprio autor (2023)

- Métodos de Redução de Dimensionalidade: Os métodos de redução de dimensionalidade seguem o processo de fazer transformações de dados num espaço de alta dimensão para um novo espaço com uma dimensão muito menor (SHAI; SHAI, 2014). Assim, o objetivo destes métodos se trata da criação de um número menor de variáveis representativas (Figura 15).

Figura 15 – Redução de Dimensionalidade



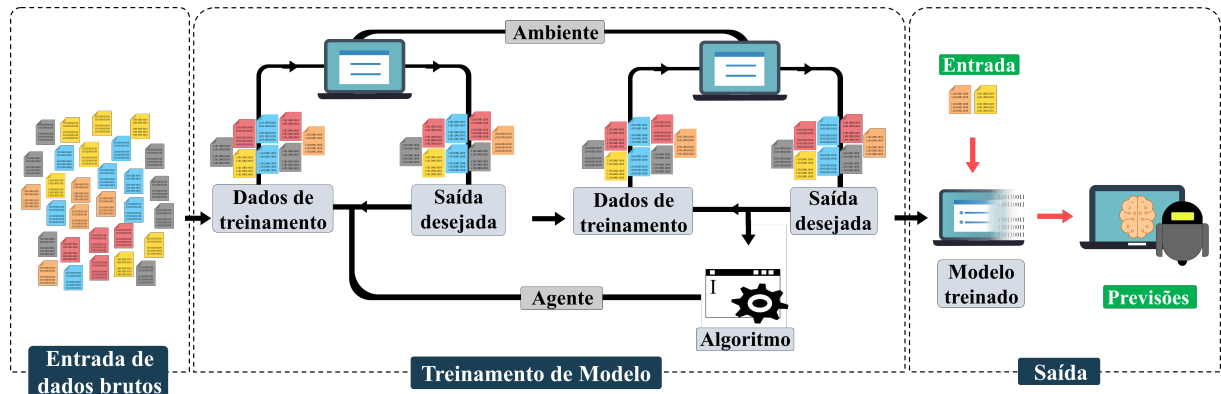
Fonte: Próprio autor (2023)

2.2.3 Aprendizado por Reforço

O AR, pelo contrário, é conhecido como "aprendizado com um crítico", sendo o crítico muito diferente do que um instrutor, no sentido de que o crítico não informa com antecedência o que fazer, apenas o que foi realizado bem no passado (ALPAYDIN, 2020).

Geralmente, este tipo de aprendizado é descrito em função da interação entre algum agente e o entorno, procedimento de análise ilustrado por RODRÍGUEZ (2020) na Figura 16, onde o agente tem a tarefa de aprender e o entorno é o ambiente onde ele está aprendendo, permitindo fornecer informações sobre a qualidade das estratégias e testando em repetições que permitam ser aperfeiçoadas ou possam se encontrar melhores soluções (MARSLAND, 2015).

Figura 16 – Procedimento do Aprendizado por Reforço

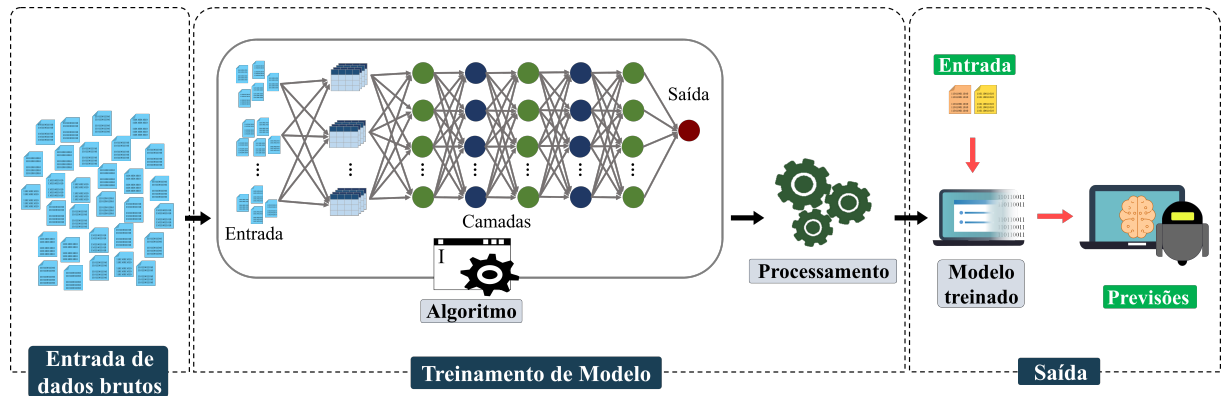


Fonte: RODRÍGUEZ (2020)

2.2.4 Aprendizado Profundo

As aplicações de AP envolvem todo tipo de dados, como estruturados e não estruturados, sejam imagens, áudios, vídeos ou textos (KOTU; DESHPANDE, 2019). Um dos principais componentes do AP são as redes neurais (JAMES *et al.*, 2021) e o processo de análise pode seguir a sequência dada por RODRÍGUEZ (2020), segundo a Figura 17.

Figura 17 – Procedimento do Aprendizado Profundo



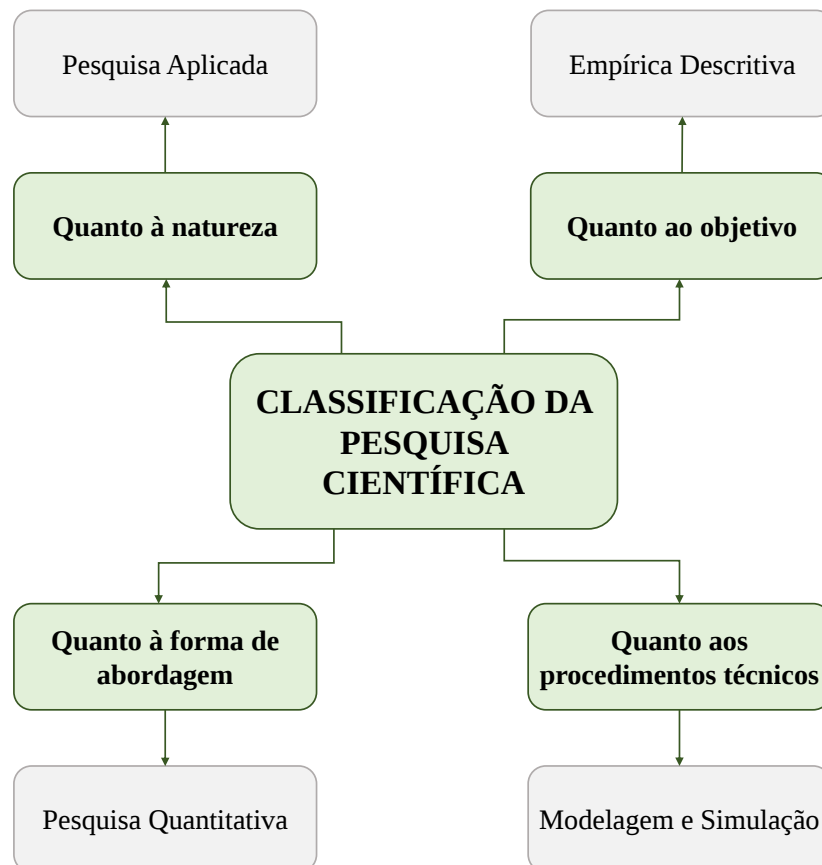
Fonte: RODRÍGUEZ (2020)

Além disso, para encontrar variáveis mais robustas, o AP utiliza o ANS em primeira instância, que podem ser refinadas e utilizadas como preditores por um modelo baseado no AS (DEO, 2015).

3 MATERIAIS E MÉTODOS

Conforme destacado por Bertrand e Fransoo (2002), Cantamessa (2003), Bell e Bryman (2007) e Choi, Cheng e Zhao (2016), este trabalho pode ser classificado como uma pesquisa aplicada, com objetivos empíricos descritivos, pois o modelo desenvolvido visa compreender relações causais que ocorrem na realidade, contribuindo para uma compreensão mais aprofundada dos problemas reais. Além disso, a abordagem utilizada é quantitativa, com o método de pesquisa centrado em modelagem e simulação, conforme ilustrado na Figura 18.

Figura 18 – Classificação da Pesquisa Científica



Fonte: Adaptação de Bertrand e Fransoo (2002), Cantamessa (2003), Bell e Bryman (2007) e Choi, Cheng e Zhao (2016)

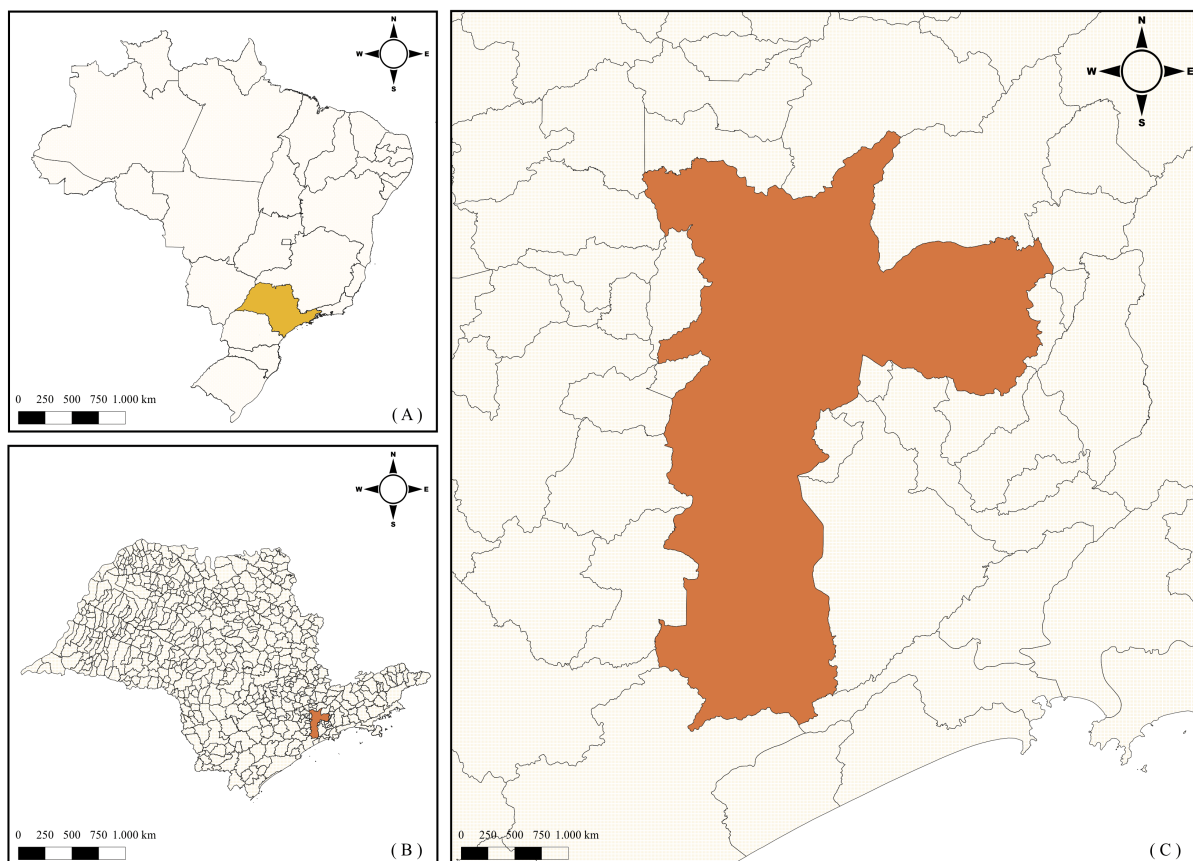
3.1 MATERIAIS

— Área Geográfica:

O município de São Paulo é uma das localidades que integram o estado de São Paulo, abrangendo uma área de 1.521,20km². Com mais de 12 milhões de habitantes, corresponde

a mais de 26% da população estadual e ultrapassa 5% da população total do Brasil, conforme dados do Instituto Brasileiro de Geografia e Estatística - IBGE (2022a). A Figura 19 ilustra o mapa digital da localização geográfica da área em análise, obtido no portal do IBGE (2022b).

Figura 19 – Mapa da localização geográfica do município de São Paulo



Fonte: IBGE, 2022b

— Banco de dados:

Os dados foram disponibilizados pelo Departamento de Informática do Sistema Único de Saúde - DATASUS (2022), mais especificamente do SINASC. Este sistema tem como base a Declaração de Nascido Vivo, visando caracterizar as condições da gestação, do parto e do nascimento (SZWARCOWALD *et al.*, 2019). Foram utilizados dados do período de 2015 a 2019, totalizando inicialmente 836.857 registros. As informações abrangem 60 variáveis relacionadas ao recém-nascido, local de ocorrência, mãe, pai, gestação, parto e possíveis anomalias congênitas no recém-nascido.

Com base no desenvolvimento de modelos para identificar o risco de nascimento prematuro, além de que os fatores mais comuns associados à prematuridade estão relacionados com as informações maternas e as informações da gestação atual ou anterior (LEONE, 2015),

foram consideradas apenas 22 variáveis do conjunto total. Assim, a Tabela 3 apresenta uma breve descrição das variáveis consideradas, organizadas em 4 grupos.

Portanto, a Tabela 3 apresenta uma breve descrição das 12 variáveis consideradas, organizadas em 3 grupos, com base no desenvolvimento de modelos para identificar o risco de nascimento prematuro com base na idade gestacional, considerando os fatores mais comuns associados à prematuridade relacionados às informações maternas e às informações da gestação atual ou anterior (LEONE, 2015).

Tabela 3: Descrição do conjunto de variáveis sobre as informações maternas e da gestação atual ou anterior

Nº	VARIÁVEL	DESCRIÇÃO
Informações maternas		
1	IDADEMAE	Idade da mãe
2	ESTCIVMAE	Estado Civil (situação conjugal da mãe)
3	ESMAE	Escolaridade da mãe (em anos de estudo concluídos)
4	RACACORMAE	Tipo de raça e cor da mãe
Informações da gestação		
5	QTDGESTANT	Número de gestações anteriores
6	QTDFILVIVO	Número de filhos vivos
7	QTDFILMORT	Número de filhos mortos, número de perdas fetais e abortos
8	GRAVIDEZ	Tipo de gravidez
9	MESPENAT	Mês de gestação em que iniciou o pré-natal
10	QTDPARTNOR	Número de partos vaginais anteriores da gravidez atual
11	QTDPARTCES	Número de partos cesáreos anteriores da gravidez atual
Variável alvo		
12	SEMAGESTAC	Número de semanas de gestação (idade gestacional)

Fonte: DATASUS (2022)

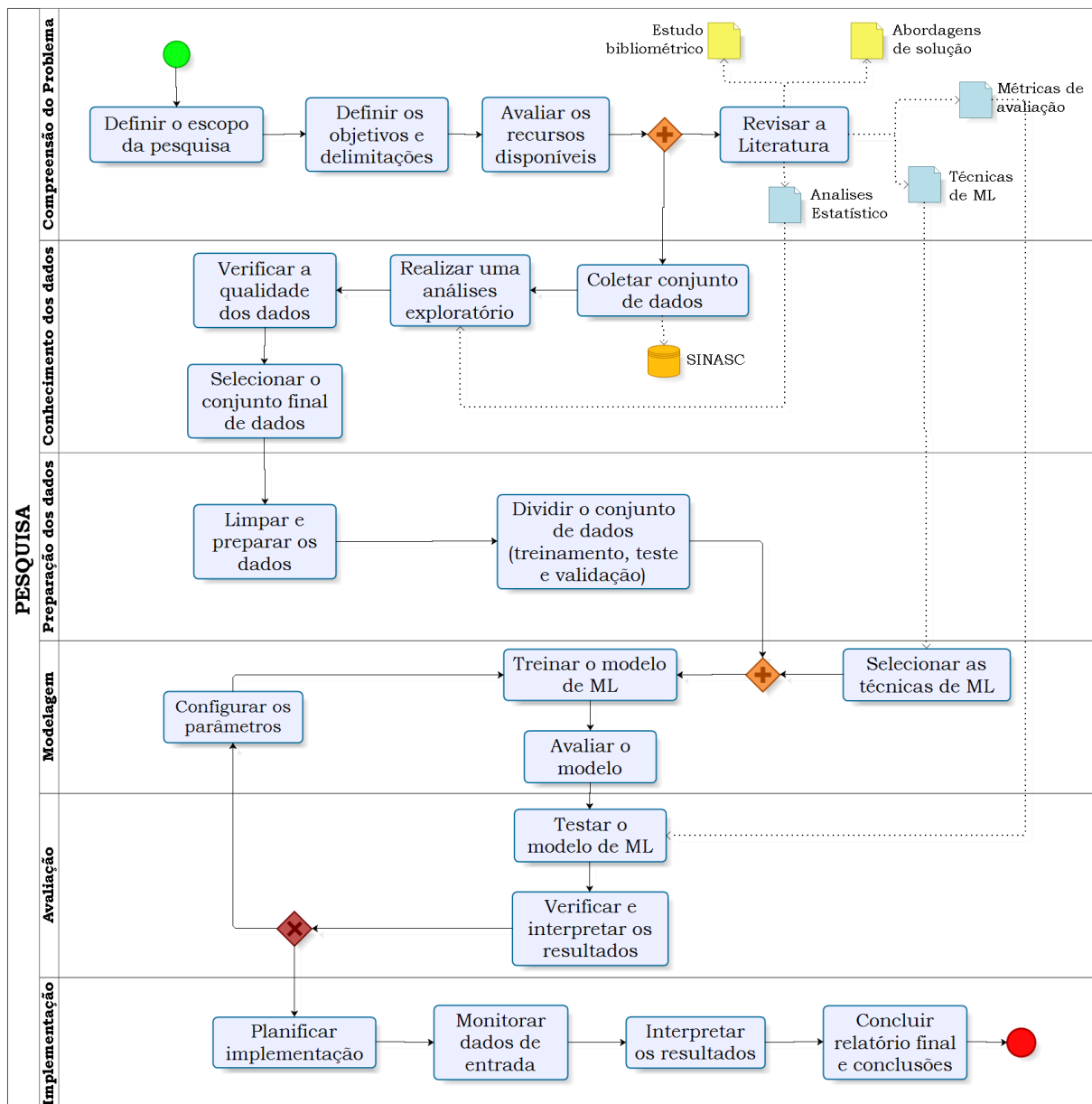
— *Software:*

Utilizou-se o *software* de código aberto Python, por meio da interface Jupyter Notebook, para a análise estatística, o desenvolvimento do modelo preditivo e a criação da interface do usuário.

3.2 MÉTODOS E PROCEDIMENTOS

A Figura 20 exibe o diagrama de processos utilizado no desenvolvimento deste trabalho, uma adaptação da metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), especialmente ajustada para projetos de Aprendizado de Máquina (CHAPMAN *et al.*, 2000; BERTRAND; FOKOUÉ; ZHANG, 2009). O fluxograma é composto por seis fases principais, cada uma delas subdividida em um conjunto de tarefas.

Figura 20 – Fluxo Metodológico



Fonte: Adaptação de Chapman *et al.* (2000) e Bertrand, Fokoué e Zhang (2009)

- Fase 1 – Compreensão do problema:

Inicia-se com a definição do escopo, objetivos e limitações do projeto, juntamente com a avaliação dos recursos disponíveis para o desenvolvimento do projeto. Realiza-se uma

revisão e levantamento da literatura sobre as técnicas de AM a serem utilizadas no modelo preditivo, assim como os métodos estatísticos para avaliação dos dados e do modelo de AM.

- Fase 2 – Conhecimento dos dados:

Envolve a coleta de dados, seguida por uma análise exploratória e descritiva do conjunto, além da verificação da qualidade dos dados. Nesta fase, seleciona-se o conjunto final de dados a ser utilizado nas etapas subsequentes.

- Fase 3 – Preparação dos dados:

O conjunto de dados é preparado utilizando técnicas estatísticas, incluindo identificação e tratamento de dados ausentes. Posteriormente, os dados são divididos para treinamento, teste e validação.

- Fase 4 – Modelagem:

Inclui o desenvolvimento do modelo preditivo, aplicando as técnicas de AM escolhidas de acordo com os objetivos da pesquisa.

- Fase 5 – Avaliação:

Testa-se o modelo com o conjunto de dados de teste, interpretando e verificando os resultados para melhorar o modelo de AM. Este processo cria um ciclo de treinamento, configuração de parâmetros, teste e avaliação.

- Fase 6 – Implementação:

Seleciona-se o modelo final, desenvolvendo a interface do usuário e planejando o monitoramento dos dados de entrada. A interface é testada, e os resultados são interpretados.

3.2.1 Pré-processamento e experimentação

3.2.1.1 Limpeza de dados

Após a coleta inicial de dados e no início do pré-processamento, o conjunto de dados foi refinado, eliminando informações irrelevantes. Foram excluídas observações com dados ausentes ou marcados como ignorados, representados pelos valores 9 e 99 para variáveis categóricas e numéricas, respectivamente. Além disso, observações com dados fora do intervalo $[0,5 ; 99,5]$ foram removidas para evitar a inclusão de *outliers*, que poderiam indicar documentação incorreta e comprometer a confiabilidade das estimativas (SOMANI *et al.*, 2020).

Para a seleção de variáveis, utilizou-se a análise de correlações entre variáveis quantitativas de entrada, como parte do método *filter* (REMESEIRO; BOLON-CANEDO, 2019). A identificação e exclusão de variáveis altamente correlacionadas contribuem para evitar o problema de colinearidade. As variáveis com coeficientes de correlação fora do intervalo $[-0,70 ; +0,70]$ (RODRÍGUEZ, 2020) foram excluídas da análise, garantindo a integridade das estimativas.

3.2.1.2 Transformação e escalonamento de dados

Alguns algoritmos de AM requerem um método apropriado para transformar variáveis compostas por categorias não numéricas em numéricas. Além disso, é crucial observar que variáveis numéricas, como contagens, podem aumentar sem limites, impactando modelos sensíveis durante o treinamento (ZHENG; CASARI, 2018). Portanto, o escalonamento de dados e a normalização podem ser empregados para mitigar problemas em modelos nos quais as variáveis possuem diferentes faixas (JAIN; SHUKLA; WADHVANI, 2018).

Considerando as 12 variáveis relacionadas às informações maternas e da gestação atual e anterior, 41,67% são categóricas. As variáveis ordinais foram transformadas usando a técnica de codificação ordinal (*Ordinal Encoding*) (PACZKOWSKI, 2021; ZHANG *et al.*, 2022), enquanto as variáveis nominais foram convertidas por meio da técnica de codificação *One-Hot* (*One-Hot Encoding*) (JAMES *et al.*, 2021; PACZKOWSKI, 2021). Quanto às variáveis de entrada numéricas, foi aplicado o escalonamento *Min-Max* individualmente, ajustando os valores das variáveis para a faixa [0 ; 1] (PACZKOWSKI, 2021; ZHANG *et al.*, 2022), conforme calculado pela Equação 3.1:

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3.1)$$

onde, x representa o valor original de uma variável, $x_{\min}$ e $x_{\max}$ são, respectivamente, os valores mínimo e máximo, e $\tilde{x}$ denota o valor escalado ou padronizado da variável.

Depois dessa etapa e com base na identificação do risco de parto prematuro por meio da idade gestacional, procedeu-se ao treinamento dos modelos de Regressão Linear, Árvore de Decisão, Floresta Aleatória, *Gradient Boosting*, e *Extreme Gradient Boosting* utilizando uma abordagem de regressão, uma vez que a variável alvo (SEMAGESTAC) é definida como quantitativa. Isso foi realizado por meio de validação cruzada de *10-Fold*, otimizando hiperparâmetros através da técnica de busca aleatória (*Random Search*) para selecionar a melhor combinação de parâmetros, visando melhorar o desempenho dos modelos (ANDONIE, 2019). Além disso, procedeu-se à comparação das diferentes métricas de desempenho para a escolha do melhor modelo.

3.3 ASPECTOS ÉTICOS

No entanto, dado que este estudo foi conduzido utilizando dados de fontes secundárias disponibilizados pelo DATASUS, um órgão de domínio público, não foi necessário submeter o projeto a um Comitê de Ética em Pesquisa. Isso está em conformidade com a Resolução do Conselho Nacional de Saúde nº. 510/16, datada de 7 de abril de 2016.

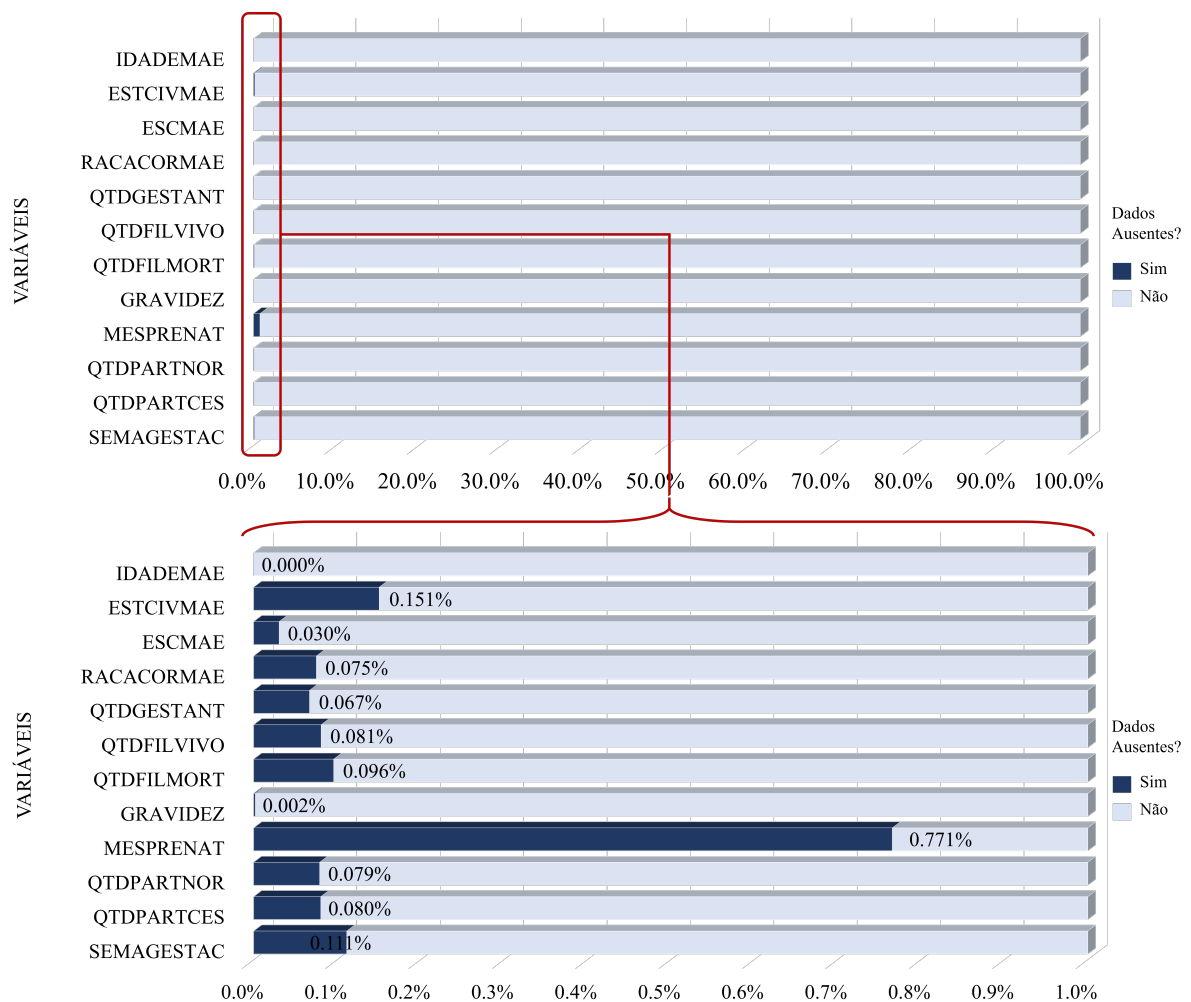
4 RESULTADOS

No período 2015-2019 foram encontrados 836.857 registros sobre os nascidos vivos no município de São Paulo, de acordo com os dados fornecidos pelo SINASC - DATASUS, além das 12 variáveis consideradas sobre as informações maternas e da gestação atual ou anterior.

4.1 ANÁLISE EXPLORATÓRIA DE DADOS E SELEÇÃO DE VARIÁVEIS

Uma primeira análise do conjunto de dados permitiu identificar dados ausentes, o que pode ser um problema no desenvolvimento de modelos de predição.

Figura 21 – Porcentagem de dados ausentes por variável



Fonte: Próprio autor (2023)

A Figura 21 ilustra a porcentagem de dados ausentes por variável, distribuídos ao longo de 12.031 instâncias ou registros de nascidos vivos. Observou-se que a porcentagem de dados ausentes por variável foi inferior a 1%, destacando-se que a variável MESPENAT apresentou o maior percentual (0,771%), e a variável IDADEMAE não apresentou dados faltantes.

Além disso, constatou-se que algumas variáveis continham informações classificadas como ignoradas, com valores de 9 e 99 para variáveis categóricas e numéricas, respectivamente.

Dessa forma, instâncias com valores ausentes ou classificadas como ignoradas foram excluídas. Adicionalmente, para não prejudicar as estimativas dos modelos, foram removidas instâncias com valores fora do intervalo [0,5 ; 99,5] nas variáveis numéricas, considerados como *outliers*.

Assim, um total de 42.161 registros foram eliminados, representando apenas 5,04% do conjunto de dados inicial. Isso possibilitou obter 794.696 registros após a limpeza dos dados, os quais foram utilizados para o desenvolvimento dos modelos de predição.

A Tabela 4 apresenta os valores atribuídos, dependendo do tipo de variável, podendo ser numérico ou categórico.

Tabela 4: Tipo de variável e valores atribuídos

Nº	VARIÁVEL	TIPO	VALOR
Informações maternas			
1	IDADEMAE	Numérico	≥ 0
2	ESTCIVMAE	Categórico	1- Com parceiro (Casada / União Estável) 2- Sem parceiro (Solteira / Viúva / Separada-Divorciada)
3	ESCMAC	Categórico	1- 0 a 3 anos 2- 4 a 7 anos 3- 8 a 11 anos 4- 12 e mais
4	RACACORMAE	Categórico	1- Branca 2- Preta 3- Amarela 4- Parda 5- Indígena
Informações da gestação			
5	QTDGESTANT	Numérico	≥ 0
6	QTDFILVIVO	Numérico	≥ 0
7	QTDFILMORT	Numérico	≥ 0
8	GRAVIDEZ	Categórico	1- Única 2- Múltipla

Fonte: Próprio autor (2023)

Continuado na próxima página

Tabela 4: Tipo de variável e valores atribuídos (Continuado)

Nº	VARIÁVEL	TIPO	VALOR
9	MESPRENAT	Categórico	1- Terceiro trimestre (Inadequado) 2- Segundo trimestre (Intermediário) 3- Primeiro trimestre (Adequado)
10	QTDPARTNOR	Numérico	≥ 0
11	QTDPARTCES	Numérico	≥ 0
Variável alvo			
12	SEMAGESTAC	Numérico	≥ 0

Fonte: Próprio autor (2023)

Como mostrado na Tabela 5, a idade da mãe (IDADEMAE) oscilou de 15 a 44 anos, com uma média de pelo menos 28 anos. O número de gestações anteriores (QTDGESTANT) por mãe variou de 0 a 7, com uma média de aproximadamente 1 gestação anterior. Da mesma forma, o número de filhos vivos (QTDFILVIVO) e de filhos mortos ou perdas fetais/abortos (QTDFILMORT) em gestações anteriores por mãe variou de 0 a 6 e de 0 a 3, respectivamente. O número de partos normais (QTDPARTNOR) e cesáreos (QTDPARTCES) anteriores variou de 0 a 5 e 0 a 3, respectivamente. Também, a variável alvo, idade gestacional (SEMAGESTAC) dos recém-nascidos, variou entre 27 e 42 semanas, com uma média de aproximadamente 39 semanas.

Tabela 5 – Estatísticas descritivas das variáveis numéricas

VARIÁVEL	MÉDIA	DESVIO-PADRÃO	MÍN	Q1	Q2	Q3	MÁX
IDADEMAE	28,2	6,7	15,0	23,0	28,0	33,0	44,0
QTDGESTANT	1,0	1,2	0,0	0,0	1,0	2,0	7,0
QTDFILVIVO	0,8	1,0	0,0	0,0	1,0	1,0	6,0
QTDFILMORT	0,2	0,5	0,0	0,0	0,0	0,0	3,0
QTDPARTNOR	0,5	0,9	0,0	0,0	0,0	1,0	5,0
QTDPARTCES	0,3	0,6	0,0	0,0	0,0	1,0	3,0
SEMAGESTAC	38,5	1,8	27,0	38,0	39,0	40,0	42,0

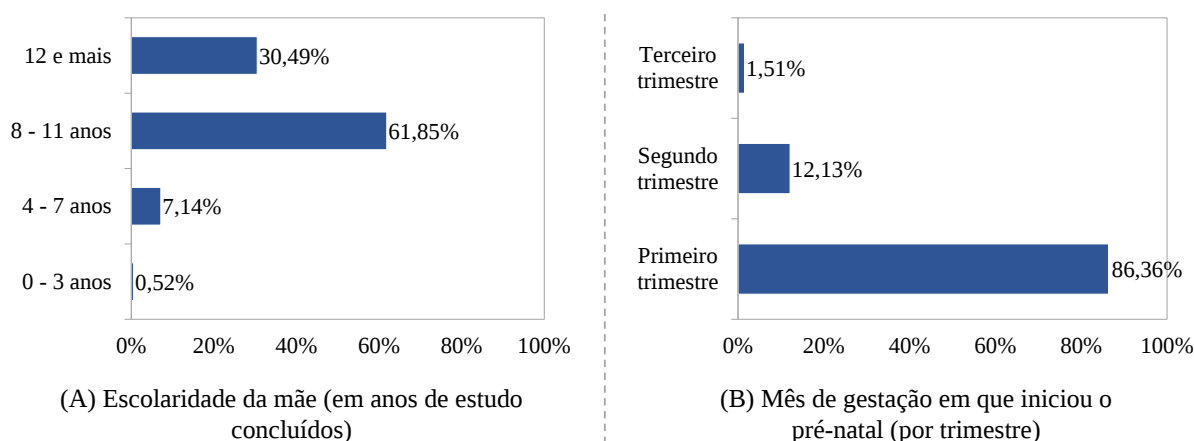
Fonte: Próprio autor (2023)

Por outro lado, as variáveis categóricas podem ser classificadas em 2 tipos: variáveis ordinais (ESCMAE e MESPRENAT) e variáveis nominais (ESTCIVMAE, RACACORMAE e GRAVIDEZ).

A Figura 22 apresenta a análise gráfica das variáveis categóricas ordinais. Com relação à escolaridade da mãe, como mostrado na Figura 22A, observa-se que 30,49% das mães estudaram

por no mínimo por 12 anos, 61,85% estudaram de 8 a 11 anos, 7,14% estudaram de 4 a 7 anos, e os 0,52% restantes estudaram no máximo por 3 anos. Também, com base no mês da gestação em que foi feita a primeira consulta pré-natal (Figura 22B), destaca-se que 86,36% das mães realizaram a consulta no primeiro trimestre, 12,13% no segundo trimestre e 1,51% no terceiro trimestre.

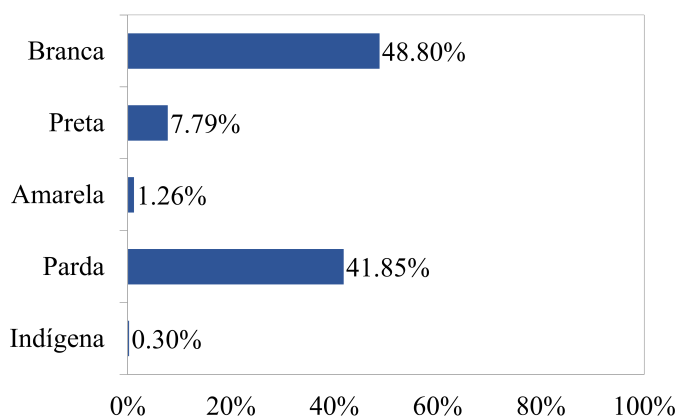
Figura 22 – Variáveis categóricas ordinais



Fonte: Próprio autor (2023)

Dentre as variáveis nominais, destaca-se a raça/cor da mãe, que possui mais de duas categorias. Podemos observar que 48,80% das mães são brancas, 41,85% são pardas, 8,79% são pretas, enquanto amarelas e indígenas representam 1,26% e 0,30%, respectivamente (Figura 23).

Figura 23 – Variável nominal com mais de duas categorias

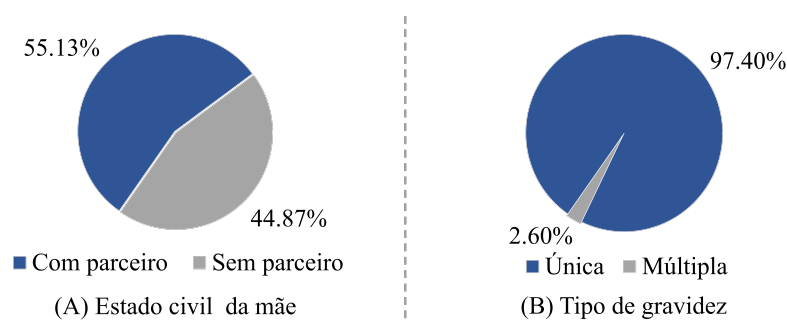


Tipo de raça e cor da mãe

Fonte: Próprio autor (2023)

Da mesma forma, com base nas variáveis nominais de apenas duas categorias observou-se que 55,13% das mães têm parceiro (Figura 24A) e 97,40% apresentaram uma gravidez única (Figura 24B).

Figura 24 – Variáveis nominais de duas categorias



Fonte: Próprio autor (2023)

Além disso, a Tabela 6 mostra as características das informações maternas e da gestação atual ou anterior, de acordo com a classificação de nascidos vivos pré-termo e à termo.

Tabela 6: Características sobre os nascidos vivos pré-termo e à termo (n = 794.696)

VARIÁVEL	PRÉ-TERMO (n = 76.925)	À TERMO (n=717.771)	P-VALOR
Informações maternas			
Idade da mãe (anos), m ± dp	28,7 ± 7,0	28,1 ± 6,6	< 0.001 ^a
Estado civil da mãe, n (%)			< 0.001 ^b
Com parceiro	42.884 (55,75%)	395.230 (55,06%)	
Sem parceiro	34.041 (44,25%)	322.541 (44,94%)	
Escolaridade da mãe (em anos de estudo concluídos), n (%)			< 0.001 ^b
0 a 3 anos	453 (0,59%)	3.658 (0,50%)	
4 a 7 anos	5.746 (7,47%)	51.007 (7,11%)	
8 a 11 anos	46.073 (59,89%)	445.422 (62,06%)	
12 a mais	24.653 (32,05%)	217.684 (30,33%)	
Tipo de raça e cor da mãe, n (%)			< 0.001 ^b
Branca	38.966 (50,65%)	348.853 (48,60%)	
Preta	5.966 (7,76%)	55.918 (7,79%)	
Amarela	889 (1,16%)	9.141 (1,27%)	
Parda	30.932 (40,21%)	301.657 (42,03%)	
Indígena	172 (0,22%)	2.202 (0,31%)	

m: Média; dp: Desvio-Padrão; ^a t de Student; ^b Qui-quadrado; ^c U de Mann-Whitney

Fonte: Próprio autor (2023)

Continuado na próxima página

Tabela 6: Características sobre os nascidos vivos pré-termo e à termo (n = 794.696) (Continuado)

VARIÁVEL	PRÉ-TERMO (n = 76.925)	À TERMO (n=717.771)	P-VALOR
Informações da gestação			
Número de gestações anteriores, m $\pm$ dp	1,1 $\pm$ 1,3	1,0 $\pm$ 1,2	< 0.001 ^c
Número de filhos vivos, m $\pm$ dp	0,8 $\pm$ 1,1	0,8 $\pm$ 1,0	< 0.001 ^c
Número de filhos mortos, número de perdas fetais e abortos, m $\pm$ dp	0,3 $\pm$ 0,6	0,2 $\pm$ 0,5	< 0.001 ^c
Tipo de gravidez, n (%)			< 0.001 ^b
Única	64.445 (83,78%)	709.615 (98,86%)	
Múltipla	12.480 (16,22%)	8.156 (1,14%)	
Mês de gestação em que iniciou o pré-natal (por trimestre), n (%)			< 0.001 ^b
1ro trimestre	65.286 (84,87%)	621.030 (86,52%)	
2do trimestre	10.384 (13,50%)	85.985 (11,98%)	
3ro trimestre	1.255 (1,63%)	10.756 (1,50%)	
Número de partos vaginais anteriores da gravidez atual, m $\pm$ dp	0,6 $\pm$ 1,0	0,5 $\pm$ 0,9	< 0.001 ^c
Número de partos cesáreos anteriores da gravidez atual, m $\pm$ dp	0,3 $\pm$ 0,6	0,3 $\pm$ 0,6	< 0.001 ^c

m: Média; dp: Desvio-Padrão; ^a t de Student; ^b Qui-quadrado; ^c U de Mann-Whitney

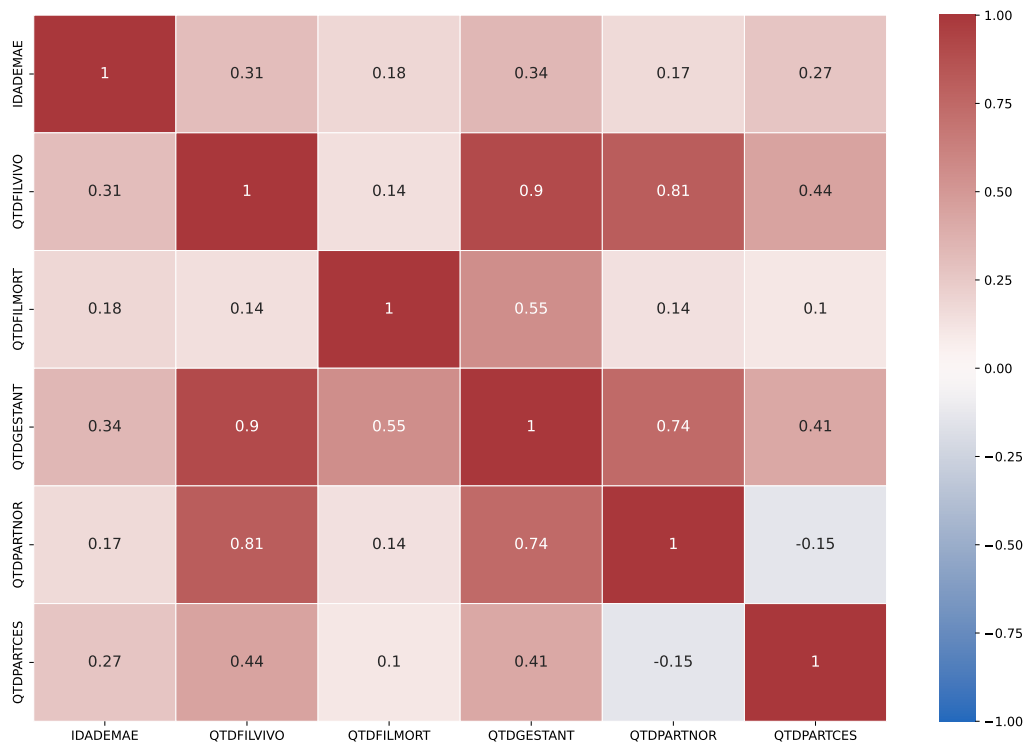
Fonte: Próprio autor (2023)

Na Tabela 6 observou-se que a idade média das mães que tiveram parto prematuro foi de aproximadamente 29 anos. A maioria delas coabitava com um parceiro (55,75%), tinha concluído mais de 8 anos de escolaridade (91,94%), e pertencia predominantemente às categorias de raça branca (50,65%) ou parda (40,21%). Além disso, observou-se que essas mães relataram, em média, pelo menos uma gravidez anterior, aproximadamente um nascimento vivo, e nenhuma perda fetal ou aborto.

Também, a maioria das mães que teve parto prematuro apresentou gravidez única (83,78%), realizou a primeira consulta pré-natal no primeiro trimestre (84,87%), e tinha uma média de partos vaginais e cesarianas anteriores de aproximadamente 1 e 0, respectivamente (Tabela 6).

Adicionalmente, uma análise crucial antes do treinamento dos modelos de AM é a identificação de variáveis de entrada altamente correlacionadas. A presença dessas correlações pode resultar em problemas de colinearidade, comprometendo a confiabilidade das estimativas. A Figura 25 ilustra as correlações (r) entre as variáveis de entrada quantitativas no conjunto de dados.

Figura 25 – Correlações entre variáveis de entrada quantitativas



Fonte: Python (2023)

Assim, na Figura 25, pode-se observar uma significativamente alta (fora do limite $[-0,70 ; +0,70]$) entre as variáveis QTDFILVIVO e QTDGESTANT ($r = 0,90$), QTDFILVIVO e QTDPARTNOR ($r = 0,81$), e QTDGESTANT e QTDPARTNOR ($r = 0,74$), indicando a presença de colinearidade. Diante disso, optamos por excluir as variáveis QTDGESTANT e QTDFILVIVO, resultando em novas correlações (r) dentro do limite de $\pm 0,70$ (Figura 26).

4.2 PRE-PROCESSAMENTO, EXPERIMENTAÇÃO E AVALIAÇÃO

Nesta seção, adotou-se a abordagem de regressão para identificar o risco de parto prematuro, estimando a idade gestacional com base em informações pré-natais, incluindo dados sobre a mãe e a gestação atual ou anterior. O processo de treinamento e validação dos modelos envolveu dados do período de 2015 a 2018 (643.325 registros de nascidos vivos), enquanto os dados de 2019 (151.371 registros de nascidos vivos) foram reservados para o teste e a avaliação do modelo.

Figura 26 – Correlações entre variáveis de entrada, excluindo QTDGESTANT e QTDFILVIVO



Fonte: Python (2023)

4.2.1 Abordagem de regressão para estimar a idade gestacional

Esta abordagem foi conduzida com 10 variáveis relacionadas às informações maternas e à gestação atual ou anterior, excluindo as variáveis QTDGESTANT e QTDFILVIVO. Incluiu-se também a variável alvo referente ao número de semanas de gestação ou idade gestacional (SEMAGESTAC), conforme indicado nas Tabelas 3 e 4.

As variáveis categóricas nominais (ESTCIVMAE, RACACORMAE e GRAVIDEZ) foram decompostas em variáveis *dummy* usando a técnica *One-Hot encoding*, enquanto as variáveis categóricas ordinais (ESCMAE e MESPENAT) foram transformadas com a técnica *Ordinal encoding*. Além disso, para normalizar as variáveis numéricas (IDADEMAE, QTDFILMORT, QTDPARTNOR e QTDPARTCES), foi aplicada a escala *Min-Max* para evitar que possíveis diferenças nos intervalos das variáveis impactassem o desempenho dos modelos de predição.

Dessa forma, durante o processo de treinamento, foram aplicados cinco modelos de AM, tais como: Regressão Linear, Árvore de Decisão, Floresta Aleatória, *Gradient Boosting*, e *Extreme Gradient Boosting*. Os modelos passaram por validação interna usando a técnica de validação cruzada de *10-Fold*, sendo ajustados para obter a melhor configuração de hiperparâmetros. A busca por esses hiperparâmetros foi realizada por meio da técnica de busca aleatória (*Random Search*), cujos resultados estão apresentados na Tabela 7.

Durante o processo de teste dos modelos, foram comparadas diversas métricas de desempenho, incluindo o Erro Médio Absoluto (EMA), o Erro Quadrático Médio (EQM), a Raiz do Erro

Tabela 7 – Otimização de hiperparâmetros dos modelos de regressão

Modelo	Hiperparâmetros
Regressão Linear	-
Árvore de Decisão	<i>splitter = 'random', min_samples_split = 2, min_samples_leaf = 9</i>
Floresta Aleatória	<i>n_estimators = 84, min_samples_split = 6, min_samples_leaf = 9, max_features = 'sqrt', bootstrap = True</i>
<i>Gradient Boosting</i>	<i>n_estimators = 89, max_leaf_nodes = 5, max_features = 'auto', max_depth = 10, learning_rate = 1</i>
<i>Extreme Gradient Boosting</i>	<i>n_estimators = 75, min_child_weight = 100, max_depth = 8, learning_rate = 0.1, booster = 'gbtree'</i>

Fonte: Próprio autor (2023)

Quadrático Médio (REQM) e o Erro Percentual Absoluto Médio (EPAM). Os valores dessas métricas para os modelos testados estão apresentados na Tabela 8.

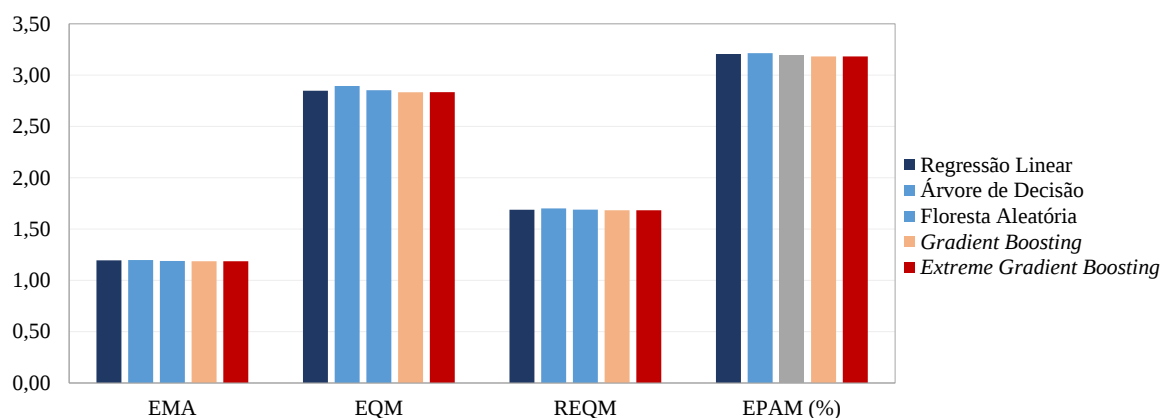
Tabela 8 – Valores das métricas de desempenho dos modelos de regressão

Modelo	EMA	EQM	REQM	EPAM (%)
Regressão Linear	1,195	2,848	1,688	3,206
Árvore de Decisão	1,198	2,894	1,701	3,214
Floresta Aleatória	1,189	2,853	1,689	3,190
<i>Gradient Boosting</i>	1,186	2,833	1,683	3,182
<i>Extreme Gradient Boosting</i>	1,186	2,834	1,683	3,182

Fonte: Próprio autor (2023)

Conforme indicado na Tabela 8 e na Figura 27, observa-se uma pequena diferença entre os modelos avaliados. Entretanto, destaca-se que o modelo de Árvore de Decisão obteve os valores mais altos nas métricas de desempenho, enquanto os modelos *Gradient Boosting* e *Extreme Gradient Boosting* apresentaram os valores mais baixos.

Figura 27 – Valores das métricas de desempenho dos modelos de regressão

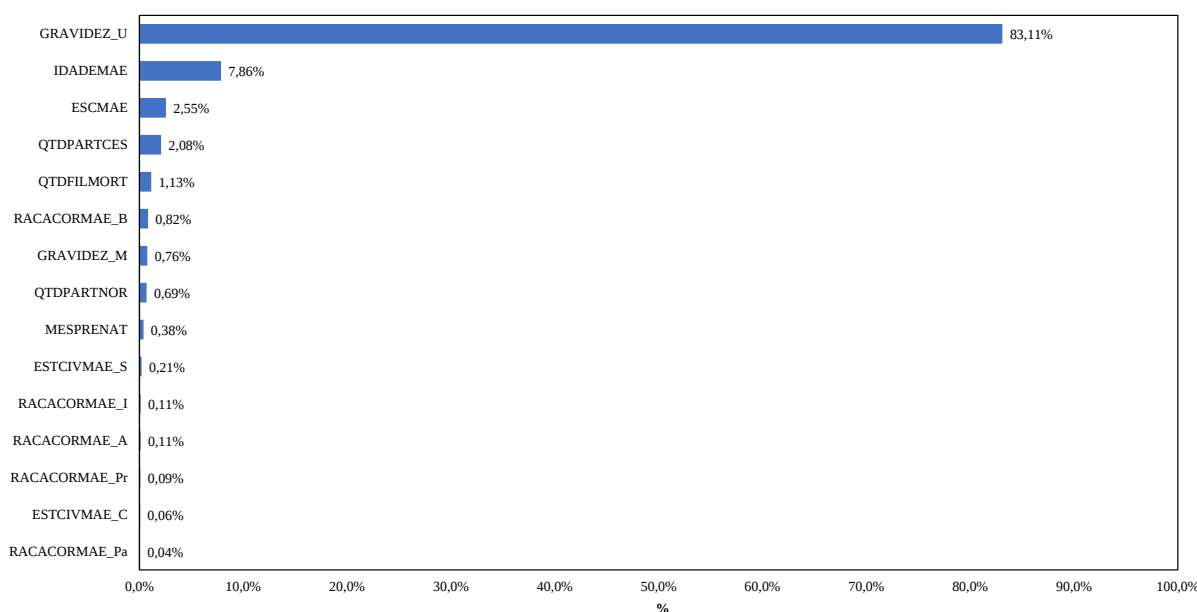


Fonte: Próprio autor (2023)

A comparação das métricas de desempenho permitiu selecionar o modelo com os menores valores como o melhor modelo, e, nesse contexto, o modelo escolhido foi o *Gradient Boosting*. Este modelo, por ser mais simples que o *Extreme Gradient Boosting*, apresentou $EMA = 1,186$, $EQM = 2,833$, $REQM = 1,683$ e $EPAM = 3,182$.

Em seguida, ao avaliar a importância das variáveis no modelo *Gradient Boosting* selecionado, observou-se que o tipo de gravidez única (GRAVIDEZ_U) e a idade da mãe (IDADEMAE) foram as principais variáveis, com a maior importância relativa. Estas foram seguidas pela escolaridade da mãe (ESCMAE), o número de partos cesáreos anteriores (QTDPARTCES), o número de filhos mortos anteriores da gestação atual (QTDFILMORT) e a raça branca (RACACORMAE_B) (ver Figura 28).

Figura 28 – Importância das variáveis no modelo de *Gradient Boosting* selecionado para regressão



Fonte: Próprio autor (2023)

4.3 IMPLEMENTAÇÃO DO MODELO SELECIONADO

Na fase de implementação, foi criada uma interface de usuário (Figura 29) com base no modelo avaliado e selecionado como o melhor neste estudo. Essa interface pode ser utilizada pelos especialistas da área da saúde como suporte na tomada de decisões.

A interface de usuário desenvolvida é composta por 2 colunas. A coluna do lado esquerdo contém as características que devem ser preenchidas pelo especialista com base nas informações maternas e da gestação atual ou anterior, no momento em que uma gestante está realizando o pré-natal, permitindo assim prever uma estimativa da idade gestacional (em semanas) para o parto. Por outro lado, a coluna do lado direito exibe os resultados das estimativas realizadas, considerando o modelo de regressão selecionado para prever a idade gestacional.

Figura 29 – Interface de usuário desenvolvida para estimar a idade gestacional

Fonte: Próprio autor (2023)

Para ilustrar e verificar os resultados obtidos neste trabalho, foi utilizada a informação aleatória de apenas 2 registros da declaração de nascidos vivos no município de São Paulo para o ano de 2020.

Figura 30 – Estimação da idade gestacional - Caso 1

Fonte: Próprio autor (2023)

No primeiro caso, fazendo uma simulação com base nas informações coletadas, temos uma mãe gestante que iniciou o pré-natal no segundo trimestre, informando ter 24 anos, coabitando com um parceiro e sendo de estado civil, apresentando 12 ou mais anos de estudos concluídos, e se identificando como parda. Além disso, a gestante não teve filhos mortos anteriores, apresentou gravidez atual única, e não tem histórico de ter realizado parto vaginal ou cesáreo antes da gestação atual. Assim, a Figura 30 ilustra a utilização da interface de acordo com as características

apontadas anteriormente, estimando como resultado que a gestação dure 39 semanas, um resultado muito próximo do valor real, já que o parto foi realizado na semana 38 de gestação.

No segundo caso, temos outra gestante que iniciou o pré-natal no primeiro trimestre, informando ter 34 anos, estado civil de coabitação, entre 8 a 12 anos de estudos concluídos e identificando-se como raça branca. A gestante indicou não ter filhos mortos e nem ter realizado parto cesáreo anteriormente, porém, registrou a realização de parto vaginal antes da gestação atual e apresentou gravidez múltipla. A Figura 31 ilustra a estimativa da duração da gestação em 36 semanas (premature), resultado que apresentou uma diferença notória com o valor real de 27 semanas. É possível que outros fatores externos tenham contribuído para o nascimento ocorrer muito antes.

Figura 31 – Estimativa da idade gestacional - Caso 2

INTERFACE PARA A IDENTIFICAÇÃO DO RISCO DE PARTO PREMATURO

Idade da mãe
(anos)
0 34 100

Estado Civil da mãe
☐ Com parceiro (Casada / União Estável)
☒ Sem parceiro (Solteira / Viúva / Separada-Divorciada)

Escolaridade da mãe
(em anos de estudo concluídos)
 8 a 11 anos

Tipo de raça e cor da mãe
☒ Branca
☐ Preta
☐ Amarela
☐ Parda

RESULTADO
ESTIMAÇÃO DA IDADE GESTACIONAL
 Estima-se que a gestação dure: 36 semanas, sendo classificada como Pré-termo

Fonte: Próprio autor (2023)

Dessa forma, com base no modelo de AM selecionado devido aos bons resultados apresentados, a ferramenta desenvolvida oferece estimativas valiosas que servirão de suporte na tomada de decisões dos especialistas em saúde materno-infantil. Além disso, ela destaca a importância de possíveis melhorias no sistema social e de saúde, promovendo a implementação de recursos direcionados para aprimorar as condições de vida e a qualidade da assistência fornecida às mães e aos recém-nascidos. Esse enfoque proativo visa não apenas identificar riscos precocemente, mas também contribuir para a implementação de medidas preventivas e intervenções oportunas, visando um impacto positivo na saúde da população.

5 DISCUSSÕES

Este trabalho analisou dados sobre os registros de nascidos vivos no município de São Paulo, no período de 2015 a 2019, adquiridos por meio do SINASC - DATASUS. Foram consideradas apenas variáveis relacionadas às informações maternas e da gestação atual ou anterior para estimar a idade gestacional antes do parto utilizando métodos de regressão. Cinco algoritmos de Aprendizado de Máquina (AM) foram treinados e testados, tais como: Regressão Linear, Árvore de Decisão, Floresta Aleatória, *Gradient Boosting*, e *Extreme Gradient Boosting*, usando validação cruzada de *10-Fold* e otimização de hiperparâmetros. Comparando as diferentes métricas de desempenho, o modelo de *Gradient Boosting* foi selecionado como o melhor.

A prematuridade é considerada como um problema epidemiológico que afeta os recém-nascidos, associando-se a diversas complicações físicas e neuropsicomotoras (RODRÍGUEZ *et al.*, 2024). Entre essas complicações, destacam-se síndromes de desconforto respiratório, doenças pulmonares crônicas, lesões intestinais, doenças cardiovasculares, problemas de audição e visão, bem como lesões neurológicas. Além disso, a prematuridade é identificada como uma das principais causas de mortalidade neonatal (BEHRMAN; BUTLER, 2007).

Estima-se que, no Brasil, o custo aproximado associado aos nascimentos prematuros alcance cerca de R\$8 bilhões, podendo aumentar devido ao tempo que os recém-nascidos prematuros podem permanecer na unidade de terapia intensiva (REDE NACIONAL PRIMEIRA INFÂNCIA, 2019). Diante desse cenário, o diagnóstico precoce de partos prematuros desempenha um papel crucial na prevenção de complicações ao longo da vida dos recém-nascidos, contribuindo para evitar mortes prematuras. Nesse sentido, para prevenir ou controlar essa situação, a intervenção do governo e de alguns especialistas relacionados à área médica é necessária (TASCÓN *et al.*, 2016; RODRÍGUEZ *et al.*, 2024).

Para auxiliar na tomada de decisões dos especialistas, diversas técnicas, como a Aprendizagem de Máquina (AM), emergiram como recursos valiosos na área médica (VOLLMER *et al.*, 2020). Essas técnicas desempenham um papel crucial no diagnóstico, controle e redução de nascimentos prematuros. Assim, diversos estudos têm utilizado métodos de AM para estimar a idade gestacional de recém-nascidos, principalmente por meio de modelos de regressão (SAZAWAL *et al.*, 2021; ROCHA *et al.*, 2021; SAZAWAL *et al.*, 2022; SHEN *et al.*, 2022; REIS *et al.*, 2022).

Em alguns desses estudos, identificou-se a estimativa da idade gestacional por meio do processamento de imagens. Sazawal *et al.* (2021) analisaram dados de ultrassonografia na triagem metabólica de 1.318 recém-nascidos no sul da Ásia e na África Subsaariana, optando pelo modelo de Floresta Aleatória como o mais eficaz, com resultados de EMA = 0,74 e REQM = 1,02. Adicionalmente, Shen *et al.* (2022) investigaram 741 exames de ressonância magnética ponderados em T2, concluindo que o modelo *Convolutional Neural Network* (CNN) apresentou o desempenho mais superior, com EMA = 6,7 (dias).

No contexto brasileiro, encontrou-se que Reis *et al.* (2022) adotaram um modelo fotobiológico de maturidade da pele em recém-nascidos prematuros, associado a dados clínicos para a estimativa da idade gestacional. Dessa forma, utilizando recursos relacionados à terapia corticosteróide pré-natal para maturidade pulmonar fetal e refletância da pele, constatou-se que o modelo *Extreme Gradient Boosting* apresentou um desempenho satisfatório, com EMA = 1,688 semanas. Adicionalmente, ao incorporar o peso ao nascer como variável preditora, o modelo demonstrou uma melhoria no desempenho, atingindo EMA = 1,147 semanas (REIS *et al.*, 2022).

Além disso, foi constatado que Rocha *et al.* (2021) examinaram dados provenientes de 3.876.666 registros de nascidos vivos distribuídos em 3.929 municípios brasileiros. Além disso, foi constatado que Rocha *et al.* (2021) examinaram dados provenientes de 3.876.666 registros de nascidos vivos distribuídos em 3.929 municípios brasileiros. Os resultados apontaram que o modelo *Extreme Gradient Boosting* apresentou o melhor desempenho, com REQM = 2,094 (ROCHA *et al.*, 2021).

Assim, foi possível observar que a utilização de imagens para testar algoritmos de Aprendizado de Máquina (AM) resultou em desempenho satisfatório, conforme relatado em alguns estudos (SAZAWAL *et al.*, 2021; SHEN *et al.*, 2022; SAZAWAL *et al.*, 2022).

No entanto, o modelo de regressão *Gradient Boosting* selecionado neste estudo apresentou desempenho superior em comparação com o modelo avaliado por Rocha *et al.* (2021). Vale ressaltar a diferença no conjunto de variáveis preditoras, uma vez que o estudo anterior incorporou informações relacionadas aos procedimentos de assistência pré-natal e à infraestrutura física em unidades de atenção primária. Além disso, o modelo *Gradient Boosting* selecionado demonstrou melhor desempenho em relação ao modelo avaliado por Reis *et al.* (2022), mesmo após a exclusão da variável peso ao nascer.

Sob outra perspectiva, é crucial destacar a complexidade da relação entre múltiplos fatores associados ao risco de um recém-nascido ser prematuro. Isso abrange determinantes sociais, econômicos e de serviços de saúde, bem como determinantes gestacionais, incluindo informações sobre parto prematuro anterior, gestações múltiplas, entre outros. Além disso, fatores maternos, como idade, estado civil, nível de escolaridade ou raça/etnia, desempenham papel fundamental nessa complexa rede de influências (TASCÓN *et al.*, 2016). Outros fatores relacionados à prematuridade incluem o baixo peso ao nascer (RODRÍGUEZ *et al.*, 2023) e valores anormais no índice APGAR, frequentemente associados ao óbito neonatal (OLIVEIRA *et al.*, 2012; QUEIROZ; GOMES; MOREIRA, 2018).

Neste estudo, identificou-se que informações maternas, como idade, escolaridade e etnia branca, além de informações relacionadas à gestação, como o tipo de gravidez única e o número de partos cesáreos anteriores, destacaram-se como características com maior importância relativa. Esses resultados estão em consonância com a literatura, que relata que determinantes maternos e gestacionais, tanto atuais quanto anteriores, são fatores de risco comuns fortemente associados a partos prematuros (LEONE, 2015; TASCÓN *et al.*, 2016).

Por outro lado, a implementação de ferramentas computacionais, como interfaces de usuário,

possibilita com maior frequência a realização de diagnóstico ou controle de doenças, além do monitoramento médico dos pacientes (MOREIRA *et al.*, 2018; ROCHA *et al.*, 2022; NAKAGAWA; ANTONINO, 2023). Essas ferramentas desempenham um papel crucial na moderna prática médica, oferecendo uma abordagem mais eficiente e personalizada para o acompanhamento de pacientes.

Assim, a elaboração da interface neste estudo não apenas visa contribuir para um monitoramento oportuno das gestantes, mas também representa um avanço na integração de métodos computacionais na área médica. Ao fornecer informações em tempo real e ferramentas analíticas, a interface desenvolvida pode servir como um valioso suporte aos especialistas na área médica para a identificação do risco de prematuridade, permitindo intervenções preventivas e personalizadas.

6 CONCLUSÕES

Com base nos resultados gerados nesta pesquisa, é fundamental ressaltar que este estudo não visa substituir o trabalho dos especialistas da área da saúde. Pelo contrário, o uso de algoritmos de AM para a estimação da duração da gestação ou a identificação de recém-nascidos prematuros pode desempenhar um papel crucial como suporte às decisões dos especialistas. Essa abordagem busca agilizar a tomada de ações em situações envolvendo mães e recém-nascidos em risco.

Dessa forma, os resultados obtidos podem fornecer alertas importantes para os governos, destacando a necessidade de melhorias nos serviços de saúde. Além disso, contribuem para a implementação de programas sociais com estratégias de intervenção destinadas à prevenção, controle e redução dos partos prematuros.

Assim, as respostas dadas para as questões de pesquisa e objetivos propostos neste trabalho nas Seções 1.2 e 1.3, são apresentadas a seguir:

- Como desenvolver um modelo preditivo para a tomada de decisão com Aprendizado de Máquina para a identificação de mães com risco de parto prematuro?

O procedimento para desenvolver o modelo preditivo foi detalhado neste trabalho. É essencial destacar que foi utilizada uma abordagem com técnicas e/ou algoritmos de regressão para estimar a duração da gestação em semanas (resultados contínuos). Também, o software de código aberto Python, por meio da interface Jupyter Notebook, desempenhou um papel fundamental como ferramenta computacional para o desenvolvimento do modelo, visualização de dados e elaboração da interface, graças à sua ampla variedade de funções.

- Quais dados são relevantes para este estudo e como devem ser analisados e tratados?

Os dados utilizados foram obtidos do DATASUS por meio do SINASC, contendo os registros da declaração de nascidos vivos. Durante a análise, observou-se a presença de instâncias com valores ausentes ou classificadas como ignoradas, bem como instâncias com valores considerados como *outliers* (fora do intervalo [0,5 ; 99,5]), representando 5,04% do total de registros. Esses *outliers* foram removidos para evitar a geração de estimativas pouco confiáveis.

- Quais variáveis de entrada devem ser consideradas no modelo preditivo?

Dado que a análise segue a sequência para estimar a duração da gestação com base nos métodos de regressão, foram empregadas como variáveis preditoras as características relacionadas às informações maternas e à gestação atual ou anterior.

- Quais algoritmos de Aprendizado de Máquina Supervisionado são adequados para o desenvolvimento do modelo preditivo?

Dentre os algoritmos clássicos, a Regressão Linear foi utilizada. Além disso, a abordagem contou com o uso do algoritmo de Árvore de Decisão, além dos algoritmos de *ensemble* baseados em múltiplas árvores, como Floresta Aleatória, *Gradient Boosting* e *Extreme Gradient Boosting*. Esses métodos foram escolhidos pela robustez demonstrada em previsões anteriores discutidas nesta pesquisa.

- Qual seria o impacto da integração de um sistema de tomada de decisão na avaliação do risco de uma mãe ter parto prematuro?

A integração de um sistema de tomada de decisão, como a utilização de uma interface, pode oferecer estimativas que servirão de suporte às decisões dos especialistas da área médica. Isso prepara esses profissionais para agir prontamente na presença de mães em risco de parto prematuro. Além disso, o impacto da utilização pode ser observado em todos os níveis, seja municipal, regional ou nacional, pois pode alertar os governos sobre a situação, incentivando melhorias nos sistemas social e de saúde destinados a mães e recém-nascidos.

- Como validar os resultados obtidos por meio desta abordagem?

Ao validar os resultados obtidos, é importante destacar que os dados do ano 2015 até 2018 foram utilizados para o treinamento, enquanto os dados do ano de 2019 foram empregados para testar os modelos. A validação foi realizada através da comparação das diferentes métricas de desempenho. Outra forma de validar os resultados é o uso da interface com novos dados, como descrito nos resultados deste trabalho, evidenciando previsões muito próximas aos valores reais.

Como limitação deste estudo, o uso de dados de fonte secundária, como o DATASUS, pode comprometer a qualidade dos dados devido à presença de valores incompletos ou preenchidos incorretamente, indicando uma oportunidade de melhoria nas fontes de informação disponibilizadas.

Finalmente, o modelo de AM de regressão testado neste trabalho apresentou um bom desempenho, destacando a utilidade da técnica de AM como uma ferramenta valiosa na área da saúde para a tomada oportuna de decisões em relação à identificação de partos prematuros. Dessa forma, é possível implementar medidas de controle visando a redução de nascimentos prematuros, alinhando-se ao objetivo 3 de Saúde e Bem-Estar dos Objetivos de Desenvolvimento Sustentável (ODS).

6.1 SUGESTÕES PARA FUTURAS PESQUISAS

- Treinar e testar outros modelos de AM, incluindo abordagens de Inteligência Artificial, como redes neurais.

- Explorar métodos alternativos de reamostragem (validação) e otimização de hiperparâmetros para os modelos.
- Ampliar a inclusão de dados mais recentes na análise.
- Estender a aplicação do estudo para outras cidades do Brasil.
- Adotar a metodologia proposta para abordar outros potenciais problemas relacionados ao parto.

REFERÊNCIAS

- ADRIANO, A. P. S. *et al.* Neonatal mortality related to prematurity. **Research, society and development**, Vargem Grande Paulista, v. 11, n. 4, p. e27511421565, 2022. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/21565/23906>. Acesso em: 20 abr. 2022.
- ALPAYDIN, E. **Introduction to machine learning**. 4th. ed. Massachusetts: Mit Press, 2020.
- ANDONIE, R. Hyperparameter optimization in learning systems. **Journal of membrane computing**, Singapore, v. 1, n. 4, p. 279–291, 2019. Disponível em: <https://link.springer.com/article/10.1007/s41965-019-00023-0>. Acesso em: 24 mar. 2023.
- BEHRMAN, R. E.; BUTLER, A. S. **Preterm birth: causes, consequences, and prevention**. Washington: National Academies Press (US), 2007. Disponível em: <https://nap.nationalacademies.org/catalog/11622/preterm-birth-causes-consequences-and-prevention>. Acesso em: 16 jun. 2022.
- BELL, E.; BRYMAN, A. The ethics of management research: an exploratory content analysis. **British journal of management**, Chichester, v. 18, n. 1, p. 63–77, 2007. Disponível em: https://www.researchgate.net/publication/227792535_The_Ethics_of_Management_Research_An_Exploratory_Content_Analysis. Acesso em: 21 abr. 2021.
- BERGER, A. Z. *et al.* Parto prematuro: características das gestantes de uma população da zona sul de São Paulo. **Revista brasileira de saúde materno infantil**, Recife, v. 16, n. 4, p. 427–435, 2016. Disponível em: <https://www.scielo.br/j/rbsmi/a/qydbLr63LzHVF6w4nZH8P4B/?lang=pt>. Acesso em: 20 jan. 2022.
- BERRAR, D. Cross-validation. **Encyclopedia of bioinformatics and computational biology**, Cambridge, v. 1, p. 542–545, 2019. Disponível em: https://www.researchgate.net/publication/324701535_Cross-Validation. Acesso em: 01 set. 2022.
- BERTRAND, C.; FOKOUÉ, E.; ZHANG, H. H. **Principles and theory for data mining and machine learning**. New York: Springer, 2009.
- BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International journal of operations and production management**, Bradford, v. 22, n. 2, p. 241–264, 2002. Disponível em: https://www.researchgate.net/publication/235291349_Operations_management_research_methodologies_using_quantitative_modeling. Acesso em: 21 abr. 2021.
- BODNAR, L. M. *et al.* Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. **The american journal of clinical nutrition**, New York, v. 111, n. 6, p. 1235–1243, 2020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0002916522011303>. Acesso em: 25 abr. 2021.
- CANTAMESSA, M. An empirical perspective upon design research. **Journal of engineering design**, London, v. 14, n. 1, p. 1–15, 2003. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/0954482031000078126>. Acesso em: 22 abr. 2021.

CHAPMAN, P. *et al.* **CRISP-DM 1.0: step-by-step data mining guide**. Chicago: SPSS Inc., 2000. 78 p. Disponível em: <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>. Acesso em: 14 abr. 2021.

CHAR, D. S.; SHAH, N. H.; MAGNUS, D. Implementing machine learning in health care — addressing ethical challenges. **The New England journal of medicine**, Boston, v. 378, n. 11, p. 981–983, 2018. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5962261/>. Acesso em: 03 jun. 2021.

CHEN, T.; GUESTRIN, C. Xgboost: a scalable tree boosting system. In: KRISHNAPURAM, B.; SHAH, M. (Ed.). *In: ACM SIGKDD INTERNATIONAL COMPUTING MACHINERY*, 22, 2022, San Francisco. **Proceedings [...]**. New York: Association for Computing Machinery, 2016. p. 785–794. Disponível em: <https://arxiv.org/pdf/1603.02754.pdf>. Acesso em: 04 jan. 2023.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. **Peerj compututer science**, London, v. 7, n. 5, p. e623, 2021. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8279135/>. Acesso em: 10 jan. 2023.

CHOI, T. M.; CHENG, T. C. E.; ZHAO, X. Multi-methodological research in operations management. **Production and operations management**, Baltimore, v. 25, n. 3, p. 379–389, 2016. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/poms.12534>. Acesso em: 22 abr. 2021.

DEO, R. C. Machine learning in medicine. **Circulation**, Baltimore, v. 132, n. 20, p. 1920–1930, 2015. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5831252/>. Acesso em: 03 jun. 2021.

DEPARTAMENTO DE INFORMÁTICA DO SISTEMA ÚNICO DE SAÚDE. **Ministério da saúde: informações de saúde**. Brasília: Departamento de informática do Sistema Único de Saúde (DATASUS), 2022. Disponível em: <https://datasus.saude.gov.br/>. Acesso em: 15 abr. 2022.

DOUPE, P.; FAGHMOUS, J.; BASU, S. Machine learning for health services researchers. **Value health**, Amsterdam, v. 22, n. 7, p. 808–815, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1098301519301469>. Acesso em: 05 jun. 2021.

ECK, N. J.; WALTMAN, L. **VOSviewer manual**. Leiden: Universiteit Leiden, 2023. 54 p. Disponível em: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.19.pdf. Acesso em: 21 abr. 2023.

FLATH, C. M.; STEIN, N. Towards a data science toolbox for industrial analytics applications. **Computers in industry**, Amsterdam, v. 94, p. 16–25, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0166361517303457>. Acesso em: 15 ago. 2022.

GAO, C. *et al.* Deep learning predicts extreme preterm birth from electronic health records. **Journal of biomedical informatics**, San Diego, v. 100, p. 103334, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046419302539>. Acesso em: 10 jun. 2021.

GOIN, D. E. *et al.* Mediation of firearm violence and preterm birth by pregnancy complications and health behaviors: addressing structural and postexposure confounding. **American journal of epidemiology**, Baltimore, v. 189, n. 8, p. 820–831, 2020. Disponível em: <https://academic.oup.com/aje/article/189/8/820/5812651>. Acesso em: 10 jun. 2021.

GOUDARD, M. J. F. *et al.* The role of skin-to-skin contact in exclusive breastfeeding: a cohort study. **Revista de saúde pública**, São Paulo, v. 56, p. 71, 2022. Disponível em: <https://www.scielo.br/j/rsp/a/kXvZFGtyZwrPTTSwZb55QCG/?lang=en>. Acesso em: 30 jul. 2022.

HAMILTON, E. F. *et al.* Estimating risk of severe neonatal morbidity in preterm births under 32 weeks of gestation. **Journal of maternal-fetal and neonatal medicine**, London, v. 33, n. 1, p. 73–80, 2020. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/14767058.2018.1487395>. Acesso em: 07 jul. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2nd. ed. New York: Springer, 2009.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Cidades e estados**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística (IBGE), 2022a. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>. Acesso em: 03 ago. 2022.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Malha municipal**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística (IBGE), 2022b. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=acesso-ao-produto>. Acesso em: 13 jan. 2022.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina**: uma abordagem estatística. São Carlos: Rafael Izbicki, 2020. Disponível em: <https://tiagoms.com/publications/ame/AME.pdf>. Acesso em: 10 jan. 2022.

JAIN, S.; SHUKLA, S.; WADHVANI, R. Dynamic selection of normalization techniques using data complexity measures. **Expert systems with applications**, New York, v. 106, p. 252–262, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741741830232X>. Acesso em: 10 ago. 2022.

JAIN, V.; CHATTERJEE, J. M. **Machine learning with health care perspective**: machine learning and healthcare. Switzerland: Springer, 2020.

JAMES, G. *et al.* **An introduction to statistical learning**: with applications in R. 2nd. ed. New York: Springer, 2021. Disponível em: <https://www.statlearning.com/>. Acesso em: 20 fev. 2022.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, Washington, v. 349, n. 6245, p. 255–260, 2015. Disponível em: https://www.science.org/doi/10.1126/science.aaa8415?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%20pubmed. Acesso em: 09 jul. 2021.

KHATIBI, T.; KHEYRIKOOCHAKSARAYEE, N.; SEPEHRI, M. M. Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features. **Archives of gynecology and obstetrics**, Berlin, v. 300, n. 6, p. 1565–1582, 2019. Disponível em: <https://link.springer.com/article/10.1007/s00404-019-05325-3>. Acesso em: 10 jul. 2021.

- KOTU, V.; DESHPANDE, B. **Data science: concepts and practice**. 2nd. ed. Massachusetts: Elsevier, 2019. Disponível em: <https://asolanki.co.in/wp-content/uploads/2019/04/Data-Science-Concepts-and-Practice-2nd-Edition-3.pdf>. Acesso em: 14 set. 2022.
- LEE, K. S.; AHN, K. H. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. **Journal of korean medical science**, Seoul, v. 34, n. 16, p. e128, 2019. Disponível em: <https://jkms.org/DOIx.php?id=10.3346/jkms.2019.34.e128>. Acesso em: 15 ago. 2021.
- LEE, K. S. *et al.* Determinants of spontaneous preterm labor and birth including gastroesophageal reflux disease and periodontitis. **Journal of korean medical science**, Seoul, v. 35, n. 14, p. e105, 2020. Disponível em: <https://jkms.org/DOIx.php?id=10.3346/jkms.2020.35.e105>. Acesso em: 15 ago. 2021.
- LEONE, C. R. Determinantes da prematuridade. **Boletim do instituto de saúde**, São Paulo, v. 16, n. 1, p. 98–104, 2015. Disponível em: <https://periodicos.saude.sp.gov.br/bis/article/view/37377/35495>. Acesso em: 16 jun. 2022.
- LOPES, M. L. B.; BARBOSA, R. M.; FERNANDES, M. A. C. Unsupervised learning applied to the stratification of preterm birth risk in Brazil with socioeconomic data. **International journal of environmental research and public health**, Basel, v. 19, n. 9, p. 5596, 2022. Disponível em: <https://www.mdpi.com/1660-4601/19/9/5596>. Acesso em: 21 jul. 2022.
- MALIK, A.; BAARDA, A. **Preterm birth**. New York: Nova, Medicine and Health, 2020.
- MARSLAND, S. **Machine learning: an algorithmic perspective**. 2nd. ed. Florida: CRC Press, 2015.
- MATSUSHITA, F. Y.; KREBS, V. L. J.; CARVALHO, W. B. Identifying clinical phenotypes in extremely low birth weight infants-an unsupervised machine learning approach. **European journal of pediatrics**, Berlin, v. 181, n. 3, p. 1085–1097, 2022. Disponível em: <https://link.springer.com/article/10.1007/s00431-021-04298-3>. Acesso em: 15 fev. 2022.
- MOREIRA, M. W. *et al.* A preterm birth risk prediction system for mobile health applications based on the support vector machine algorithm. In: JAJSZCZYK, A.; MARQUARDT, R. (Ed.). *In: ICC COMMUNICATIONS FOR CONNECTING HUMANITY*, 53, 2018, Kansas City. **Proceedings [...]**. New York: IEEE Communications Society, 2018. p. 1–5. Disponível em: <https://ieeexplore.ieee.org/document/8422616>. Acesso em: 15 set. 2023.
- MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with python: a guide for data scientists**. California: O'Reilly Media, 2016.
- NAIDU, K.; FREDLUND, K. L. **Gestational age assessment**. Treasure Island: StatPearls Publishing, 2023. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK526000/>. Acesso em: 23 ago. 2023.
- NAKAGAWA, E. Y.; ANTONINO, P. O. **Reference architectures for critical domains: industrial uses and impacts**. London: Springer Nature, 2023.
- NGIAM, K. Y.; KHOR, I. W. Big data and machine learning algorithms for health-care delivery. **The lancet: oncology**, London, v. 20, n. 5, p. e262–e273, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1470204519301494>. Acesso em: 17 ago. 2021.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future - big data, machine learning, and clinical medicine. **The New England journal of medicine**, Boston, v. 375, n. 13, p. 1216–1219, 2016. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070532/>. Acesso em: 17 ago. 2021.

OLIVEIRA, T. G. *et al.* Escore de apgar e mortalidade neonatal em um hospital localizado na zona sul do município de São Paulo. **Einstein**, São Paulo, v. 10, p. 22–28, 2012. Disponível em: <https://www.scielo.br/j/eins/a/j8tgY9YfZvCRn6M4Kjf98qv/?lang=pt>. Acesso em: 10 jul. 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Preterm birth**. Genebra: Organização Mundial da Saúde (OMS), 2022. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/preterm-birth>. Acesso em: 21 jan. 2022.

PACZKOWSKI, W. R. **Advanced data handling**: preprocessing methods. Switzerland: Springer, 2021. 127–157 p.

PORTUGAL, I.; ALENCAR, P.; COWAN, D. The use of machine learning algorithms in recommender systems: a systematic review. **Expert systems with applications**, New York, v. 97, p. 205–227, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417417308333>. Acesso em: 05 jan. 2022.

QUEIROZ, M. N.; GOMES, T. G. A. C. B.; MOREIRA, A. C. G. Idade gestacional, índice de apgar e peso ao nascer no desfecho de recém-nascidos prematuros. **Comunicação em ciências da saúde**, Brasília, v. 29, n. 04, 2018. Disponível em: <https://revistaccs.escs.edu.br/index.php/comunicacaoemcienciasdasaude/article/view/294/242>. Acesso em: 10 jul. 2023.

REDE NACIONAL PRIMEIRA INFÂNCIA. **O custo da prematuridade para a saúde pública ultrapassa R\$8 bilhões por ano no país**. Recife: Rede Nacional Primeira Infância, 2019. Disponível em: <http://primeirainfancia.org.br/o-custo-da-prematuridade-para-a-saude-publica-ultrapassa-r-8-bilhoes-por-ano-no-pais/>. Acesso em: 15 abr. 2021.

REIS, Z. S. N. *et al.* Newborn skin maturity medical device validation for gestational age prediction: clinical trial. **Journal of medical internet research**, Richmond, v. 24, n. 9, p. e38727, 2022. Disponível em: <https://www.jmir.org/2022/9/e38727/>. Acesso em: 13 ago. 2022.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in biology and medicine**, Elmsford, v. 112, p. 103375, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482519302525?via%3Dihub>. Acesso em: 23 jan. 2023.

RITTENHOUSE, K. J. *et al.* Improving preterm newborn identification in low-resource settings with machine learning. **Plos one**, San Francisco, v. 14, n. 2, p. e0198919, 2019. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198919>. Acesso em: 23 out. 2021.

ROCHA, A. S. *et al.* Determinantes do nascimento prematuro: proposta de um modelo teórico hierarquizado. **Ciência e saúde coletiva**, Rio de Janeiro, v. 27, n. 8, p. 3139–3152, 2022. Disponível em: <https://www.scielo.br/j/csc/a/Q4tbwN8FfVcmY8xpXWkz8JK/>. Acesso em: 26 out. 2022.

ROCHA, T. A. H. *et al.* Data-driven risk stratification for preterm birth in Brazil: a population-based study to develop of a machine learning risk assessment approach. **The lancet regional health – Americas**, Amsterdam, v. 3, p. 100053, 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2667193X21000454>. Acesso em: 16 ago. 2023.

RODRÍGUEZ, E. *et al.* Machine learning techniques to predict overweight or obesity. In: SHAKHOVSKA, N. *et al.* (Ed.). *In: IDDM - INTERNATIONAL CONFERENCE ON INFORMATICS AND DATA-DRIVEN MEDICINE*, 4, 2021, Valencia. **Proceedings** [...]. Aachen: CEUR Workshop Proceedings, 2021. p. 190–204. Disponível em: <https://ceur-ws.org/Vol-3038/paper13.pdf>. Acesso em: 10 jan. 2023.

RODRÍGUEZ, E. C. A. *et al.* Spatial patterns of prematurity and its determinants in the metropolitan region of São Paulo, Brazil, 2010-2019. **Revista brasileira de epidemiologia**, São Paulo, v. 27, p. e240008, 2024. Disponível em: <https://www.scielo.br/j/rbepid/a/mgTkJcJYwP8RbDcT3MvBSnz/>. Acesso em: 26 fev. 2024.

RODRÍGUEZ, E. Y. A. **Técnicas de aprendizado de máquina para predição do custo da logística de transporte: uma aplicação em empresa do segmento de autopeças**. Dissertação (Mestrado em Engenharia de Produção) — Faculdade de Engenharia e Ciências, Universidade Estadual Paulista, Guaratinguetá, 2020. Disponível em: <https://repositorio.unesp.br/items/750c1858-58ec-4d54-8c3a-9346c854ddf0>. Acesso em: 15 jan. 2023.

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. e230034, 2023. Disponível em: <https://www.scielo.br/j/rbepid/a/Lc5YT6XwsqDvKk5fBQNQVzM/?lang=en>. Acesso em: 29 ago. 2023.

SANTOS, H. G. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. Tese (Doutorado em Epidemiologia) — Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018. Disponível em: https://teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/publico/HellenGeremiasdosSantos_DR_ORIGINAL.pdf. Acesso em: 10 abr. 2021.

SAZAWAL, S. *et al.* Machine learning prediction of gestational age from metabolic screening markers resistant to ambient temperature transportation: facilitating use of this technology in low resource settings of South Asia and East Africa. **Journal of global health**, London, v. 12, p. 04021, 2022. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9022771/>. Acesso em: 30 mar. 2023.

SAZAWAL, S. *et al.* Machine learning guided postnatal gestational age assessment using new-born screening metabolomic data in South Asia and sub-Saharan Africa. **BMC pregnancy and childbirth**, London, v. 21, p. 1–11, 2021. Disponível em: <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-021-04067-y>. Acesso em: 30 mar. 2023.

SHAI, S. S.; SHAI, B. D. **Understanding machine learning: from theory to algorithms**. New York: Cambridge University Press, 2014. 1–416 p. Disponível em: <https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>. Acesso em: 11 out. 2021.

SHEN, L. *et al.* Attention-guided deep learning for gestational age prediction using fetal brain MRI. **Scientific reports**, London, v. 12, n. 1, p. 1408, 2022. Disponível em: <https://www.nature.com/articles/s41598-022-05468-5>. Acesso em: 11 abr. 2023.

SOMANI, S. S. *et al.* Characterization of patients who return to hospital following discharge from hospitalization for COVID-19. **Journal of general internal medicine**, New York, v. 35, n. 10, p. 2838–2844, 2020. Disponível em: <https://link.springer.com/article/10.1007/s11606-020-06120-6>. Acesso em: 10 ago. 2022.

SZWARCWALD, C. L. *et al.* Avaliação das informações do sistema de informações sobre nascidos vivos (SINASC), Brasil. **Cadernos de saúde pública**, Rio de Janeiro, v. 35, n. 10, 2019. Disponível em: <https://www.scielo.br/j/csp/a/dqy9gbh3k8fJBjYHVBwbMnS>. Acesso em: 11 out. 2021.

TASCÓN, L. A. M. *et al.* Epidemiología de la prematuridad, sus determinantes y prevención del parto prematuro. **Revista chilena de obstetricia y ginecología**, Santiago de Chile, v. 81, n. 4, p. 330–342, 2016. Disponível em: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-75262016000400012. Acesso em: 17 set. 2022.

UNITED NATIONS. **Millennium development goals report 2015**. New York: United Nations, 2015a. 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: 21 abr. 2021.

UNITED NATIONS. **Transforming our world: the 2030 agenda for sustainable development**. New York: United Nations, 2015b. Disponível em: https://www.un.org/ga/search/view_doc.asp?symbol=A/%20RES/70/1&Lang=E. Acesso em: 21 abr. 2021.

VOLLMER, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. **BMJ**, London, v. 368, p. l6927, 2020. Disponível em: <https://www.bmj.com/content/368/bmj.l6927>. Acesso em: 14 nov. 2021.

ZHANG, H. *et al.* Research on the construction and realization of data pipeline in machine learning regression prediction. **Mathematical problems in engineering**, London, v. 2022, 2022. Disponível em: <https://www.hindawi.com/journals/mpe/2022/7924335/>. Acesso em: 30 jan. 2023.

ZHENG, A.; CASARI, A. **Feature engineering for machine learning: principles and techniques for data scientists**. Sebastopol: O'Reilly, 2018.

ZHOU, L. *et al.* Machine learning on big data: opportunities and challenges. **Neurocomputing**, Amsterdam, v. 237, p. 350–361, 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231217300577>. Acesso em: 02 dez. 2021.