CADERNOS DE ESTUDOS SOCIAIS
v.37, n.1, 2022
e-ISSN:2595-40911

IMPACTO DA PANDEMIA PELA COVID-19 E MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE NASCIMENTOS PREMATUROS NAS CAPITAIS DA REGIÃO NORDESTE DO BRASIL, 2018-2021

Autor 1: **José Mauricio Matapi da Silva**
ORCID: 0000-0002-3184-6450
Filiação:CIN/UFPE
jmms2@cin.ufpe.br

Autor 2: **Heitor Victor Veiga da Costa**
ORCID: 0000-0003-2525-6689
Filiação: CIN/UFPE
hvvc@cin.ufpe.br

Autor 3: **Fernando Maciano de Paula Neto**
ORCID: 0000-0003-4264-1124
Filiação: CIN/UFPE
fernando@cin.ufpe.br

Trabalho submetido em 22/09/2022 e aprovado em 14/01/2023.
DOI: 10.33148/CESv37n1(2022)2122

RESUMO

O nascimento prematuro é um problema global devido a suas implicações para a morbidade e mortalidade. Consiste em um dos principais fatores de risco para a mortalidade neonatal e infantil. O parto pré-termo é definido como aquele cuja gestação termina entre a 20ª e a 37ª semanas ou entre 140 e 257 dias após o primeiro dia da última menstruação. Para este estudo, utilizou-se dados do Sistema de Informações sobre Nascidos Vivos (SINASC) das capitais da região Nordeste do Brasil, entre 2018 e 2021. Foi Verificado se os dois primeiros anos da pandemia pela covid-19 trouxeram impactos significativos para as distribuições das métricas de performance, em comparação ao que foi utilizado para treinamento e validação dos modelos. Foram aplicados seis algoritmos de aprendizado de máquina (Regressão Logística, Análise Discriminante Linear, Perceptron Multicamadas, AdaBoost, Árvore de decisão e Floresta Aleatória) para predição de prematuridade. Os modelos apresentaram como resultado queda na métrica Area Under the roc Curve (AUC) nos anos de 2020 e 2021 em relação a 2018 e 2019, com ênfase para os modelos Adaboost, Floresta Aleatória e Árvore de decisão, com quedas superiores a 10% atestadas pelos testes estatísticos de Kruskal-Wallis e Nemenyi. Como causadores da queda de performance dos modelos, foi identificado que as variáveis mês do início do pré-natal e idade perderam aderência em relação à base de treino. Os modelos apresentaram boa performance preditiva, contudo, a utilização de modelos baseados em árvores deve ser feita com cautela, visto que estes são mais instáveis e que a covid-19 trouxe impacto na distribuição das variáveis idade e mês de início de pré-natal. Para treinamento de novos modelos, atenção às variáveis de entrada e ao período utilizado para treinamento. Para soluções já estabelecidas, considerar o seu retreinamento.

PALAVRAS-CHAVE:

Prematuridade. Saúde. Inteligência Artificial. Aprendizado de Máquina. covid-19.

IMPACT OF THE COVID-19 PANDEMIC AND MACHINE LEARNING MODELS FOR PREDICTING PREMATURE BIRTHS IN THE CAPITALS OF THE NORTHEAST REGION OF BRAZIL, 2018-2021

ABSTRACT

Premature birth is a global problem due to its implications for morbidity and mortality. It is one of the main risk factors for neonatal and infant mortality. Preterm delivery is defined as one whose pregnancy ends between the 20th and 37th weeks or between 140 and 257 days after the first day of the last menstrual period. For this study, data from the Information System on Live Births (SINASC) from the capitals of the Northeast region of Brazil, between 2018 and 2021, were used. of the performance metrics, compared to what was used for training and validation of the models. Six machine learning algorithms (Logistic Regression, Linear Discriminant Analysis, Multilayer Perceptron, AdaBoost, Decision Tree and Random Forest) were applied to predict prematurity. The models showed a drop in the Area Under the Roc Curve (AUC) metric in the years 2020 and 2021 compared to 2018 and 2019, with emphasis on the Adaboost, Random Forest and Decision Tree models, with drops greater than 10% attested by the statistical tests of Kruskal-Wallis and Nemenyi. As causes of the drop in performance of the models, it was identified that the variables month of beginning of prenatal care and age lost adherence in relation to the training base. The models showed good predictive performance, however, the use of tree-based models should be done with caution, since they are more unstable and that covid-19 had an impact on the distribution of the variables age and month of beginning of prenatal care. For training new models, pay attention to the input variables and the period used for training. For already established solutions, consider your retraining.

KEYWORDS: Depersonalization. Emotional Exhaustion. Professional Achievement. Predictors.

IMPACTO DE LA PANDEMIA POR COVID-19 Y MODELOS DE APRENDIZAJE DE MÁQUINA PARA PREDICCIÓN DE NACIMIENTOS PREMATUROS EN LAS CAPITALES DE LA REGIÓN NORDESTE DE BRASIL, 2018-2021

RESUMEN

El parto prematuro es un problema mundial por sus implicaciones en la morbimortalidad. Es uno de los principales factores de riesgo de mortalidad neonatal e infantil. El parto prematuro se define como aquel cuyo embarazo finaliza entre las semanas 20 y 37 o entre 140 y 257 días después del primer día de la última menstruación. Para este estudio, se utilizaron datos del Sistema de Información sobre Nacidos Vivos (SINASC) de las capitales de la región Nordeste de Brasil, entre 2018 y 2021. de las métricas de desempeño, en comparación con lo que se utilizó para el entrenamiento y validación de los modelos. . Se aplicaron seis algoritmos de aprendizaje automático (regresión logística, análisis discriminante lineal, perceptrón multicapa, AdaBoost, árbol de decisión y bosque aleatorio) para predecir la prematuridad. Los modelos mostraron una caída en la métrica Area Under the Roc Curve (AUC) en los años 2020 y 2021 en comparación con 2018 y 2019, con énfasis

en los modelos Adaboost, Random Forest y Decision Tree, con caídas superiores al 10% atestiguadas por el Pruebas estadísticas de Kruskal-Wallis y Nemenyi. Como causas de la caída en el desempeño de los modelos, se identificó que las variables mes de inicio del prenatal y edad perdieron adherencia en relación a la base de formación. Los modelos mostraron un buen desempeño predictivo, sin embargo, el uso de modelos basados en árboles debe hacerse con precaución, ya que son más inestables y que el covid-19 tuvo impacto en la distribución de las variables edad y mes de inicio del prenatal. Para entrenar nuevos modelos, preste atención a las variables de entrada y al período utilizado para el entrenamiento. Para soluciones ya establecidas, considere su reentrenamiento.

PALABRAS CLAVE: Prematuridad. Salud. Inteligencia artificial. Aprendizaje automático. COVID-19.

Para citar este artigo: Matapi, J. M. S.; Veiga, H. V. C.; Maciano, F. P. Impacto da pandemia pela covid-19 e modelos de aprendizagem de máquina para predição de nascimentos prematuros nas capitais da região Nordeste do Brasil, 2018-2021 **Cadernos de Estudos Sociais,** v. 37, n. 1, Jan./Jun., 2022.
DOI:10.33148/CESv37n1(2022)2122
Disponível em: http://periodicos.fundaj.gov.b r/index. php/CAD. Acesso em: dia mês, ano.

Esta obra está licenciada com uma Licença Creative Commons Atribuição 4.0 Internacional, sendo permitido que outros distribuam, remixem, adaptem e criem a partir deste trabalho, desde que seja dado ao autor o devido crédito pela criação original e reconhecida a publicação nesta revista.

1 INTRODUÇÃO

A prematuridade, um dos desafios globais na área da saúde perinatal, (definida pela Organização Mundial da Saúde como nascimento antes da 37ª semana de idade gestacional), é uma das principais causas de mortalidade infantil até os cinco anos de idade (PERIN et al, 2022). Globalmente estima-se que, em 2014, nasceram 15 milhões de prematuros, representando de 10,6% do total nascidos vivos (CHAWANPAIBOON, SAIFON, et al, 2014).

No Brasil, um estudo que avaliou a prematuridade entre o período de 2012 a 2019 foi observado um comportamento de redução, variando de 10,87% para 9,95% (MARINELLI et al., 2021). Pesquisas apontam que a ocorrência da prematuridade no Brasil pode ocorrer por fatores associados como: gravidez na adolescência, vulnerabilidade social, baixos níveis de escolaridade e pré-natal inadequado (LEAL et al., 2016).

Já em março de 2020, a covid-19 foi definida como pandemia pela Organização Mundial de Saúde. Com intuito de conter a sua propagação, foram adotadas iniciativas de bloqueios tanto das movimentações geográficas e hábitos e contatos interpessoais (Organização Pan-Americana de Saúde OPAS, 2020). As mudanças nos hábitos ocasionadas pela pandemia de covid-19 também afetaram a vida das gestantes, fato este que pode ter uma relação com a mudança do padrão de exposição aos fatores de risco relacionados à prematuridade, como hábitos de convívio, níveis de estresse, entre outros (CARDOSO, 2021, HEDERMANN et al., 2021).

Com o fim de obter meios que auxiliem a prevenção/redução da prematuridade, abordagens utilizando aprendizagem de máquina para previsão de nascimentos prematuros podem ser elaboradas. Tais métodos buscam identificar, com certa antecedência, se um parto será prematuro ou não. Diferente dos modelos estatísticos inferências, que buscam detectar os fatores que mais influenciam a prematuridade, a aprendizagem de máquina é focada na acurácia das estimativas, a qual busca sempre maximizar alguma métrica que mede assertividade. Na revisão de Lee e Ki Hoon (2020), são discutidas abordagens que fazem uso de algoritmos inteligentes tradicionais como Florestas Aleatórias, k-vizinhos mais próximos, Regressão logística, Máquinas de Vetores de Suporte e outros mais robustos como: Redes Neurais Artificiais e Redes Neurais Convolucionais.

Este artigo tem como objetivo utilizar algoritmos de aprendizado de máquina para predição de prematuridade em gravidez do tipo I (apenas um embrião). A população de

estudo foi concentrada nos nascimentos ocorridos nas capitais dos estados da região Nordeste do Brasil com o intuito de verificar se a pandemia de covid-19 trouxe impacto às estimativas dos modelos de aprendizado de máquina ao longo de 2020 e 2021 ao comparar-se com os anos de 2018 e 2019.

2 MATERIAIS E MÉTODOS

2.1 ÁREA DE ESTUDO

A área de estudo em avaliação neste trabalho consiste de todas as capitais da região Nordeste do Brasil. A região Nordeste é formada por nove estados, totalizando uma área territorial de aproximadamente 1,5 milhão de km², o que equivale a 18% do território Brasileiro, com densidade demográfica de 39,64 habitantes/km. Optou-se por utilizar as capitais, considerando a similaridade populacional e o acesso aos serviços de saúde.

2.2 POPULAÇÃO DE ESTUDO E PERÍODO DE REFERÊNCIA

Os partos prematuros, provenientes de gestação do tipo única, de residentes das capitais da região nordeste, registrados no Sinasc no período de 2018 até 2021. O parto pré-termo (ou prematuro) é definido como aquele cuja gestação termina entre a 20ª e a 37ª semanas ou entre 140 e 257 dias após o primeiro dia da última menstruação.

2.3 FONTE DE DADOS

De acordo com a Secretaria de Vigilância em Saúde, o Sistema de Informações sobre Nascidos Vivos (SINASC), foi implantado oficialmente a partir de 1990, com o objetivo de coletar dados sobre os nascimentos informados em todo território nacional e fornecer dados sobre natalidade para todos os níveis do Sistema de Saúde. Os microdados do Sinasc são coletados por meio da declaração de nascidos vivos (DN), as quais são preenchidas pelos diferentes profissionais de saúde ou parteiras tradicionais responsáveis pela assistência ou parto dos recém-nascidos. As DNs são então encaminhas as Secretarias Municipais de Saúde (SMS) as quais são digitadas, processadas, criticadas e consolidadas no Sinasc local, cujos microdados estão disponíveis no site do próprio Sinasc, contudo, eles precisam de tratamento quanto à descrição das categorias dos dados entre outras características. Por conta disso, para facilitar a sua obtenção, os dados foram adquiridos a partir da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS).

A PCDaS é uma plataforma que facilita a obtenção de dados públicos da área de saúde, com foco principalmente em conjuntos de dados de alto volume e dimensionalidade (número de variáveis). Fornece estrutura para análise e visualização de dados de forma facilitada além da agregação de microdados em uma única sessão, disponibilizando acesso a dados de forma direta por meio de Interface de Programação de Aplicações (APIs) sem a necessidade de *download* direto de arquivos.

Para os anos de 2018 a 2020, a obtenção foi feita na Plataforma de Ciência de Dados aplicada à Saúde (PCDaS), na data de acesso de agosto de 2022, quando os dados já foram tratados pela Fundação Oswaldo Cruz (FIOCRUZ) por metodologia própria (FIOCRUZ, 2018). Para o ano de 2021, os dados (preliminares) foram obtidos diretamente no endereço eletrônico do SINASC (Secretaria de Vigilância em Saúde – Ministério da Sáude, 2022).e, explicando o que é, para que serve etc.

2.4 TRATAMENTO DOS DADOS

Iniciou-se o tratamento, sendo realizada a seleção com relação ao tipo de gravidez, entre os três tipos (I: Única; II: Dupla; III: Tripla e mais) apresentados, por critério de inclusão foi definido o tipo I de gravidez (apenas um embrião). Um processamento prévio foi realizado para o tratamento inicial dos dados com intuito de eliminar os registros com dados faltantes em alguma das variáveis utilizadas na análise.

2.5 VARIÁVEIS DO ESTUDO

As variáveis selecionadas para análise foram aquela de fontes secundárias disponibilizadas no Sinasc e baseadas nos estudos de Lee et al (2020) e Koivu e Sairanen (2020) que também fazem uso de dados demográficos e socioeconômicos, assim como dados clínicos. Uma série de filtragens e regras de truncamento (para variáveis com dados discrepantes) foram feitas para garantir a qualidade dos dados de entrada para a modelagem. Novas variáveis foram criadas com base em variáveis já existentes.

Para a idade gestacional, está sendo considerada de 22 a 31 semanas, levando em conta as seguintes opções: 1: Menos de 22 semanas; 2: 22 a 27 semanas; 3: 28 a 31 semanas; 4: 32 a 36 semanas; 5: 37 a 41 semanas; 6: 42 semanas e mais. Considerando estas, apenas os tipos 2 e 3 foram incluídos. Algumas variáveis de acordo com a distribuição como: idade da mãe, quantidade de gestações, filhos (mortos/vivos) e parto (cesárea / normal) sofreram um truncamento (que é a limitação da amplitude das variáveis numéricas) com relação a distribuição dos dados, conforme a tabela 1. As

observações em que a variável raça/cor da pele foi indígena ou amarela foram retiradas do estudo devido a sua frequência ínfima, o que iria acarretar problemas de estimação nos modelos de aprendizado de máquina.

Tabela 1: Regras de truncamento utilizados nas variáveis em análise que apresentaram dados discrepantes.

Variáveis	Regra de truncamento
Idade	Registro maior que 50 recebe valor 50
Quantidade de gestações anteriores	Registro maior que 5 recebe valor 5
Quantidade de filhos mortos	Registro maior que 4 recebe valor 4
Quantidade de filhos vivos	Registro maior que 5 recebe valor 5
Quantidade de partos do tipo cesárea	Registro maior que 2 recebe valor 2
Quantidade de partos do tipo normal	Registro maior que 6 recebe valor 6

Fonte: Elaborada pelos autores, 2022

Com relação à variável dependente, que serão os partos considerados prematuros, conforme a documentação das variáveis (PCDaS, 2018), a sua definição é dada de acordo com as opções a seguir:

0: Não há indícios de prematuridade;

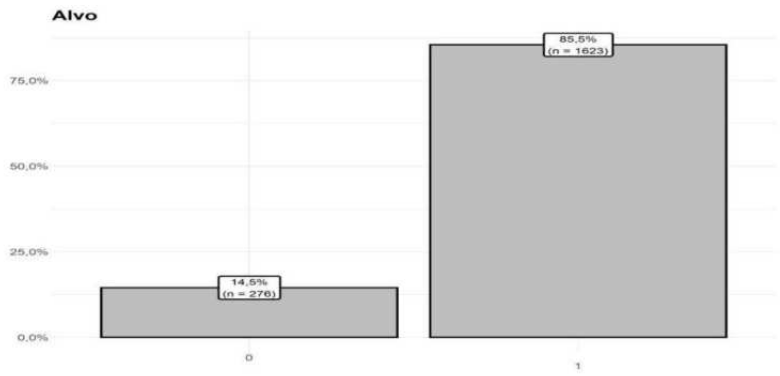
1: Indício de prematuridade dado pela idade gestacional (GESTACAO <= 4);

2: Indício de prematuridade dado pelo peso ao nascer (PESO < 2500);

3: Pré-termo indicado pela idade gestacional e o peso ao nascer.

Dessa forma, foi realizada uma transformação assumindo, assim, 1 na variável dependente, os registros de parto prematuro preenchidos com 3, e 0 caso contrário. Portanto, na definição adotada, temos a variável dependente categorizada da seguinte forma: 0: Não Prematuro e 1: Prematuro. A distribuição da variável dependente (figura 1) foi de 85,5% para a classe alvo (1: Prematuro).

Figura 1 – Frequência absoluta e relativa da variável partos prematuros na base de dados referente ao ano de 2018



Fonte: Elaborada pelos autores (2022)

Também foram criadas novas variáveis como:

$$PropQtdFilhoVivo = \frac{QtdFilhoVivo}{QtdFilhoVivo + QtdFilhoMorto}$$

$$PropQtdFilhoMorto = \frac{QtdFilhoMorto}{QtdFilhoVivo + QtdFilhoMorto}$$

$$SomaQtdFilhoVivoMorto = QtdFilhoVivo + QtdFilhoMorto$$

$$RazaoQtdFilhoVivoMorto = \frac{QtdFilhoVivo}{QtdFilhoMorto}$$

Outras variáveis também foram consideradas no estudo: Quantidade de gestações anteriores, Idade e mês do início do pré-natal e escolaridade. As variáveis proporção de filhos vivos e proporção de filhos mortos, quando divididas por zero (que é quando a mãe nunca teve nenhum filho), receberam o valor de -1. Já quando a mãe não fez pré-natal, na variáveis mês de ínicio do pré-natal, também recebeu valor -1.

O total amostral de cada base analisada e o percentual da variável dependente foram de: 1.899 (85,5% de prematuros) no ano de 2018, 1868 (87,5% de prematuros) no ano de 2019, 1.449 (86,3% de prematuros) no ano de 2020 e 1285 (88,9% de prematuros) em 2021. Um resumo estatístico é apresentado na Tabela 2 para as variáveis numéricas e a Tabela 3 contendo as categóricas.

Tabela 2 - Descrição das variáveis numéricas

Variável	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo	Desvio padrão	Coeficiente de variação
Quantidade de gestações anteriores	0	0	1	1,17	0	5	1,38	1,18
Quantidade de filhos mortos	0	0	0	0,36	0	4	0,71	1,95
Quantidade de filhos vivos	0	0	0	0,85	0	5	1,19	1,40
Quantidade de partos do tipo cesáreo	0	0	0	0,28	0	2	0,55	1,98

Quantidade de partos do tipo normal	0	0	0	0,63	0	6	1,18	1,88
Idade	12	21	27	27,07	21	47	7,31	0,27
Soma da quantidade de filhos vivos e mortos	0	0	1	1,22	0	9	1,48	1,22
Proporção de quantidade de filhos vivos	-1	-1	0	-0,01	-1	1	0,89	-92,04
Proporção de quantidade de filhos mortos	-1	-1	0	-0,24	-1	1	0,71	-2,93
Mês do ínicio do pré-natal	-1	1	2	2,31	1	8	1,75	0,76

Fonte: Elaborada pelos autores (2022)

Tabela 3: - Descrição das variáveis categóricas

Variável	Valores	Quantidade	%
Partos prematuros	0	276	14,5
	1	1623	85,5
Raça/cor da pele*	0Não informado	143	7,5
	01.Branca	234	12,3
	02.Preta	180	9,5
	04.Parda	1342	70,7
Situação conjugal	0.Nenhuma	23	1,2
	01.Solteira/Viúva	948	49,9
	02.Casada	468	24,6
	03.Separada	13	0,7
	04.Em união consensual	447	23,5
Razão quantidade de filhos vivos/mortos	1.Nunca teve filhos	792	41,7
	2.Todos os filhos nasceram mortos	207	10,9
	3.Teve mais filhos mortos que vivos	39	2,1
	4.Teve igual quantidade de filhos mortos e vivos	125	6,6
	5.Teve mais filhos vivos que mortos	736	38,8

Nota: * A raça/cor da pele indígena/amarelo foi retirado da análise devido a % ínfima de observações, menos de 0,5%.

Fonte: Elaborada pelos autores (2022)

3 EXPERIMENTOS

3.1 CONFIGURAÇÃO DE EXPERIMENTOS

Os ambientes de desenvolvimento e linguagens, para a realização das análises exploratórias, ajustes nas bases de dados e as execuções dos modelos foram: a linguagem R (4.1.2) e Python (3.10.4). O sistema operacional Linux, cujas bibliotecas que se destacam são: pandas, tidyverse, scikit-learn e caret. Com relação aos algoritmos de aprendizado de máquina, foram utilizados: Árvore de decisão (DT), Impulso adaptativo (AdaBoost), Regressão logística (LR), Floresta aleatória (RF), Perceptron multicamadas (MLP), Análise Discriminante Linear (LDA). Já os parâmetros foram definidos conforme a Tabela 4, os demais que não foram listados, assumiram o valor padrão da biblioteca scikit-learn (PEDREGOSA et al., 2011).

Para a aplicação dos algoritmos nas bases de dados, um novo processamento foi necessário. Foi realizada a normalização dos dados com intuito da padronização dos valores do conjunto de dados em uma escala comum, sem que se tenha distorções nos intervalos de valores ou perda de informações, utilizando a técnica *Min-Max* scaler.

Com relação às variáveis categóricas, para esta foi usada a técnica *One Hot Encoding* que transforma os atributos categóricos representados como números, dessa forma evitando que os algoritmos interpretem esses valores como sendo numéricos, assim os mesmos são binarizados e cada valor representado assume um valor de 0 ou 1.

Seguindo, a divisão das bases foi feita, sendo inicialmente fixado para o treinamento o ano de 2018, validação o ano de 2019 e por fim os testes nos anos de 2020 e 2021. Uma lista de hiperparâmetros foi definida na Tabela 4 para otimizar a performance dos modelos. Foi feita uma busca intensiva *Grid Search* na base de validação com o objetivo de encontrar uma melhor combinação entre um conjunto de hiperparâmetros.

Tabela 4 – Parâmetros utilizados no Grid search

Parâmetros	Descrição	Valores
Mtry	Número de variáveis amostradas aleatoriamente como candidatas em cada divisão	[1, 2, 3, ..., 60]
Ntree	Número de árvores a crescer	[5, 10, 15, ..., 120]
Size	Número de unidades nas(s) camadas(s) escondida(s)	[5, 10, 15, ..., 120]
learnFuncParams	Taxa de aprendizagem	[0.01, 0.02, ..., 0.1]
Maxit	Máximo de iterações a serem feitas independente da convergência do algoritmo	[100, 150, 200, ...,300]
Minprob	Proporção de observações necessárias para estabelecer um nó terminal	[0.01, 0.02, ..., 0.05]
Maxvar	Número máximo de variáveis que a árvore pode dividir	[3,4,5,...,36]
Maxdepth	Profundidade máxima da árvore	[3, 5, 8, 12, sem restrições (Inf)]

Fonte: Elaborada pelos autores (2022)

3.2 MÉTRICAS DE AVALIAÇÃO

Com relação às diferentes métricas, são incluídas na análise: Acurácia, Precisão, Recall, F1-Score e AUC são avaliados para cada conjunto de dados: treinamento, validação e teste. A definição das métricas são dadas abaixo:

$$Acurácia = \frac{Verdadeiros\ positivos + Verdadeiros\ negativos}{Total\ da\ amostra}$$

$$Precisão = \frac{Verdadeiros\ positivos}{Total\ da\ amostra\ para\ valores\ preditos\ como\ parto\ prematuro}$$

$$Recall = \frac{Verdadeiros\ positivos}{Total\ da\ amostra\ para\ valores\ observados\ como\ parto\ prematuro}$$

$$F1 - Score = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

A AUC é uma métrica estatística que mede a área sob a curva formada entre o Recall (eixo y) e 1- especificidade (eixo x). A especificidade é dada por:

$$Especificidade = \frac{Verdadeiros\ negativos}{Total\ da\ amostra\ para\ valores\ preditos\ como\ parto\ não - prematuro}$$

Podendo ocorrer de serem as métricas bastante semelhantes, o teste *Kruskal-Wallis* será aplicado para identificar se, pelo menos, um dos anos foi diferente dos demais. O teste de *Kruskal-Wallis* é um teste não paramétrico utilizado na comparação de três ou mais amostras independentes (MYLES HOLLANDER; DOUGLAS A. WOLFE, 1973),. Ele indica se há diferença entre pelo menos duas delas. Sua aplicação utiliza os valores numéricos transformados em postos e agrupados num só conjunto de dados. A comparação dos grupos é realizada por meio da média dos postos (HOLLANDER; WOLFE; CHICKEN, 2013).

Com a rejeição da hipótese nula, o teste de *Nemenyi* será feito a fim de realizar as comparações de um ano para o outro (comparações dois a dois ou teste post-hoc). O teste de Nemenyi é um teste post-hoc tem como objetivo encontrar grupos de dados que diferem após um teste estatístico ter rejeitado a hipótese nula e o desempenho das comparações nos grupos de dados é semelhante (HOLLANDER; WOLFE; CHICKEN, 2013).

Os dados utilizados para aplicação dos testes de hipótese mencionados foram feitos com base em amostras da média bootstrap de cada modelo para cada ano. Cada amostra bootstrap continha 30 observações e foi computada 100 vezes para cada modelo e ano. Essa abordagem foi necessária pois alguns modelos utilizados não têm diferença ao realizar as predições quando a base de dados é a mesma, o que não gera variabilidade para possibilitar aplicação dos testes de hipótese, todas as amostras foram feitas com reposição.

A base de dados em análise pode ser classificada como desbalanceado. Essa característica pode superestimar/subestimar algumas métricas de performance como a acúracia e AUC, devendo-se tratar esse problema com alguma técnica de imputação de dados para conter este desbalanceamento (Maione, 2020). Devido ao desbalanceamento das bases, também será realizada a execução do experimento usando técnicas de balanceamento de dados. Para tanto, o pacote ROSE (LUNARDON; MENARDI; TORELLI, 2014) para linguagem R é usado, pois fornece funções para lidar com problemas de classificação binária com classes desbalanceadas. Amostras artificiais balanceadas são geradas de acordo com uma abordagem *bootstrap* suavizada e permitem auxiliar tanto nas fases de estimativa quanto na avaliação da precisão de um classificador binário na presença de uma classe minoritária (MENARDI; TORELLI, 2014).

O estudo foi dispensado de apreciação do Comitê de Ética e Pesquisa por ter, como fonte de informações, dados secundários, agregados de acesso público, que não

possibilitam a identificação individual. Foram respeitados todos os preceitos éticos da Resolução 510, de 2016, da Comissão Nacional de Ética e Pesquisa.

4 RESULTADOS

Os resultados obtidos pelos modelos de aprendizagem de máquina aplicados e suas respectivas descrições são descritos a seguir. Na Tabela 4, são listados os melhores parâmetros em conformidade com cada modelo que foi aplicada à técnica de Grid Search, incluindo as bases balanceadas e desbalanceadas.

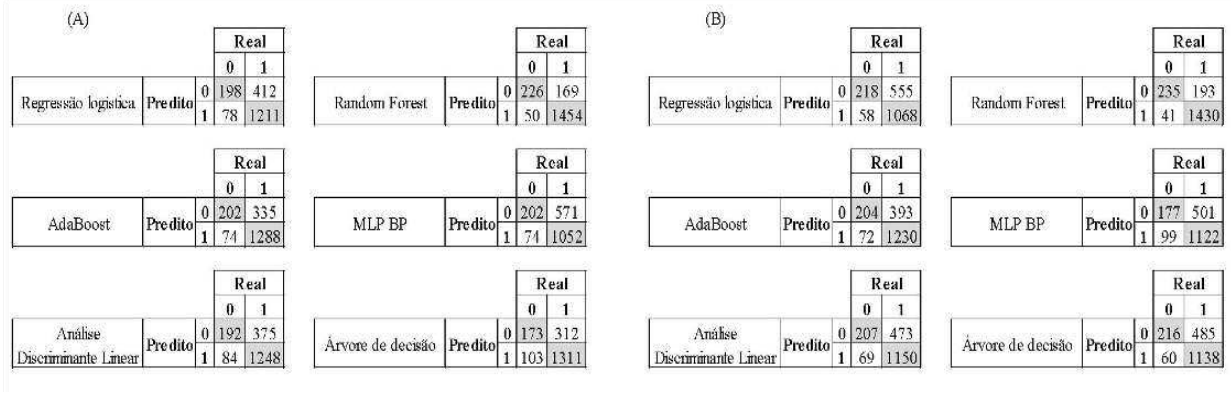
Tabela 5 - Melhores parâmetros Grid search

Parâmetro (Modelo)	Valores	Base balanceada?
mtry (RF)	2	Não
mtry (AdaBoost)	5	
ntree (RF)	150	
size (MLP)	5	
maxit (MLP)	100	
learnFuncParams (MLP)	0.09	
maxdepth (DT)	5	
maxvar (DT)	6	
minprob (DT)	0.02	
mtry (RF)	27	Sim
mtry (AdaBoost)	5	
ntree (RF)	150	
size (MLP)	55	
maxit (MLP)	200	
learnFuncParams (MLP)	0.1	
maxdepth (DT)	8	
maxvar (DT)	6	
minprob (DT)	0.03	

Fonte: Elaborada pelos autores (2022)

A Figura 2 exibe as matrizes de confusão para cada modelo na base de treino e teste, que é a tabulação cruzada entre a classificação predita através de um ponto de corte e a situação real, em que a diagonal principal destacada em cor cinza representa as classificações corretas (parto prematuro e não prematuro) e os valores fora dessa diagonal correspondem a erros de classificação.

Figura 2 – Matriz de confusão base de treinamento (2018) - (A) Base desbalanceada (B) base balanceada



Fonte: Elaborada pelos autores (2022)

A partir destas matrizes, foram obtidas as métricas de performance abordadas, as quais são apresentadas nas Tabelas 5 e 6 para, as bases de treinamento e validação respectivamente.

Tabela 6 -Métricas de performance para a base de dados treinamento

Base balanceada	Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Não	Regressão logística	0,74	0,94	0,75	0,83	0,77
	AdaBoost	0,78	0,95	0,79	0,86	0,81
	Análise Discriminante Linear	0,76	0,94	0,77	0,84	0,78
	Floresta Aleatória	0,88	0,97	0,90	0,93	0,92
	Multilayer Perceptron	0,66	0,93	0,65	0,77	0,74
	Árvore de decisão	0,78	0,93	0,81	0,86	0,76
Sim	Regressão logística	0,68	0,95	0,66	0,78	0,77
	AdaBoost	0,76	0,94	0,76	0,84	0,81
	Análise Discriminante Linear	0,71	0,94	0,71	0,81	0,77
	Floresta Aleatória	0,88	0,97	0,88	0,92	0,94
	Multilayer Perceptron	0,68	0,92	0,69	0,79	0,72
	Árvore de decisão	0,71	0,95	0,70	0,81	0,78

Fonte: Elaborada pelos autores (2022)

Tabela 7: -Métricas de performance para a base de dados validação

Base balanceada	Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Não	Regressão logística	0,76	0,95	0,76	0,85	0,80
	AdaBoost	0,77	0,94	0,79	0,86	0,74
	Análise Discriminante Linear	0,76	0,95	0,77	0,85	0,79
	Floresta Aleatória	0,75	0,93	0,77	0,85	0,76
	Multilayer Perceptron	0,67	0,95	0,66	0,78	0,77
	Árvore de decisão	0,76	0,93	0,78	0,85	0,71

Sim	Regressão logística	0,68	0,95	0,67	0,78	0,78
	AdaBoost	0,74	0,94	0,75	0,84	0,77
	Análise Discriminante Linear	0,72	0,95	0,72	0,82	0,79
	Floresta Aleatória	0,81	0,92	0,85	0,88	0,78
	Multilayer Perceptron	0,70	0,93	0,71	0,81	0,75
	Árvore de decisão	0,69	0,94	0,69	0,80	0,74

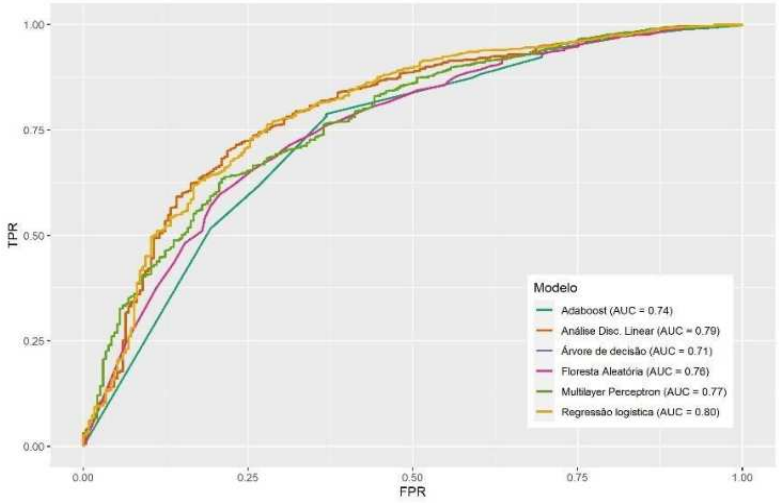
Fonte: Elaborada pelos autores (2022)

Para a base de treinamento desbalanceada (Tabela 5), a métrica acurácia teve uma variação de 0,66 (Multilayer Perceptron) à 0,88 (Floresta aleatória), enquanto para a base de validação desbalanceada (tabela 6), essa variação foi de 0,67 (Multilayer Perceptron) à 0,77 (Adaboost). O modelo Floresta aleatória teve diferenças notórias também nas demais métricas.

O comportamento das métricas de performance são similares também para Precisão, Recall, F1 score e AUC. A precisão foi métrica que obteve a menor variação dentre as demais, de 0,93 (Multilayer Perceptron e Árvore de decisão) à 0,97 (Floresta aleatória) na base de 2018 desbalanceada e 0,93 (Floresta aleatória e Árvore de decisão) à 0,95 (Regressão logística, Análise discriminante linear e Multilayer Perceptron).

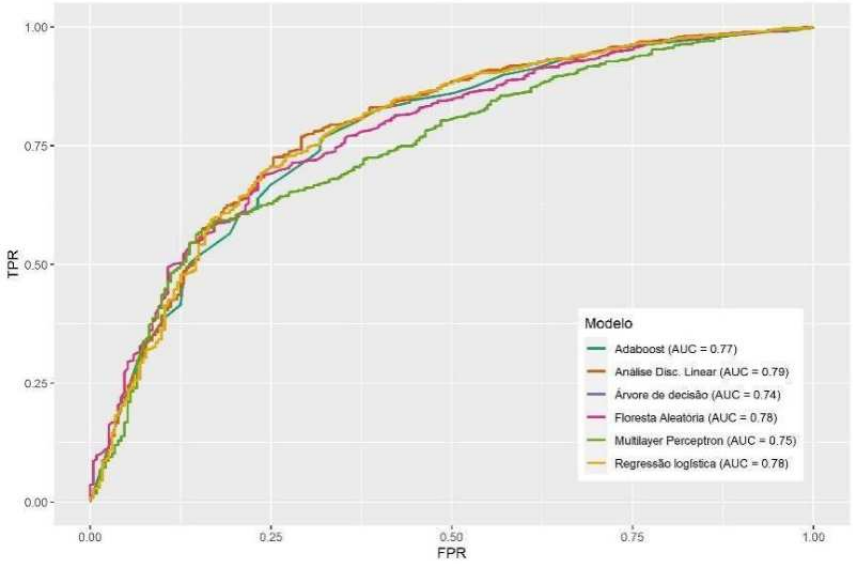
A distribuição das métricas de performance são bastante similares quando observadas para as bases balanceadas, com quedas oscilando em torno de 10% nas métricas acurácia, recall e F1 Score para o modelo de regressão logística quando comparou-se os resultados da base balanceada e desbalanceada, tanto para a base de treino quanto de validação. Nas Figura 3 e 4, são exibidas a área sob a curva (AUC), que relaciona a taxa de falsos positivos (especificidade) e a taxa de verdadeiros positivos (sensibilidade).

Figura 3 – AUC - Validação (base desbalanceada)



Fonte: Elaborada pelos autores(2022)

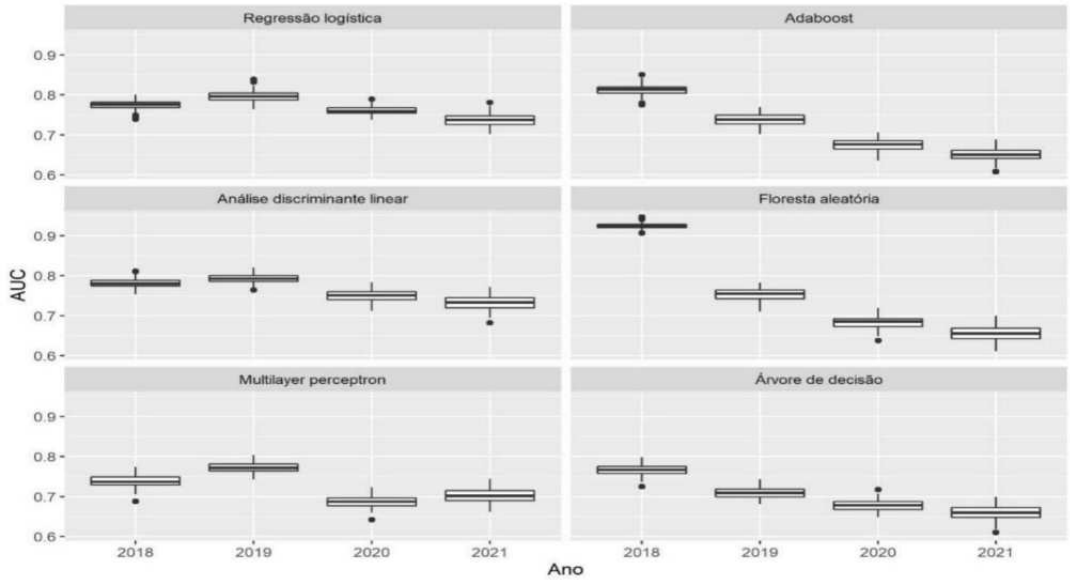
Figura 4 – AUC - Validação (base balanceada)



Fonte: Elaborada pelos autores(2022)

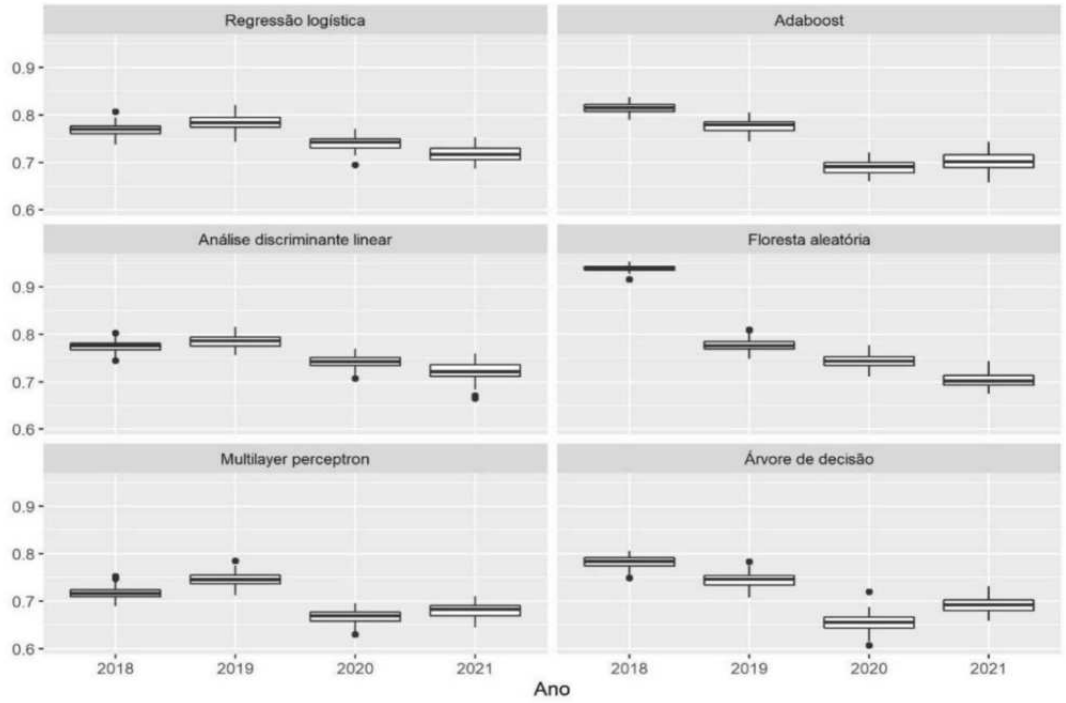
A partir dos gráficos boxplot que contém os resultados da amostra bootstrap da métrica AUC (Figura 5, base desbalanceada e figura 6, base balanceada) para todas as bases de dados (2018, 2019, 2020 e 2021), pode-se verificar que a distribuição dos dados para as bases de dados de 2020 e 2021 para todos os modelos apresentam quedas notórias em sua mediana. Ainda, os modelos Adaboost, Floresta Aleatória e Árvore de Decisão apresentam uma queda já a partir de 2019, mantendo-se em 2020 e 2021. Quanto à Regressão Logística, Análise Discriminante Linear e Multilayer Perceptron, 2019 apresenta um aumento em relação a 2018 e para 2020 e 2021, queda.

Figura 5 – AUC (base desbalanceada)



Fonte: Elaborada pelos autores(2022)

Figura 6 – AUC (base balanceada)



Fonte: Elaborada pelos autores (2022)

Para complementar a análise das Figuras 5 e 6 e verificar a hipótese de pesquisa, será testado estatisticamente se houve diferença entre a AUC ano a ano. O teste de Kruskal-Wallis para todos os modelos alegou diferença (Tabelas 7 e 8). A partir do teste post-hoc de Nemenyi destacaram-se os achados referentes aos modelos: Adaboost, de 2020 e 2021 para 2018, queda de respectivamente, 17% (valor-p < 0,001) e 20% (valor-p < 0,001) enquanto que de 2020 e 2021 para 2019, 9% (Valor-p = 0,001) e 12% (valor-p < 0,001). Floresta Aleatória de 2020 e 2021 para 2018, queda de respectivamente, 26% (valor-p < 0,001) e 29% (valor-p = 0,01) e de 2021 para 2019, queda de 13% (valor-p = 0,04). Na Árvore de decisão houve queda de 2020 para 2018 de 11% (valor-p = 0,01), de 2021 para 2019 de 7% (valor-p = 0,02) e de 2021 para 2020 de 3% (valor-p < 0,001).

Tabela 7: Resultados dos testes de hipótese aplicados a amostra bootstrap da métrica AUC para a base de dados desbalanceada

Ano	AUC bootstrap				Kruskal-Wallis	Dif18-All	Dif19-All	Dif20-All	Modelo
	Média	Mediana	Desvio Padrão	Coeficiente de variação (%)					
2018	0,77	0,78	0,012	1,6	**0,00**				Regressão logística
2019	0,80	0,80	0,014	1,8		0,022			
2020	0,76	0,76	0,011	1,4		**-0,014**	-0,036		
2021	0,74	0,74	0,015	2,0		**-0,038**	**-0,060**	-0,025	

2018	0,81	0,81	0,013	1,6	**0,00**				Adaboost
2019	0,74	0,74	0,015	2,0		-0,074			
2020	0,67	0,68	0,015	2,2		**-0,137**	**-0,064**		
2021	0,65	0,65	0,016	2,4		**-0,162**	**-0,088**	-0,024	
2018	0,78	0,78	0,013	1,7	**0,00**				Análise Discriminante Linear
2019	0,79	0,79	0,012	1,5		0,012			
2020	0,75	0,75	0,014	1,8		**-0,030**	-0,042		
2021	0,73	0,73	0,018	2,4		**-0,049**	-0,061	-0,019	
2018	0,92	0,92	0,008	0,8	**0,00**				Floresta Aleatória
2019	0,75	0,75	0,016	2,1		-0,171			
2020	0,68	0,69	0,016	2,3		**-0,241**	-0,070		
2021	0,66	0,66	0,018	2,7		**-0,268**	**-0,098**	-0,028	
2018	0,74	0,74	0,015	2,0	**0,00**				Multilayer Perceptron
2019	0,77	0,77	0,013	1,7		**0,034**			
2020	0,69	0,69	0,013	1,9		-0,051	-0,086		
2021	0,70	0,70	0,017	2,4		-0,037	**-0,071**	0,015	
2018	0,77	0,77	0,013	1,7	**0,00**				Árvore de decisão
2019	0,71	0,71	0,012	1,7		-0,058			
2020	0,68	0,68	0,013	2,0		**-0,088**	-0,030		
2021	0,66	0,66	0,018	2,8		-0,107	**-0,050**	**-0,020**	
Resultados em negrito significam que foram estatisticamente significantes									
As colunas com prefixo Dif-Ano-All representam a diferença entre as médias da AUC do respectivo ano e os demais									

Fonte: Elaborada pelos autores (2022)

A partir do teste post-hoc de Nemenyi, constatou-se que as diferenças mais relevantes para a base balanceada (Tabela 8) foram: Modelo Adaboost, de 2021 para 2018, queda de respectivamente, 14% (valor-p < 0,001); Multilayer Perceptron obteve aumento de 4% de 2019 para 2018 (valor-p < 0,001) e queda de 10% de 2020 para 2019 (valor-p < 0,001) e de 9% de 2021 para 2019 (valor-p < 0,001); Na Árvore de decisão houve queda de 2019 para 2018 de 5% (valor-p = 0,01), de 2020 para 2018 de 16% (valor-p < 0,001), de 2021 para 2019 de 11% (valor-p < 0,001) e de 2020 para 2019 de 12% (valor-p = 0,01).

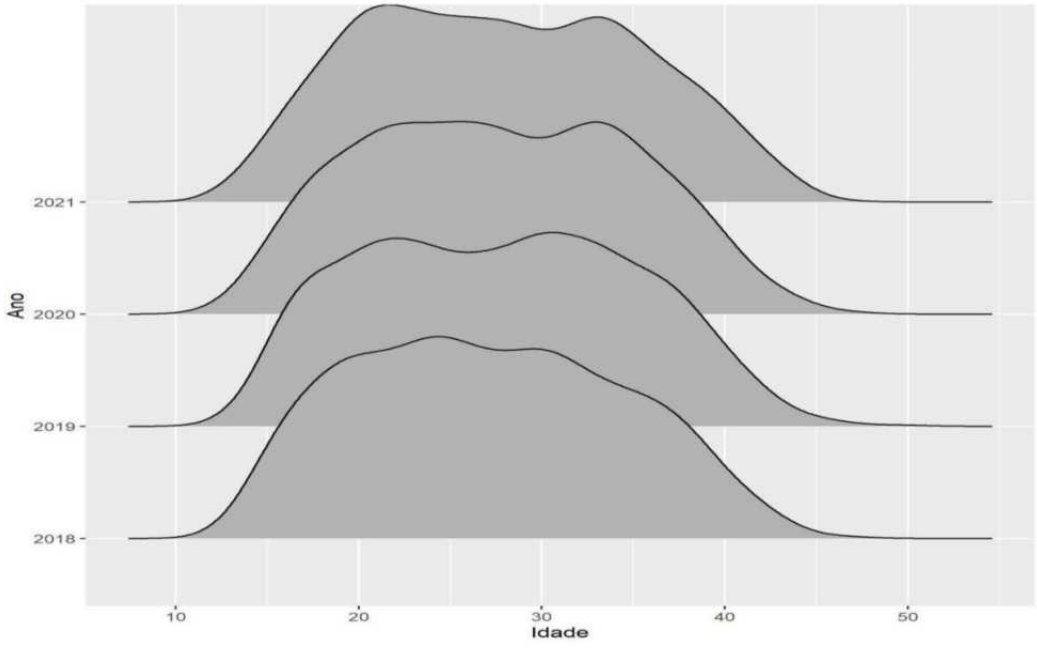
Tabela 8: Resultados dos testes de hipótese aplicados à amostra bootstrap da métrica AUC para a base de dados balanceada

Ano	AUC bootstrap				Kruskal-Wallis	Dif18-All	Dif19-All	Dif20-All	Modelo
	Média	Mediana	Desvio padrão	Coeficiente de variação (%)					
2018	0,77	0,77	0,012	1,6	**0,00**				Regressão logística
2019	0,78	0,78	0,016	2,1		0,015			
2020	0,74	0,74	0,013	1,8		-0,028	**-0,044**		

2021	0,72	0,72	0,016	2,2		-0,051	**-0,067**	-0,023	
2018	0,81	0,81	0,011	1,3	**0,00**				Adaboost
2019	0,78	0,78	0,013	1,7		-0,038			
2020	0,69	0,69	0,015	2,1		-0,126	-0,088		
2021	0,70	0,70	0,018	2,5		**-0,113**	-0,075	0,013	
2018	0,77	0,78	0,013	1,6	**0,00**				Análise Discriminante Linear
2019	0,78	0,79	0,013	1,6		0,010			
2020	0,74	0,74	0,012	1,7		**-0,032**	-0,042		
2021	0,72	0,72	0,018	2,4		**-0,053**	-0,064	-0,021	
2018	0,94	0,94	0,006	0,7	**0,00**				Floresta Aleatória
2019	0,78	0,78	0,013	1,7		-0,162			
2020	0,74	0,74	0,014	1,8		-0,196	-0,034		
2021	0,70	0,70	0,014	2,0		-0,236	-0,073	-0,040	
2018	0,72	0,72	0,013	1,8	**0,00**				Multilayer Perceptron
2019	0,75	0,75	0,013	1,8		**0,028**			
2020	0,67	0,67	0,014	2,1		-0,049	**-0,077**		
2021	0,68	0,68	0,015	2,2		-0,038	**-0,066**	0,012	
2018	0,78	0,78	0,012	1,5	**0,00**				Árvore de decisão
2019	0,74	0,75	0,014	1,9		**-0,039**			
2020	0,65	0,66	0,017	2,6		**-0,128**	**-0,089**		
2021	0,69	0,69	0,016	2,4		**-0,089**	-0,050	0,039	
Nota: Resultados em negrito significam que foram estatisticamente significantes.									
Nota 2: As colunas com prefixo Dif-Ano-All representam a diferença entre as médias da AUC do respectivo ano e os demais									

Fonte: Elaborada pelos autores(2022)

Figura 7 – Distribuição da variável idade de acordo com o ano das bases de dados utilizadas.



Fonte: Elaborada pelos autores (2022)

Ao observar a Tabela 9, é possível identificar que a variável idade e mês do início do pré-natal perderam aderência em relação a distribuição do ano de 2018, isto é, a distribuição estatística entre os anos para estas variáveis mudou de forma significativa. Para a variável idade, houve uma não aderência a partir de 2020, onde pode-se identificar, a partir da figura 7 que a distribuição desta variável está se tornando bimodal (quando há duas modas). Isto quer dizer que está havendo dois grupos principais de mães, um grupo mais jovem (com cerca de 20 anos) e outro mais velho (com cerca de 35 anos), sendo o grupo mais jovem predominante. Tal característica não era mostrada na distribuição da idade para o ano de 2018, onde havia uma concentração de mães com idade em cerca de 26 anos.

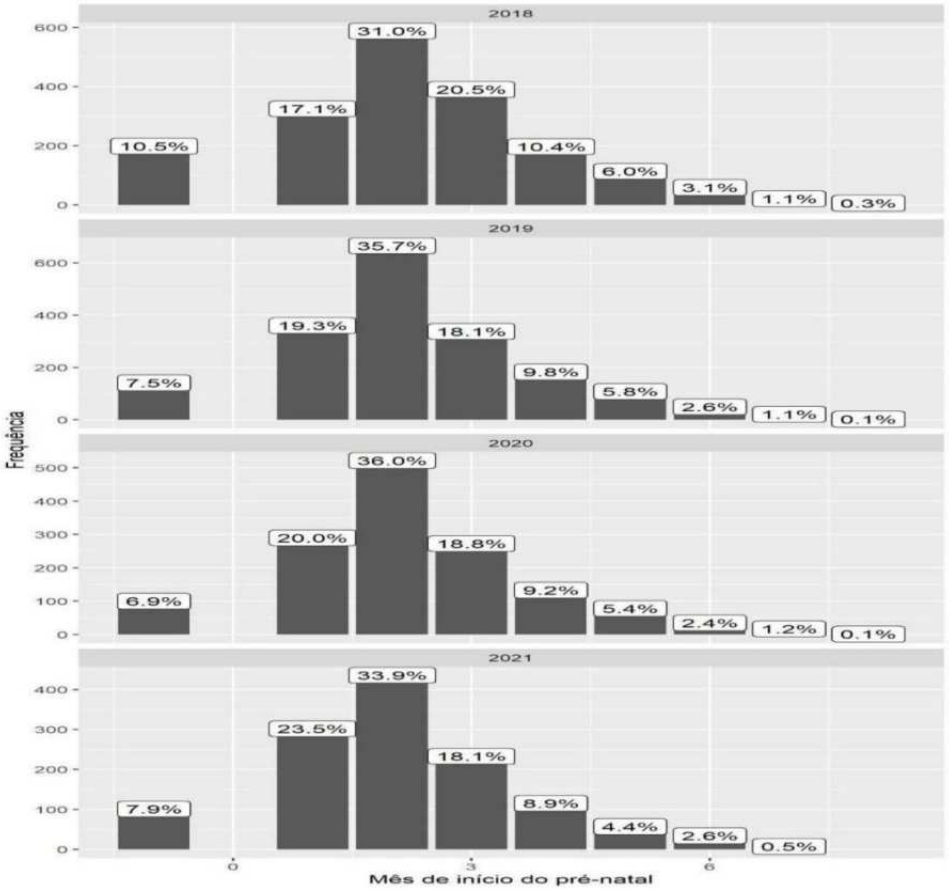
Tabela 9: Testes de aderência entre as variáveis utilizadas nos modelos preditivos, com referência para o ano de 2018 (base de treino)

Variável	Tipo*	2018 vs 2019		2018 vs 2020		2018 vs 2021	
		Estatís-tica	Valor-p	Estatística	Valor-p	Estatística	Valor-p
Quantidade de gestações anteriores	Númerica	0,01	1,00	0,01	1,00	0,01	1,00
Quantidade de filhos mortos	Númerica	0,02	0,85	0,01	1,00	0,02	0,77
Quantidade de filhos vivos	Númerica	0,02	0,78	0,01	1,00	0,01	1,00
Quantidade de partos cesária	Númerica	0,01	1,00	0,02	0,87	0,02	0,90
Quantidade de partos normais	Númerica	0,01	1,00	0,01	1,00	0,02	0,98
Idade	Númerica	0,04	0,07	**0,05**	**0,04**	**0,05**	**0,02**
Soma quantidade de filhos vivos e mortos	Númerica	0,01	1,00	0,01	1,00	0,01	1,00
Razão quantidade de filhos vivos e mortos	Categórica	20,00	0,22	20,00	0,22	20,00	0,22
Proporção da quantidade de filhos vivos	Númerica	0,02	0,95	0,01	1,00	0,02	0,97
Proporção da quantidade de filhos mortos	Númerica	0,03	0,50	0,02	0,84	0,03	0,67
Mês do ínicio do pré-natal	Númerica	0,04	0,11	0,04	0,09	**0,07**	**0,00**
Raça/cor da pele	Categórica	12,00	0,21	8,00	0,24	12,00	0,21
Situação conjugal	Categórica	20,00	0,22	20,00	0,22	20,00	0,22
Escolaridade	Categórica	20,00	0,22	20,00	0,22	20,00	0,22

Nota: *Se o tipo da variável é numérica, foi aplicado o teste de Kolmogorov-Smirnorv e, se categórica, o teste qui-quadrado de Pearson para comparar proporções.

Fonte: Elaborada pelos autores (2022)

Figura 8 – Distribuição da variável mês de início do pré-natal de acordo com o ano, segundo as capitais da região Nordeste, Brasil, período



Fonte: Elaborada pelos autores (2022)

Quanto ao mês de início do pré-natal, só perde aderência a partir de 2021, sendo que o gráfico da Figura 8 deixa bem evidente a mudança das distribuições. Em 2018, havia mais mães que não haviam realizado nenhum pré-natal, quando a maioria iniciava no mês 2 ou 3. Já em 2019 e 2020, a quantidade de mães que não fizeram pré-natal diminui e a quantidade que iniciou o pré-natal no mês 1 ou 3 torna-se mais homogênea. Já para 2021 que é quando constata-se a não aderência, a quantidade que não fez pré-natal se estabiliza em relação a 2019/2020, porém, a quantidade que iniciou o pré-natal no primeiro mês torna-se maior que o quantitativo que iniciou no terceiro mês, divergindo do que era apresentado em 2018.

5 Discussão

Este trabalho propôs a aplicação de modelos de aprendizagem de máquina para classificação, com intuito de identificar (baseado principalmente nas informações das gestantes) se as mulheres com gravidez do tipo I residentes nas capitais nordestinas será prematuro ou não. Também buscou-se identificar se o período que abrange a

pandemia pela covid-19 (em específico para os anos de 2020 e 2021) trouxe impactos significativos para a performance preditiva destes modelos, sendo tal verificação feita por meio dos testes de *Kruskal-Wallis* e Post-Hoc de *Nemenyi* a partir de amostras bootstrap da AUC.

Os modelos apresentaram resultados competitivos com relação às performances preditivas, onde esta constatação é feita ao observar estudo recente como o de Dresse (2022). A métrica AUC teve, em alguns casos, resultados acima do que foi observado em outros trabalhos da área com dados similares (dados não clínicos). Em outro trabalho relacionado, a AUC variou de 0,54 a 0,76 (LEE et al.2020), enquanto que, neste trabalho, alguns modelos obtiveram AUC acima de 0,77, como a Regressão logística (0,80 na base desbalanceada e 0,78 na balanceada) e Análise Discriminante Linear (0,79 em ambas as bases)

Tratando-se dos efeitos da pandemia pela covid-19 sobre os classificadores, as evidências apontam para um impacto sobre todos os classificadores, com ênfase aos mais instáveis, que são aqueles baseados em árvores de decisão: Adaboost, Floresta Aleatória e Árvore de Decisão, os quais apresentaram queda de 2019 para 2018, porém, com um decréscimo maior de 2021 e 2020 em comparação a 2018 e 2019, onde o teste de *Nemenyi* encontrou diferenças 2 a 2 principalmente para a base de treinamento desbalanceada. Tal resultado mostra que esses modelos já eram instáveis antes da pandemia, e que o período pandêmico, agravou mais essa situação.

Os modelos que tiveram estabilidade, em 2019, ou aumento, em relação a 2018, são bastante consolidados na literatura: Regressão Logística, Análise Discriminante linear e Multilayer Perceptron. Estes modelos, ainda tiveram queda em 2020 e 2021, sendo que em 2019, em alguns casos, o teste de Nemenyi alegou que houve um aumento na métrica AUC, o que não é o ideal, mas é melhor do que se houvesse um declínio.

Um outro ponto que deve ser considerado é a causa da perda de performance dos modelos. Tal queda pode ser evidenciada pela perda de aderência das variáveis idade e mês de início do pré-natal à base de treinamento. A variável idade começou a apresentar distribuição bimodal, quando as concentrações de idade são em torno de 20 e 35 anos, mostrando que está havendo uma queda de mães com idade no intervalo entre as modas e, por sua vez, na formação de dois grupos distintos. Já a variável do mês de início de pré-natal, apresentou uma melhora no quesito de precaução, já que a

quantidade de mães que não informaram os dados sobre esse evento diminuiu, e a proporção de mães que iniciou o pré-natal no primeiro mês aumentou.

6 CONCLUSÃO

Conclui-se então que os modelos baseados em árvores de decisão devem ser utilizados com cuidado, visto que a maioria não são aderentes à base de treinamento. Já em análise para todos os modelos, há indícios de que a covid-19 afetou a estabilidade deles. Portanto, as soluções elaboradas com bases de dados antes dessa pandemia devem ser monitoradas com cautela, e caso necessário, a realização de seu re-treinamento deve ser considerada. Tendo em conta como opção para trabalhos futuros, são elegíveis: abranger uma maior cobertura da idade gestacional; adicionar mais variáveis (sociodemográficas e/ou clínicas) e considerar outros modelos estatísticos ou de aprendizagem de máquinas.

REFERÊNCIAS

CARDOSO, Paôla. **Prematuridade durante a pandemia de covid-19 em vigência de medidas restritivas: uma revisão integrativa.**, 2021.

CHAWANPAIBOon, Saifon, et al. **Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis.** The Lancet Global Health 7.1 (2019): e37-e46. (LINHA 104).

DO CARMO LEAL, M., Esteves-Pereira, A.P., Nakamura-Pereira, M., Torres, J.A., Theme-Filha, M., Domingues, R.M.S.M., Dias, M.A.B., Moreira, M.E., Gama, S.G.: **Prevalence and risk factors related to preterm birth in brazil**. Reproductive health 13(3), 163–174 (2016).

FIOCRUZ. **Plataforma de ciência de dados aplicada à saúde. laboratório de informação em saúde (lis). instituto de comunicação e informação científica e tecnológica em saúde (icict). fundação oswaldo cruz (fiocruz).** 2018. Disponível em: "https://bigdata-metadados.icict.fiocruz.br/dataset/sistema-de-informacoes-de-nascidos-vivos-sinasc"

HARRISON, M.S., Goldenberg, R.L.:**Global burden of prematurity. *In*: Seminars in fetal and neonatal medicine.** vol. 21, pp. 74–79. Elsevier (2016).

HEDERMANN G. et al. **Danish premature birth rates during the COVID-19 lockdown.** Archives of Disease in Childhood-Fetal and Neonatal Edition, v. 106, n. 1, p. 93-95, 2021.

LEE, Kwang-Sig, and Ki Hoon Ahn. **Application of artificial intelligence in early diagnosis of spontaneous preterm labor and birth.** Diagnostics 10.9 (2020): 733.

MAIONE, Camila. **Balanceamento de dados com base em oversampling em dados transformados.** (2020).LEE, Kwang-Sig et al. Determinants of spontaneous preterm

labor and birth including gastroesophageal reflux disease and periodontitis. Journal of Korean medical science, v. 35, n. 14, 2020.

MARTINELLI KG. et al. **Prematuridade no Brasil entre 2012 e 2019: dados do Sistema de Informações sobre Nascidos Vivos.** *Revista Brasileira de Estudos de População* [online]. 2021, v.38 acesso em: 12 ago.2022, e0173. Disponível em: https://doi.org/10.20947/S0102-3098a0173.

MYLES Hollander and Douglas **A. Wolfe** (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 115–120.

OPAS/OMS. **Folhetim OMS - Folha informativa – COVID-19: doença causada pelo novo coronavíru**s. Organização Pan-Americana de Saúde, 12 de Março de 2020.

PCDAS. **Sistema de Informações de Nascidos Vivos - SINASC.** 2018. Disponível em:"https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacao-sobre-nascidos-vivos/dicionario-de-variaveis/".

PEDREGOSA, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel,O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: **Machine learning in Python.** ***Journal of Machine Learning Research***, v. 12, p. 2825–2830, 2011.

PERIN J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong KL, et al. **Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the Sustainable Development Goals.** Lancet Child Adolesc Health. 2022;6(2):106-15. doi:10.1016/S2352-4642(21)00311-4

SVC / MS. **Sistema de Informações de Nascidos Vivos - SINASC.** Disponível em: "url:https://svs.aids.gov.br/daent/cgiae/sinasc/apresentacao/". Acesso em: 2022.

VOGEL, Joshua P., et al. **The global epidemiology of preterm birth.** Best Practice & Research Clinical Obstetrics & Gynaecology 52 (2018): 3-12.