**UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO**

**JUAN VENDRAMI**

# APLICAÇÃO DE MÉTODOS DE *MACHINE LEARNING* PARA PREDIÇÃO E SELEÇÃO DE ATRIBUTOS NA CLASSIFICAÇÃO DE BAIXO PESO AO NASCER

**Guaratinguetá**
**2025**

**JUAN VENDRAMI**

# APLICAÇÃO DE MÉTODOS DE *MACHINE LEARNING* PARA PREDIÇÃO E SELEÇÃO DE ATRIBUTOS NA CLASSIFICAÇÃO DE BAIXO PESO AO NASCER

Trabalho de Graduação em Engenharia de Produção Mecânica apresentado ao Departamento de Engenharia e Ciências de Guaratinguetá, como exigência parcial para colação de grau e obtenção do título de bacharel em Engenharia de Produção Mecânica.
Orientador: Prof. Dr. Fernando Augusto Silva Marins.

**Guaratinguetá**
**2025**

Vendrami, Juan

V453a Aplicação de métodos de *Machine Learning* para predição e seleção de atributos na classificação de baixo peso ao nascer / Juan Vendrami - Guaratinguetá, 2025.

16 f : il.

Bibliografia: f. 16

Trabalho de Graduação em Engenharia de Produção Mecânica – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2025.

Orientador: Prof. Dr. Fernando Augusto Silva Marins.

1. Aprendizado do computador. 2. Recém-nascido - Peso baixo. 3. Processo decisório. 4. Análise de regressão logística. I. Título.

CDU 65.012.4

Luciana Máximo
Bibliotecária/CRB-8 3595

**JUAN VENDRAMI**

# APLICAÇÃO DE MÉTODOS DE *MACHINE LEARNING* PARA PREDIÇÃO E SELEÇÃO DE ATRIBUTOS NA CLASSIFICAÇÃO DE BAIXO PESO AO NASCER

Trabalho de Graduação em Engenharia de Produção Mecânica apresentado ao Departamento de Engenharia e Ciências de Guaratinguetá, como exigência parcial para colação de grau e obtenção do título de bacharel em Engenharia de Produção Mecânica.
Orientador: Prof. Dr. Fernando Augusto Silva Marins.

Data: 14 de Novembro de 2025.

Banca examinadora:

gov.br Documento assinado digitalmente
FERNANDO AUGUSTO SILVA MARINS
Data: 29/01/2026 08:29:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fernando Augusto Silva Marins

gov.br Documento assinado digitalmente
ANEIRSON FRANCISCO DA SILVA
Data: 28/01/2026 10:29:20-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Aneirson Francisco da Silva

unesp Digitally signed by Elen Yanina Aguirre Rodríguez
Date: 2026.01.28 23:42:24 -05'00'

Prof. Dr. Elen Yanina Aguirre Rodriguez

## RESUMO

Neste estudo, foram aplicadas técnicas de Machine Learning para prever o risco de baixo peso ao nascer (BPN), utilizando dados do sistema SINASC. O modelo Regressão Logística apresentou acurácia média de 93,10% com validação cruzada, e bom desempenho em métricas como precisão, recall e F1-score. A análise de importância das variáveis destacam fatores como idade gestacional e tipo de gestação como relevantes para a predição. Os resultados indicam que o modelo é capaz de identificar com confiabilidade os casos de BPN, contribuindo para ações preventivas mais eficazes e para o direcionamento estratégico de recursos na atenção pré-natal.



**Palavra-chave:** Baixo peso ao nascer; Random Forest; Machine Learning.

## ABSTRACT

In this study, Machine Learning techniques were applied to predict the risk of low birth weight (LBW) using data from the SINASC system. The Logistic Regression model achieved an average accuracy of 93.10% with cross-validation and showed good performance in metrics such as precision, recall, and F1-score. The analysis of variable importance highlighted factors such as gestational age and type of pregnancy as relevant for prediction. The results indicate that the model is able to reliably identify LBW cases, contributing to more effective preventive actions and to the strategic allocation of resources in prenatal care.



**Keywords**: Low birth weight; Random Forest; Machine Learning.

**SUMÁRIO**

## 1 INTRODUÇÃO

O Aprendizado de Máquina - *Machine Learning* (ML) é um dos principais ramos da Inteligência Artificial, sendo definida como o estudo de métodos computacionais capazes de identificar padrões complexos em vastos conjuntos de dados para construir modelos preditivos [1]. Diferentemente dos métodos tradicionais baseados em dados, os sistemas de ML utilizam dados coletados e armazenados para melhorar iterativamente seu desempenho, tornando-os altamente adaptáveis a diversos problemas complexos.

Nesse contexto, as técnicas de ML surgiram como uma alternativa promissora aos métodos estatísticos tradicionais [2], oferecendo maior flexibilidade e escalabilidade no processamento de grandes volumes de dados com estruturas diversas. A aplicação dessas metodologias permite a extração de informações úteis, além do desenvolvimento de ferramentas avançadas voltadas para o suporte ao planejamento e à tomada de decisões críticas.

Por outro lado, o baixo peso ao nascer (BPN) refere-se a recém-nascidos com peso inferior a 2.500g e é considerado um importante problema de saúde pública devido à sua associação com maiores riscos de mortalidade neonatal e infantil [3,4]. Assim, a redução da incidência de BPN tem sido estabelecida como um objetivo de saúde em diversos países, visando diminuir as taxas de mortalidade infantil [5].

Em 2012, os Estados-membros da Organização Mundial da Saúde (OMS) comprometeram-se a reduzir os casos de BPN em 30% até 2025 [6], como parte dos esforços para atingir os Objetivos de Desenvolvimento do Milênio (ODM) e os Objetivos de Desenvolvimento Sustentável (ODS). Especificamente, esse compromisso contribui para o alcance do Objetivo 3, relacionado à Saúde e Bem-Estar, que estabelece a meta de reduzir a taxa de mortalidade neonatal para, no máximo, 12 por 1.000 nascidos vivos [7].

Nesse cenário, a detecção precoce de gestações de alto risco associadas ao BPN representa uma estratégia valiosa para gestores e especialistas em saúde. A integração de algoritmos avançados de ML, como Random Forests (RF), tem ganhado destaque no setor da saúde, oferecendo aplicações promissoras para modelagem preditiva e sistemas de suporte à decisão [8].

Esses algoritmos são especialmente vantajosos na análise de conjuntos de dados complexos e multidimensionais, permitindo a identificação de fatores de risco críticos e facilitando intervenções oportunas para melhorar os desfechos maternos e neonatais.

# 2 OBJETIVOS

## 2.1 OBJETIVO GERAL

O objetivo geral deste estudo foi desenvolver e validar um modelo preditivo utilizando ML para identificar o risco de baixo peso ao nascer (BPN) em gestantes, visando apoiar especialistas em saúde nos processos de tomada de decisão para a prevenção e redução dos casos de BPN.

## 2.2 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral, uma série de objetivos específicos foram estabelecidos, conforme descrito a seguir:

a) Coletar e criar um banco de dados com informações sobre nascidos vivos no Estado de São Paulo;
b) Limpar e pré-processar os dados utilizando técnicas de engenharia de atributos;
c) Treinar e testar o algoritmo *Random Forest*, avaliando sua eficiência em termos de acurácia preditiva, robustez e desempenho geral;
d) Validar os resultados por meio de métodos estatísticos para garantir confiabilidade e funcionalidade.

## 3 MÉTODO DE PESQUISA

Para alcançar os objetivos desta proposta de pesquisa, três etapas principais e sequenciais foram realizadas:

Etapa 1 - Exploração:

A fase inicial do estudo envolveu uma revisão abrangente da literatura para identificar pesquisas anteriores sobre modelos de previsão de baixo peso ao nascer (BPN). Em seguida, foi realizada a aquisição de dados. As informações sobre nascidos vivos de mães residentes foram obtidas do Departamento de Informática do Sistema Único de Saúde (DATASUS), especificamente do Sistema de Informações sobre Nascidos Vivos (SINASC) [9]. O SINASC, baseado em declarações de nascidos vivos, forneceu detalhes sobre a gestação, o parto e as condições do nascimento [10]. Esse conjunto de dados serviu como base para o desenvolvimento do modelo.

Etapa 2 - Desenvolvimento:

Os dados coletados do SINASC passaram por um rigoroso processo de controle de qualidade, incluindo reformatação, pré-processamento e tratamento de valores ausentes ou incorretos, garantindo a confiabilidade do conjunto de dados. Além disso, foi realizada uma análise estatística para compreender a distribuição e as características das variáveis. Técnicas de redução de atributos, como a remoção de variáveis irrelevantes ou redundantes [11] e análise de correlação [12], foram aplicadas para melhorar a eficiência do modelo.

Em seguida, foi avaliada a adequação do uso do algoritmo Random Forest, reconhecido por seu desempenho preditivo (Alpaydin, 2020). As fases de treinamento e teste permitiram o desenvolvimento de um modelo de classificação, que foi avaliado por meio da comparação de diversas métricas de desempenho, como acurácia, precisão, recall, F1-score e área sob a curva ROC [13,14]. Por fim, o modelo selecionado foi validado por meio de técnicas estatísticas robustas e fontes externas de dados para garantir sua generalização e confiabilidade [13].

Etapa 3 - Interpretação e Avaliação:

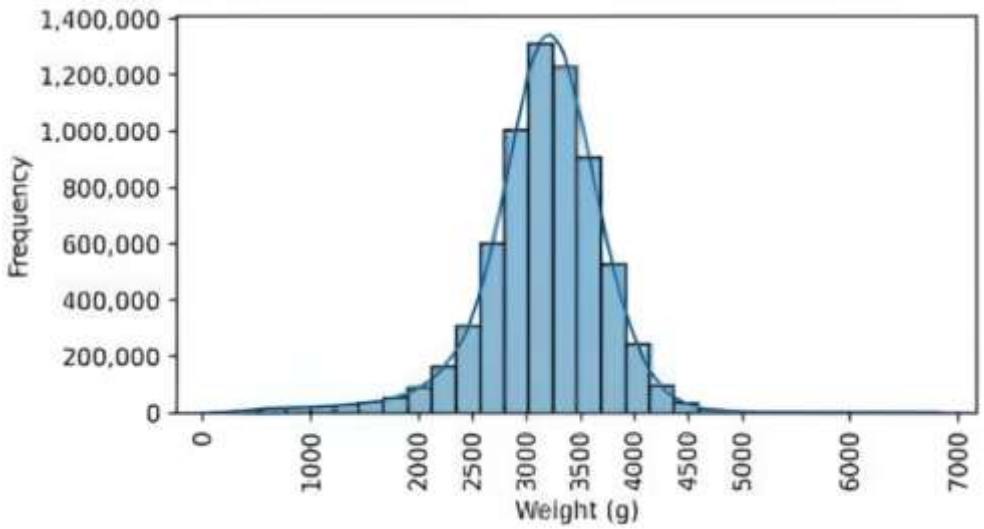
Os resultados foram analisados e interpretados para sintetizar as descobertas. Eles foram comparados com *insights* de estudos anteriores na literatura para avaliar a consistência e a relevância do modelo proposto. A interpretação final destacou as implicações da pesquisa, suas contribuições para a área e possíveis direções para trabalhos futuros.

## 4 RESULTADOS

Os dados coletados do sistema SINASC referem-se ao período de 2012 a 2023, abrangendo informações sobre mais de 6 milhões de nascidos vivos no estado de São Paulo. Após o processamento e a limpeza do conjunto de dados, os dados utilizados para o treinamento do modelo consistiram em 6.722.132 linhas, com as seguintes colunas: "IDADEMAE" (idade da mãe), "ESTCIVMAE" (estado civil da mãe), "QTDFILVIVO" (quantidade de filhos vivos), "QTDFILMORT" (quantidade de natimortos), "CODMUNRES" (município de residência), "GESTACAO" (gestação), "GRAVIDEZ" (tipo de gravidez), "PARTO" (tipo de parto), "CONSULTAS" (número de consultas pré-natais), "SEXO" (sexo), "APGAR1" (pontuação Apgar no 1º minuto), "APGAR5" (pontuação Apgar no 5º minuto), "PESO" (peso), "SEMAGESTAC" (idade gestacional), "KOTELCHUCK" (índice de Kotelchuck), "PARIDADE" (paridade) e "CONSPRENAT" (assistência pré-natal).

A Figura 2 ilustra a distribuição do peso no conjunto de dados, que foi utilizado para treinar o modelo de classificação empregando a técnica de *Random Forest*.

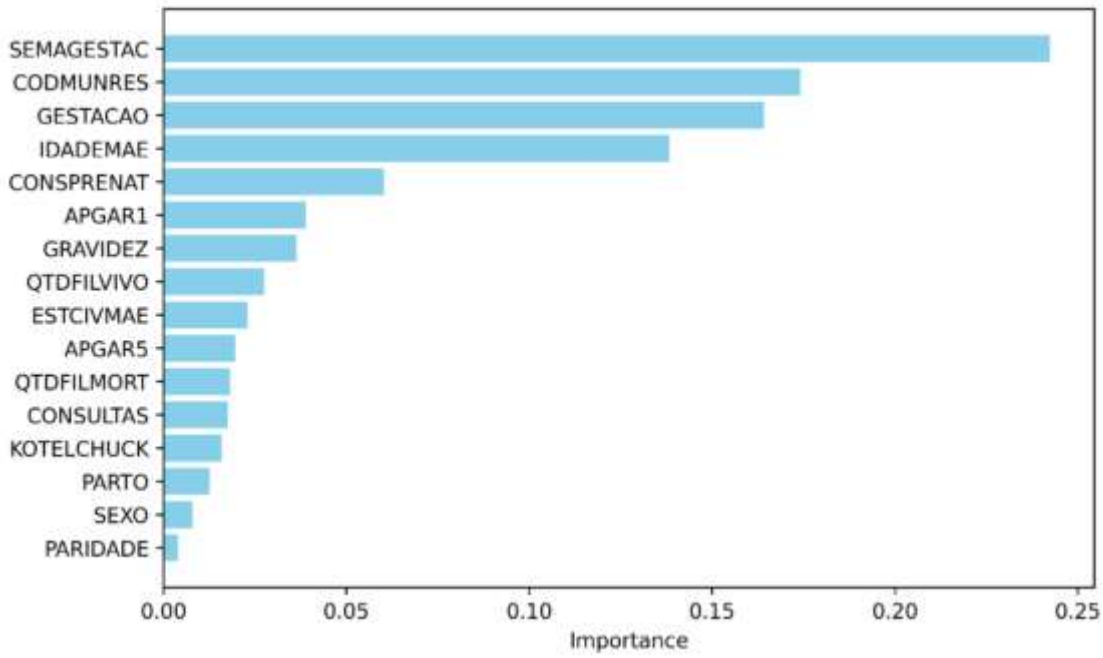
Figura 1 - Distribuição dos dados de peso dos nascidos vivos no conjunto de dados.




Fonte: Elaborado pelo autor.

Conforme ilustrado na Figura 2, o peso dos recém-nascidos variou aproximadamente entre 100 e 6.830 gramas. Por outro lado, a avaliação da importância das variáveis utilizando a correlação de Pearson permitiu identificar os atributos mais relevantes no conjunto de dados para a previsão do baixo peso ao nascer. A Figura 3 apresenta o gráfico de importância das variáveis.

Figura 2 - Importância das variáveis.




Fonte: Elaborado pelo autor.

Essa análise destaca os fatores-chave que influenciam significativamente a ocorrência de baixo peso ao nascer (Figura 3), fornecendo informações valiosas para aprimorar os modelos preditivos e entender os padrões subjacentes nos dados. A Tabela 1 exibe as variáveis juntamente com seus respectivos valores de importância.

Tabela 1 - Importância das variáveis.

| VARIÁVEIS | IMPORTÂNCIA |
|---|---|
| **SEMAGESTAC** | 0,242308 |
| **CODMUNRES** | 0,173969 |
| **GESTACAO** | 0,164152 |
| **IDADEMAE** | 0,138186 |
| **CONSPRENAT** | 0,062091 |
| **APGAR1** | 0,038823 |
| **GRAVIDEZ** | 0,036207 |
| **QTDFILVIVO** | 0,027580 |
| **ESTCIVMAE** | 0,023029 |
| **APGAR5** | 0,019607 |

(Continua)

| | |
|---|---|
| **QTDFILMORTO** | 0,018147 |
| **CONSULTAS** | 0,017649 |
| **KOTELCHUCK** | 0,015862 |
| **PARTO** | 0,012531 |
| **SEXO** | 0,007823 |
| **PARIDADE** | 0,003834 |

Fonte: Elaborado pelo autor.

A análise da importância das variáveis permitiu a eliminação de certas variáveis do conjunto de dados, com a seleção das variáveis cujos valores de importância superaram 0,010. Com a seleção das variáveis foi realizada a aplicação de 2 modelos *Random Forest*, um utilizando todas as variáveis cuja importância supera 0,010 e outro simplificado. Após a aplicação obteve-se os resultados que estão na Tabela 2.

Tabela 2 - Resultado dos Modelos *Random Forest.*

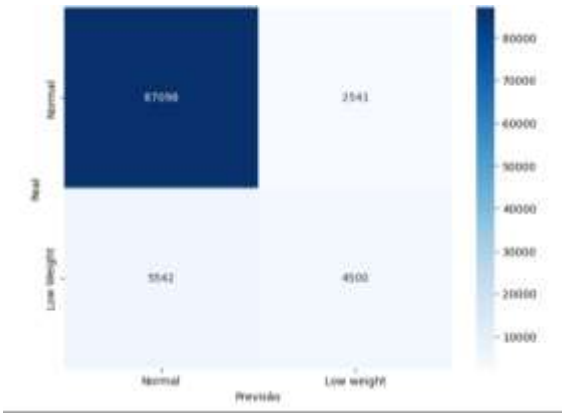
| Modelo | Precisão Peso Normal e Precisão Baixo Peso |
|---|---|
| **RF** | 0,94 0,64 |
| **RF Simplificado** | 0,94 0,66 |

Fonte: Elaborado pelo autor.

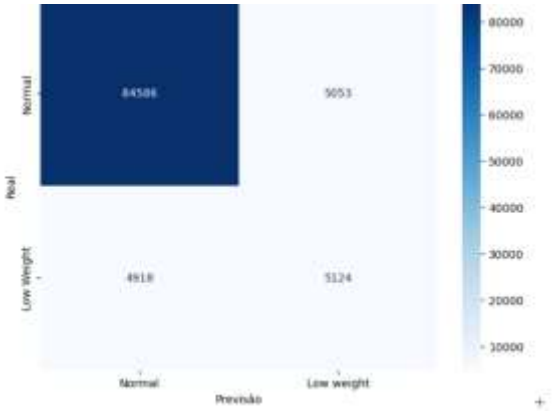
Esses dados indicaram que ambos os modelos apresentaram uma alta taxa em identificar recém-nascidos com o peso normal. Entretanto, o modelo que utiliza todas as variáveis apresentou um melhor desempenho ao prever o recém-nascido com baixo peso. Na Figura 4 tem-se a Matriz de Confusão dos 2 modelos.

Figura 3 - Matriz de Confusão.

A) RF



B) RF Simplificado

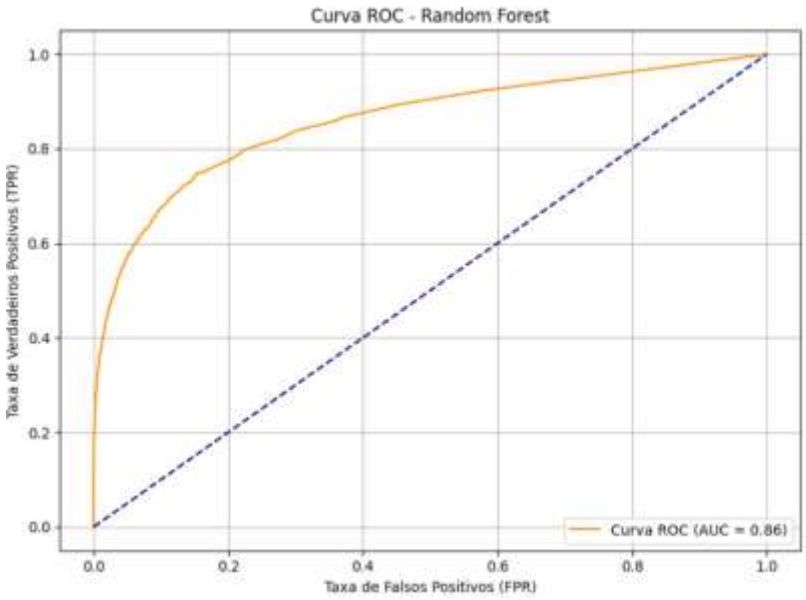



Fonte: Elaborado pelo autor.

Conforme mostrado na Figura 4 e na Tabela 1, e com base nas métricas de desempenho, o modelo que poderia ser selecionado como o melhor foi aquele treinado com o conjunto de dados completo, pois apresentou um desempenho superior na classificação dos casos de baixo peso ao nascer (BPN).

Isso indica que a manutenção de uma gama mais ampla de atributos permite que o modelo capture melhor os padrões subjacentes associados ao BPN, resultando em previsões mais precisas. O desempenho superior do modelo com o conjunto de dados completo sugere que a eliminação de variáveis pode comprometer a capacidade do modelo de distinguir efetivamente entre as diferentes classes.

Figura 4 - *Receiver Operating Characteristic* (ROC) Curva para *Random Forest.*




Fonte: Elaborado pelo autor.

Além disso, a Figura 5 exibe a curva Receiver Operating Characteristic (ROC) do melhor modelo, ilustrando seu desempenho na distinção entre as diferentes classes.

Prosseguindo com as análises do Modelo de *Random Forest* foi realizada a Validação Cruzada, uma técnica de avaliação de modelos de ML que busca medir o desempenho de um modelo de forma mais confiável e robusta. Para esse trabalho foi utilizada a validação cruzada de 5 *folds* cujos resultados podem ser vistos na Tabela 3.

Tabela 3 - Dados da Validação Cruzada com 5 *folds.*

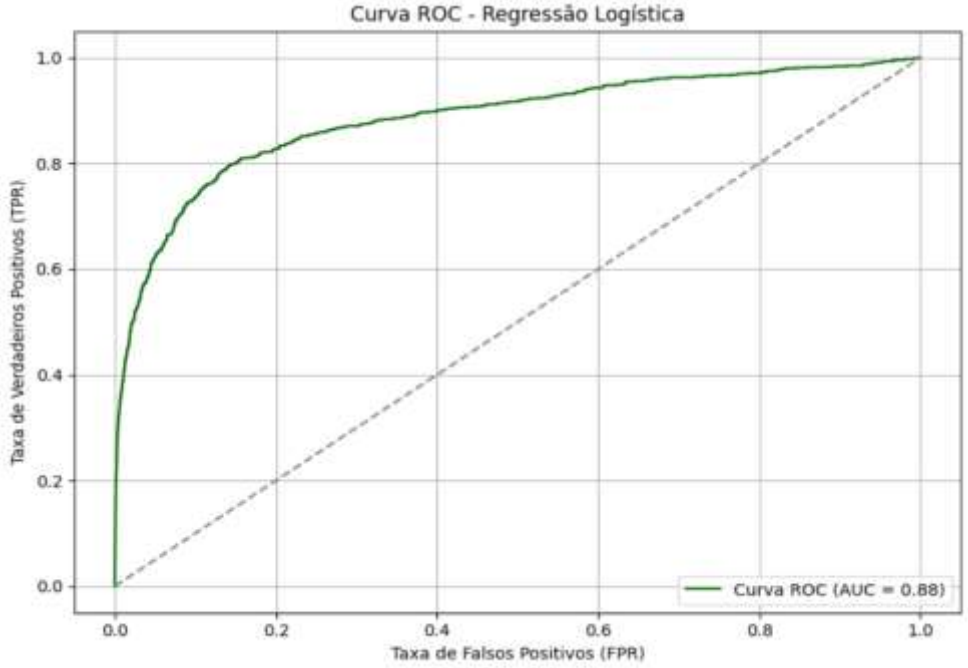
| | |
|---|---|
| Acurácia da dobra 1 | 0,9229 |
| Acurácia da dobra 2 | 0,9250 |
| Acurácia da dobra 3 | 0,9246 |
| Acurácia da dobra 4 | 0,9254 |
| Acurácia da dobra 5 | 0,9247 |
| **Média da acurácia nas 5 dobras** | **0,9245** |
| **Desvio padrão das acurácias** | **0,0009** |

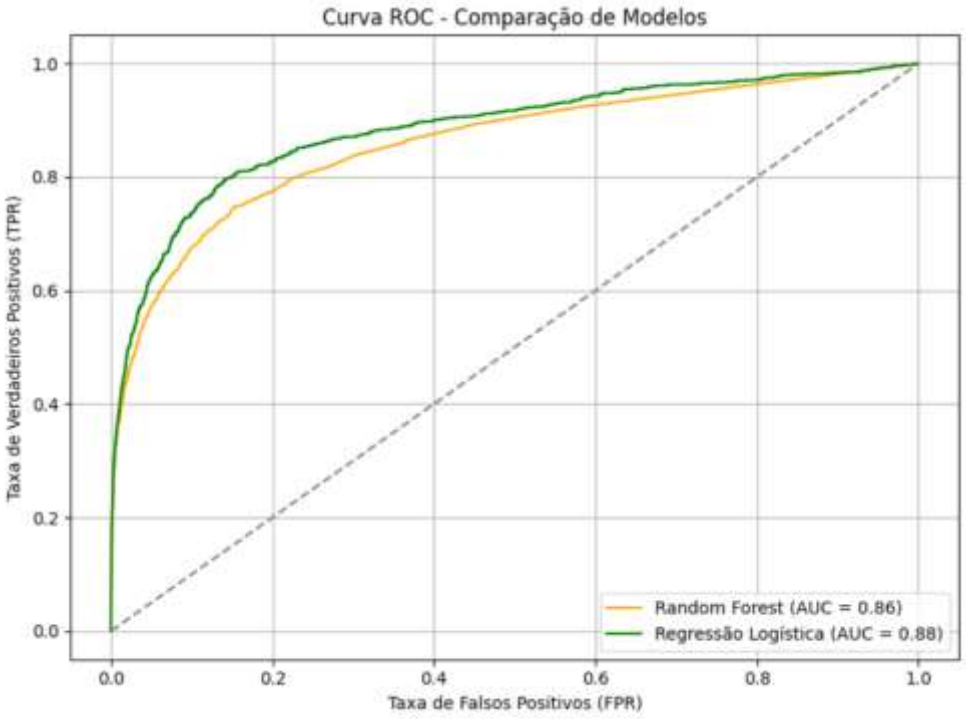
Fonte: Elaborado pelo autor.

Esses resultados indicaram que o modelo apresentou um desempenho consistente e estável, independentemente da divisão dos dados. A pequena variação entre as dobras mostra que o modelo possui boa capacidade de generalização, ou seja, ele tende a manter um bom desempenho mesmo quando aplicado a dados que não foram utilizados durante o treinamento. Além disso, a alta acurácia média mostrou que o modelo está, em geral, classificando corretamente a maior parte dos casos.

Por fim, foi testado um outro tipo de modelo de ML, a Regressão Logística. Ao aplicar esse modelo a precisão para prever um recém-nascido com baixo peso foi de 0.7828, um grande aumento quando comparado com o modelo de *Random Forest*. Também foi feita a validação cruzada de 5 *folds*, retornando uma média de 0,9310 e desvio padrão de 0,0013. Na Figura 6, pode-se observar a curva ROC para esse modelo. Quando se compara a curva ROC do modelo *Random Forest* (Figura 5) com a curva ROC da Regressão Logística (Figura 6) fica evidenciado que a Regressão Logística apresentou resultados superiores ao *Random Forest*.

Figura 5 - *Receiver Operating Characteristic* (ROC) - Curva para Regressão Logística




Fonte: Elaborado pelo autor.

## 5 CONCLUSÃO

Este estudo demonstrou o potencial da aplicação de técnicas de ML, especialmente o algoritmo Regressão Logística, na predição de baixo peso ao nascer (BPN). Utilizando uma base de dados ampla e real, proveniente do SINASC/DATASUS, foi possível construir e treinar modelos robustos, capazes de identificar padrões relevantes entre variáveis maternas, gestacionais e neonatais.

A análise de importância dos atributos evidenciou que fatores como idade gestacional (SEMAGESTAC), município de residência (CODMUNRES) e tipo de gestação (GESTACAO) estão fortemente associados ao risco de BPN, o que reforça a importância da vigilância pré-natal e da estratificação de risco gestacional.

Os modelos desenvolvidos apresentaram alta acurácia e desempenho consistente, conforme evidenciado pelos resultados da validação cruzada com 5 dobras, onde o modelo Regressão Logística alcançou uma média de acurácia superior a 93%, com baixo desvio padrão.

Apesar de ambos os modelos *Random Forest* (com todas as variáveis e simplificado) mostrarem boa performance, o modelo com maior número de atributos apresentou melhor recall na detecção de casos de baixo peso, o que é fundamental em um contexto de saúde pública.

A Regressão Logística também foi avaliada e superou o *Random Forest* em algumas métricas, especialmente na previsão de BPN, demonstrando que diferentes algoritmos podem ser complementares ou até superiores em determinados cenários.

Assim, conclui-se que os modelos baseados em ML são ferramentas promissoras para apoiar a tomada de decisão na área da saúde, particularmente na identificação precoce de gestações com risco de baixo peso ao nascer. A adoção desses sistemas pode contribuir para o direcionamento de políticas públicas, alocação de recursos e desenvolvimento de estratégias de prevenção mais eficazes.

Futuramente, recomenda-se a ampliação do estudo para outros estados brasileiros, bem como a integração de novas variáveis clínicas e socioeconômicas, visando o aprimoramento contínuo dos modelos e a maximização de seu impacto social.

## REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. 4. ed. Cambridge: MIT Press, 2020.

BRASIL. Ministério da Saúde. **Informações de nascidos vivos e óbitos infantis**. Brasília: Departamento de Informática do Sistema Único de Saúde, 2025. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos. Acesso em: 20 abr. 2025.

CAI, J. *et al.* Feature selection in machine learning: a new perspective. **Neurocomputing**, Amsterdam, v. 300, p. 70-79, 2018.

DOUPE, P.; FAGHMOUS, J.; BASU, S. Machine learning for health services researchers. **Value in Health**, Philadelphia, v. 22, n. 7, p. 808-815, 2019.

HUGHES, M. M.; BLACK, R. E.; KATZ, J. 2500-g low birth weight cutoff: history and implications for future research and policy. **Maternal and Child Health Journal**, New York, v. 21, n. 2, p. 283-289, 2017.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, Washington, v. 349, n. 6245, p. 255-260, 2015.

KATZ, J. *et al*. Mortality risk in preterm and small for gestational age infants in low-income and middle-income countries: a pooled country analysis. **The Lancet**, London, v. 382, n. 9890, p.
417-425, 2013.

KUMAR, S. N. *et al*. Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. **Reproductive Toxicology**, Philadelphia, v. 94, p. 55-63, 2020.

MARSLAND, S. **Machine learning**: an algorithmic perspective. 2. ed. Boca Raton: Chapman & Hall/CRC, 2015.

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, e230034, 2023.

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, e00214918, 2019.

TRAVERSO, A. *et al*. **Diving deeper into models**: fundamentals of clinical data. Cham: Springer International Publishing, 2019.

UNITED NATIONS. **Resolution adopted by the general assembly on 11 september 2015**: transforming our world: the 2030 agenda for sustainable development. New York: United Nations, 2015.

WORLD HEALTH ORGANIZATION. **Every newborn action plan**: progress report. Geneva: World Health Organization, 2019.