UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA - CIN
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

JOSÉ MAURÍCIO MATAPI DA SILVA

Impacto da Pandemia da COVID-19 e modelos de aprendizagem de máquina para predição de prematuridade no Brasil

Recife

2022

JOSÉ MAURÍCIO MATAPI DA SILVA

Impacto da Pandemia da COVID-19 e modelos de aprendizagem de máquina para predição de prematuridade no Brasil

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como parte dos requisitos necessários para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Inteligência computacional

Orientador (a): Fernando Maciano de Paula Neto

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586i Silva, José Maurício Matapi da
Impacto da pandemia da COVID-19 e modelos de aprendizagem de máquina para predição de prematuridade no Brasil / José Maurício Matapi da Silva. – 2022.
59 f.: il., fig.

Orientador: Fernando Maciano de Paula Neto.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Aprendizado de máquina. 3. COVID-19. I. Paula Neto, Fernando Maciano de (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-15

José Maurício Matapi da Silva

“Impacto da Pandemia da COVID-19 e modelos de aprendizagem de máquina para predição de prematuridade no Brasil”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 13/09/2022.

BANCA EXAMINADORA

_____

Prof. Dr. Adiel Teixeira de Almeida Filho
Centro de Informática / UFPE

_____

Profa. Dra. Cristine Vieira do Bonfim
Instituto de Pesquisas Socias / FUNDAJ

_____

Prof. Dr. Fernando Maciano de Paula Neto
Centro de Informática / UFPE
(**Orientador**)

Este trabalho é dedicado a toda minha família, em especial ao meu irmão Daniel Matapi da Silva.

AGRADECIMENTOS

Agradeço a Deus por me conceder o dom da vida, saúde e conhecimentos para que eu enfrentasse mais essa jornada em minha vida. Agradeço também a minha mãe Saturia Matapi e Danilo Matapi por me fazerem refletir enquanto pessoa nas tomadas de decisões que fiz.

À minha esposa Cássia Ponte pela paciência e companheirismo durante toda essa caminhada, visto que foram grandes os desafios, pois além do mestrado Deus nos presenteou com nossa linda filha Eloá que nos fortalece para sempre estarmos superando os desafios da vida.

Aos meus colegas Miguel Ortiz e Felipe pela troca de experiência durante as disciplinas, pelas conversas na tentativa de resolução de cada problema que nos eram postos. Outro grande colega é o Heitor Veiga que sempre esteve colaborando em algumas fases desta pesquisa, assim como a troca de acontecimentos quase que diária dentro e fora do universo acadêmico. Ao meu amigo Johnny Mayron que também sempre me incentivou a conclusão desta pesquisa.

Agradeço, sobretudo ao meu orientador Fernando Maciano de Paula Neto pela paciência, conhecimento e pela sua disponibilidade para seguir em frente com a pesquisa. Também gostaria de registrar meu agradecimento à banca examinadora composta pelos membros internos Adiel Teixeira De Almeida Filho e Leandro Maciel Almeida, assim como os externos, a professora Cristine Vieira do Bonfim e Conceição Maria de Oliveira, pela pronta disponibilidade em dar suas colocações nesta pesquisa.

Agradeço a FACEPE, pelo suporte estrutural dado durante um período da pós-graduação.

RESUMO

Prematuridade é quando a criança nasce com menos de 37 semanas completas de gestação, sendo considerado um problema de saúde global, e ainda uma das principais consequências de mortes em neonatais e infantis menores de cinco anos de idade. A taxa de parto prematuro pode variar de acordo com a região geográfica e o nível de renda, mantendo uma maior frequência em países subdesenvolvidos. Nos países desenvolvidos, ele é amplamente avaliado como forma de compreender as causas e na criação de ações preventivas. Nesta pesquisa, foi proposta a utilização de algoritmos de aprendizado de máquina para predição de parto prematuro em em gestantes únicas, utilizando dados das capitais brasileiras. Foi verificado se os dois primeiros anos (2020-2021) da pandemia COVID-19 trouxeram impactos significativos para as estimativas dos modelos testados, em comparação ao que foi constatado na base de treinamento. Foram utilizados 6 classificadores de aprendizagem de máquina: Árvore de Decisão, Floresta Aleatória, Regressão Logística, Adaptive Boosting, Análise de Discriminante Linear e Rede Neural do tipo Multi-layer Perceptron, analisando as métricas de acurácia, precisão, revocação, F1-SCORE e área sobre a curva *r*eceiver *o*perating *c*haracteristic. Portanto, com o processamento desses resultados, foi possível verificar a predição de parto prematuro com dados secundários no período de pandemia. A AUC dos modelos na base de validação variou de 0,7052 a 0,7729 (base sem balanceamento) e de 0,7199 a 0,7717 (base com balanceamento). Os resultados demonstraram que a COVID-19 impactou os modelos de Regressão logística, Análise discriminante linear e Multilayer perceptron (os quais são considerados estáveis), enquanto que os modelos baseados em árvore (Adaboost, Floresta aleatória e Árvore de decisão) não apresentam boa aderência à base de treino, devendo ser utilizados com cautela ou desconsiderados.

Palavras-chaves:parto prematuro; recém-nascido prematuro; nascido vivo; inteligência artificial; aprendizado de máquina; COVID-19.

ABSTRACT

Prematurity is when a child is born with less than 37 completed weeks of gestation, being considered a global health problem, and still one of the main consequences of deaths in neonatal and five-year-old children. The premature birth rate may vary according to geographic region and income level, maintaining a higher frequency in underdeveloped countries. In the countries included, it is widely considered as a way of understanding the causes and creating preventive actions. In this research, the use of machine learning algorithms was proposed to predict premature birth in singletons, using data from Brazilian capitals. It was verified whether the first two years (2020-2021) of the COVID-19 pandemic had impacts on the estimates of the tested models, compared to what was found in the training base. Six machine learning classifiers were used: Decision Tree, Random Forest, Logistic Regression, Adaptive Boosting, Linear Discriminant Analysis and Multi-layer Perceptron Neural Network, analyzing the metrics of accuracy, precision, recall, F1-SCORE and area under curve *r*eceiver *o*perational *c*haracteristic. Therefore, with the processing of these results, it was possible to verify the prediction of premature birth with secondary data in the pandemic period. The AUC of the models in the validation base ranges from 0.7052 to 0.7729 (unbalanced base) and from 0.7199 to 0.7717 (balanced base). The impressive results that COVID-19 impacted Logistic Regression, Linear Discriminant Analysis and Multilayer perceptron models (which are considered stable), while tree-based models (Adaboost, Random Forest and Decision Tree) did not show good adherence based on training, and should be used with caution or disregarded.

Keywords:premature birth; premature newborn; born alive; artificial intelligence; machine learning; COVID-19.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	Adaptive Boosting
Covid-19	Síndrome respiratória aguda grave - SARS-CoV-2
DT	Decision Tree
DUM	Data da Última Menstruação
IG	Idade Gestacional
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MLP	Multilayer Perceptron
OMS	Organização Mundial da Saúde
PCDaS	Plataforma de Ciência de Dados aplicada à Saúde
RF	Random Forest
SINASC	Sistema de Informações sobre Nascidos Vivos

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

A prematuridade está intimamente associada à morbimortalidade infantil, sendo uma das principais causas de morte no período neonatal (óbitos de 0 a 27 dias de vida) (VOGEL et al., 2018). Dos 15 milhões de bebês nascidos prematuros a cada ano em todo o mundo, mais de um milhão morreu antes dos cinco anos de idade devido à prematuridade e suas complicações. Houve um declínio geral nas mortes de crianças menores de cinco anos nas últimas duas décadas devido à redução da mortalidade relacionada a doenças infecciosas (VOGEL et al., 2018). Como resultado, as complicações relacionadas à prematuridade são agora a principal causa de morte entre crianças, representando 18% de todas as mortes em crianças menores de 5 anos (HUG; SHARROW; YOU, 2017).

Dentre as complicações de curto prazo da prematuridade estão o aumento de risco para doenças respiratórias neonatais (como síndrome do desconforto respiratório e displasia broncopulmonar), enterocolite necrosante, sepse, condições neurológicas (como leucomalácia periventricular, convulsões, hemorragia, paralisia cerebral e encefalopatia hipóxico-isquêmica), bem como dificuldades de alimentação e problemas visuais e auditivos (MWANIKI et al., 2012).

A Organização Mundial da Saúde (OMS) define o parto prematuro como o parto que ocorre antes das 37 semanas de gestação, ou que ocorre antes de 259 dias a partir da Data da Última Menstruação (DUM) (HOWSON, 2012), podendo acontecer o trabalho de parto espontâneo ou por indução, e independente do peso do recém-nascido (DAMASO, 2016), isto é, uma condição definida pela interrupção da gestação antes de atingir um determinado período de tempo, independente da presença de sinais e sintomas específicos (KRAMER et al., 2012).

A prematuridade não é caracterizada por grupo homogêneo, pois o grau de maturidade fisiológica difere com a Idade Gestacional (IG). Alguns estudos propõem uma subclassificação mais adequada dividindo os RN prematuro em: prematuro extremo (IG $<$28 semanas), muito prematuro (IG entre 28 e 32 semanas) e prematuro tardio (IG entre 32 - 37 semanas incompletas) (CHERMONT et al., 2020) (FERREIRA et al., 2011). As inferências para o nascimento antes da idade gestacional de 37 semanas perpassa por todo o ciclo de vida, em muitas das vezes esses bebês vão exigir cuidados especiais pois enfrentam maiores possibilidades de desenvolver graves problemas de saúde (VASQUEZ, 2018).

A prematuridade propicia a probabilidade de complicações respiratórias, gastrintestinais e do neurodesenvolvimento (HUSEYNOVA et al., 2021). Sendo também uma das principais causas de morbidade e mortalidade neonatal precoce e tardia, superando até mesmo os defeitos congênitos, representando assim, um grande problema de saúde pública (ARNAEZ et al., 2021).

As últimas estimativas sugerem que as complicações do parto prematuro foram a principal causa de morte em crianças menores de cinco anos em todo o mundo em 2016 e representam aproximadamente 16% de todas as mortes em crianças menores de cinco anos e 35% das mortes entre recém-nascidos (VOGEL et al., 2018).

No Brasil, em um estudo que avaliou a prematuridade entre o período de 2012 a 2019, foi observado que a proporção de prematuridade apresentou um comportamento de redução, variando de 10,87% a 9,95% (MARTINELLI et al., 2021).

A quantidade e a qualidade dos dados de nascimentos prematuros relatados na maioria dos países, inclusive no Brasil, traz uma certa dificuldade com relação a sua utilização; uma vez que quando se tem disponibilização como a feita pelo Ministério da Saúde brasileiro, o Sistema de Informações sobre Nascidos Vivos (SINASC), não há uma padronização de disponibilização dos dados (genérico). Iniciativas como a Plataforma de Ciência de Dados aplicada à Saúde (FIOCRUZ, 2018), tem como objetivo amenizar essa questão de forma a facilitar a gestão e análise de grandes quantidades de dados na área de saúde para os pesquisadores.

A prematuridade apresenta-se como um problema em todo o mundo. Em países desenvolvidos ele é amplamente avaliado como forma de compreender as causas e na criação de ações preventivas. Uma das iniciativas que vem sendo empregada nos últimos anos é a inteligência artificial, na literatura (LEE; AHN, 2020) existem trabalhos abordando o uso da inteligência artificial como proposta de solução de auxílio a prevenção da prematuridade, com resultados que fornecem suporte, com estimativas mais precisas para a tomada de decisão em práticas clínicas.

1.1.1 Prematuridade e COVID-19

Em 12 de Março de 2020 a Organização Mundial de Saúde (OMS) declarou a Síndrome respiratória aguda grave - SARS-CoV-2 (Covid-19) como uma pandemia, o que levou a inúmeros bloqueios tanto das movimentações geográficas quanto dos hábitos e contatos interpessoais (OPAS, 2020). Nesse contexto, em virtude de todas as medidas restritivas impostas pelo Estado para o período pandêmico, houve uma mudança importante da rotina diária na vida das

gestantes, fato este que pode ter uma correlação importante com a mudança do padrão de exposição aos fatores de risco relacionados a prematuridade, como hábitos de vida, níveis de estresse, esforço físico, entre outros (CARDOSO, 2021) (HEDERMANN et al., 2021).

As alterações da gravidez envolvem os sistemas cardio respiratório e imunológico, o que pode resultar em uma resposta alterada à infecção por SARS-CoV-2 na gravidez (WASTNEDGE et al., 2021). Os fetos podem ser expostos ao SARS-CoV-2 durante períodos críticos do desenvolvimento fetal (WEI et al., 2021). A natureza da associação entre o COVID-19 e os resultados na gravidez permanece incerta. Em um estudo de meta-analise que avaliou a associação entre a infecção por coronavírus 2 da síndrome respiratória aguda grave (SARS-CoV-2) durante a gravidez em 2019, no qual foram incluídos 42 estudos envolvendo 438.548 grávidas foi observado que o COVID-19 pode estar associado ao parto prematuro (WEI et al., 2021).

O aumento de natimortos pode ser resultado de efeitos indiretos, como relutância em ir ao hospital quando necessário (por exemplo, com movimentos fetais reduzidos), medo de contrair infecção, não querer sobrecarregar o sistema de saúde, ou até mesmo uma consequência direta de infecção por SARS-CoV-2 (ASHISH at al., 2020; KHALIL et al., 2020). (SENTILHES et al., 2020) também estudou sobre a correlação da COVID-19 na gravidez e observou que houve associação entre a morbidade materna e parto prematuro com a COVID-19. Sua associação com outros fatores de risco bem conhecidos para morbidade materna grave em gestantes sem infecção, incluindo idade materna acima de 35 anos, sobrepeso e obesidade (SENTILHES et al., 2020).

Além desses fatores, a infecção por SARS-CoV-2 também pode diretamente provocar respostas inflamatórias sistêmicas exacerbadas envolvidas na patogênese do parto prematuro. A má perfusão vascular fetal placentária foi encontrada em achados histopatológicos placentários em pacientes com COVID-19 no parto, (PATBERG et al., 2021)podem contribuir para o baixo crescimento fetal, natimorto e prematuridade (WEI et al., 2021).

1.2 TRABALHOS RELACIONADOS

Esta seção apresentar alguns trabalhos fazendo o uso de aprendizagem de máquina na área da saúde, que em pontos tem semelhanças com o que foi proposto neste projeto.

A aprendizagem de máquina cada vez mais vem sendo aplicada na área da saúde, auxiliando na tomada de decisão nos diversos momentos da atenção à saúde (OBERMEYER; LEE, 2017). A seguir são apresentadas algumas pesquisas que fazem uso da aprendizagem de máquina como

proposta de solução para problemas que envolve a área da saúde.

Em sua pesquisa (REZAEIAN et al., 2020), propõe um estudo comparando dois modelos: rede neural e regressão logística, para predição de mortalidade em neonatos prematuros admitido em uma unidade de tratamento intensivo neonatal. A pesquisa foi realizada com base nas informações de 1.618 neonatos cujo os dados foram coletados de prontuários de pacientes de todos recém-nascidos com menos de 37 semanas (prematuro) nascidos entre 2010 e 2017 e foram admitidos na UTIN de Ghaem Hospital, Mashhad, Irã. Seus resultados apontam que o Multilayer Perceptron(MLP) apresentou 95% de área sob a Curva ROC AUC aceitável em comparação com a regressão logística, onde foi capaz de prever com maior assertividade o risco de mortalidade neonatal.

Como pesquisa (MAS-CABO et al., 2019) propõe o desenvolvimento e comparação de diferentes classificadores, baseados em redes neurais, para prever o parto prematuro usando recursos da atividade contrátil uterina, que gera um sinal eletromiográfico (denominado de eletrohisterograma). Os dados utilizados foram no Departamento de Obstetrıcia e Ginecologia, Medical Center Ljubljana, em Ljubljana. Os registros foram coletados da população em geral bem como dos pacientes internados no hospital com o diagnóstico de prematuridade. Destes, 300 registros de Eletroencefalograma (EHG) de mulheres grávidas, 262 que tiveram parto a termo e os 38 restantes de mulheres tiveram prematuridade. Todos os registros de EHG foram realizados individualmente em mulheres entre 22 e 32 dias de idade gestacional. O desempenho dos classificadores também foi avaliado quando os dados obstétricos foram incluídos, houve melhora significativa nas métricas do classificador alcançando uma AUC 91,1% . Estes resultados são otimistas quanto ao uso do EHG para a previsão de parto prematuro.

O estudo proposto por (LEE et al., 2020) é o uso do aprendizado de máquina para predizer nascimento prematuro a partir da análise de dados de doença do refluxo gastroesofágico (DRGE) e periodontite, pois a doença a do refluxo gastroesofágico é comum durante gravidez e espera-se que esteja relacionado com periodontite, que está associada ao parto pré-termo. Os dados utilizados foram do Hospital Anam em Seul, Coréia, com 731 pacientes obstétricos entre o período de 1995 a 2018. Seis modelos de aprendizado de máquina foram aplicados e comparados. Em termos de precisão, usando algoritmo de aprendizado de máquina Random forest (0,8681%) e Logistic regression (0,8736%) obtiveram resultados semelhantes. Com base na importância variável foi constatado que a DRGE é mais importante do que a periodontite para prever e prevenir o parto prematuro.

Dentre os trabalhos analisados nessa seção, os de Mas-Cabo (MAS-CABO et al., 2019) e

Kwang-Sig Lee (LEE et al., 2020) trazem uma proposta focando também na areá de parto prematuro,no entanto, focam em alguns exames bem específicos. Já Rezaeian (REZAEIAN et al., 2020), que apesar do foco ser em mortalidade neonatal, faz uso de variáveis similares às que são propostas nesta pesquisa.

1.3 OBJETIVOS

Este estudo tem como objetivo principal aplicar algoritmos de aprendizado de máquina para predição de parto prematuro em gestantes únicas, utilizando dados das capitais brasileiras e verificar se nos dois primeiros anos da pandemia COVID-19 trouxeram impactos significativos para as distribuições das variáveis contidas na base de dados, em comparação ao que foi utilizado para treinamento dos modelos. Os objetivos específicos são:

- Aplicar os modelos de aprendizagem máquina e armazenado as métricas acurácia, precisão, revocação, F1-SCORE e área sobre a curva ROC;
- Analisar os resultados de forma comparativa, referente a aderência dos modelos na predição de parto prematuro e verificar os possíveis impactos que a COVID-19 trouxe para as estimativas dos modelos testados, em comparação ao que foi constatado na base de treinamento.

1.4 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está organizado em cinco capítulos descritos a seguir:

- Capítulo 1: Apresenta a contextualização deste trabalho, o problema a cerca dos partos pré-termo e Covid-19, trabalhos relacionados e os objetivos.
- Capítulo 2: Expõe a fundamentação teórica necessária como os modelos de aprendizagem de máquina.
- Capítulo 3: Demonstrar todo o fluxo definido para a realização dos experimentos além das métricas de avaliação dos modelos de aprendizagem de máquina.
- Capítulo 4: Traz a descrição dos resultados obtidos nos experimentos realizados.
- Capítulo 5: Faz as considerações finais desta pesquisa, algumas sugestões para trabalhos futuros e em seguida as referências.

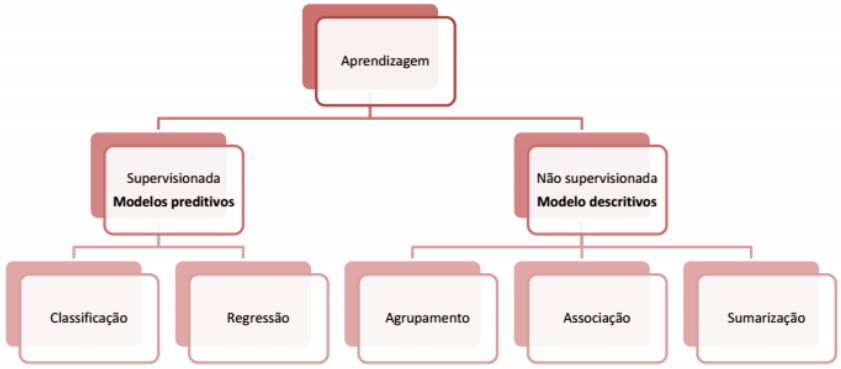
2 FUNDAMENTAÇÃO TEÓRICA

2.1 INTELIGÊNCIA ARTIFICIAL

Quando o assunto é Inteligência artificial, o pensamento inicial é voltado para ideias futuristas, como as do robôs vestidos de mordomos e servindo seu chefe, mas essa já deixou de ser há alguns anos. O filtro de spam, por exemplo, foi o primeiro aplicativo a realmente se destacar e conquistar o mundo na década de 1990 (GÉRON, 2019). Cada vez mais surgem sistemas que imitam a inteligência do homem e aprimoram suas tarefas à medida que coletam informações. Embora haja essa superstição de que vão dominar o mundo, o objetivo da Inteligência artificial não é substituir o homem, mas sim contribuir significativamente para a humanidade (ORACLE, 2014).

Sendo um dos ramos da inteligência artificial, o aprendizado de máquinas vem de uma ciência que aprende através de dados passados a elas. Podemos exemplificar a caixa de spam do e-mails. O programa usa um conjunto de de e-mails como treinamento e aprende a identificar um e-mails como spam (GÉRON, 2019). A figura 1 apresenta um modelo representativo de como é a hierarquia dos aprendizados de máquina. Além do aprendizado supervisionado e não supervisionado, também temos o aprendizado por reforço.

Figura 1 – Hierarquia dos tipos de aprendizado



Fonte: Italo (2018)

O aprendizado não-supervisionado não possui atributos de saída ou variáveis respostas. O algoritmo explora a regularidade dos dados, um exemplo de aprendizado não supervisionado é o modelo de agrupamentos. Já no aprendizado por reforço, o algoritmo descobre, por tentativa e erro, a solução que oferece melhor recompensa (ITALO, 2018). Classificar é uma tarefa típica do aprendizado supervisionado. O filtro de spam é um bom exemplo disto. Também é possível

prever variáveis de valores numéricos, como o preço de um carro a partir de suas características (quilometragem, ano, marca, etc), tipo a que chamamos de tarefa de regressão.

Alguns algoritmos de regressão também podem ser utilizados para classificação, assim como os algoritmos de classificação podem ser utilizados para problemas de regressão. Como a Regressão Logística onde o algoritmo classifica um determinado objeto de acordo a probabilidade de pertencer a ela (GÉRON, 2019).

Nesta pesquisa, será utilizado o aprendizado supervisionado, em que é fornecido ao sistema de aprendizado um conjunto de exemplos (também conhecido na literatura como *caso*, *registro* ou *dado*), que é uma tupla de valores de atributos (ou um vetor de valores de atributo), o qual cada exemplo possui um rótulo associado. Tal rótulo é o que define a classe a qual o exemplo pertence)(BARANAUSKAS et al., 2018).

2.2 ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

2.2.1 Análise Discriminante Linear

Proposta inicialmente por Ronald Fisher em 1936 para análises bivariadas, foi generalizada para a multivariada em 1948, por Calyampudi Radhakrishna Rao. Uma variável estatística discriminante é uma combinação linear de duas variáveis independentes (HAIR et al., 2009).

No momento da escolha de técnicas estatísticas, devem ser determinada a variável dependente, também conhecida como variável resposta, e as variáveis independentes. A análise discriminante é um modelo em que se tem a variável depedente como categórica.

Geralmente, a variável resposta consiste em classificar dois grupos, por exemplo: pessoas do sexo masculino e feminino. Mas, também podemos distinguir três ou mais grupos, como pessoas baixas, médias e altas. O modelo em si é descrito matematicamente como a combinação linear das variáveis independentes (HAIR et al., 2009). Veja a formulação da equação 2.1.

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_n X_{nk} \tag{2.1}$$

Em que,

Z_{jk} = escore Z discriminante da função discriminante j para o objeto k;

a = intercepto;

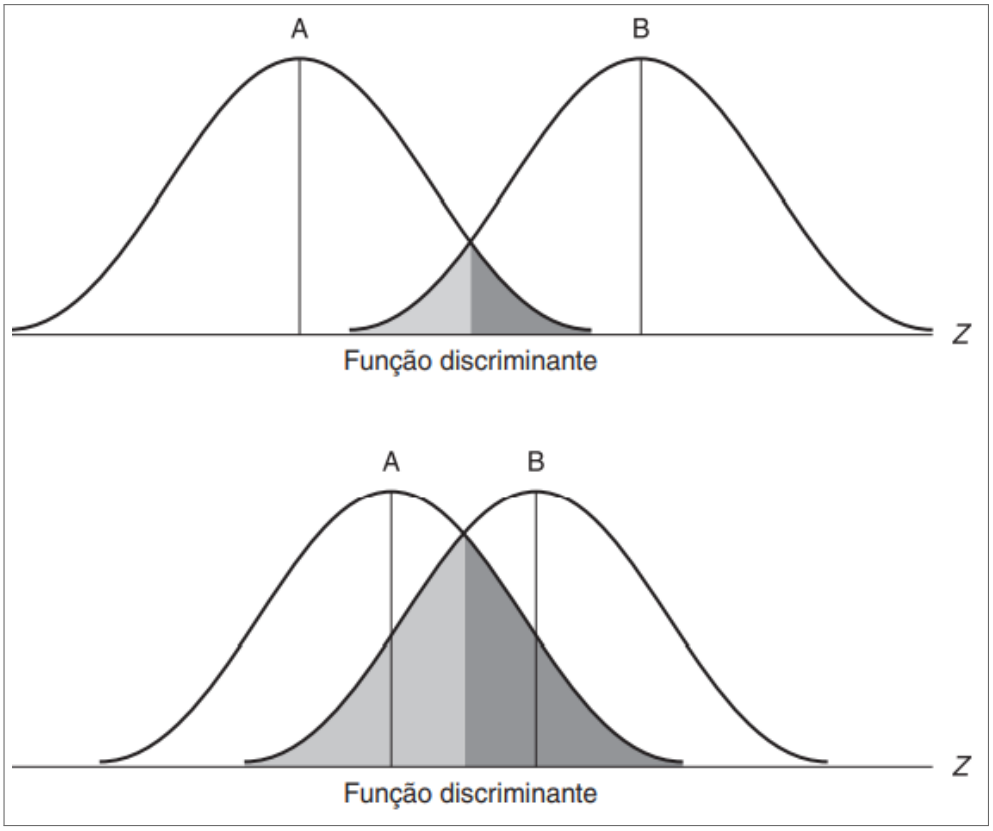
W_i = peso discriminante para a variável independente i;

X_{ik} = variável independente i para o objeto k.

Em (MINGOTI, 2007), vê-se que além de classificar grupos, a análise discriminante linear tem como objetivo reduzir a dimensionalidade dos dados para obter um melhor classificador com o menor número de dimensões.

O teste de validação para o modelo são os centroides do grupo, que é calculado por comparação de distância entre eles. Se temos dois grupos a serem comparados, será gerada uma distribuição discriminante para cada um. Se a sobreposição das distribuições forem pequenas, a função discriminante separa bem os grupos, se não a função não explica bem essa separação. A figura 2 mostra bem essa analogia, em que as partes sobreposta indicam que ali há indícios de má classificação.

Figura 2 – Representação univariada de escores Z discriminantes



Fonte: Hair et al. (2009)

Se a distribuição de probabilidade das características for conhecida é possível usar o princípio da máxima verossimilhança, (CASELLA; BERGER, 2002), para assim minimizar a probabilidade de uma classificação ruim. (FERREIRA, 1996) frisa que análise de discriminante não deve ser confundida com a análise de agrupamento, pois esta segunda deve ser aplicada a um número de grupos desconhecidos previamente, tendo como objetivo classificar os indiví-

duos. Além de que análise de agrupamento é uma técnica de aprendizado não supervisionado, enquanto a de discriminante é uma técnica e aprendizado supervisionado.

2.2.2 Regressão logística

Os modelos de regressão logística são adequados para estudar situações em que um conjunto de variáveis explicativas estão relacionadas a uma variável resposta dicotômica (SOUZA, 2014). Para entender melhor o modelo, use o experimento descrito abaixo:

Seja Ω um experimento aleatório consistindo em m repetições independentes de um evento dicotômico, ou seja, contém apenas dois resultados possíveis: 0 (falha) ou 1 (sucesso). Diz-se este um experimento de natureza binária. Com as condições sendo iguais para todas as réplicas do experimento, a probabilidade de cada resultado é $P(1) = \phi$ e $P(0) = 1 - \phi$ e constantes no decorrer das m repetições. Seja Y a variável aleatória de interesse, representando o número, nas m repetições, em que 1 ocorre, ou seja, o "sucesso", os valores que Y pode assumir são $0, 1, 2, \ldots, m$; cada um desses valores estão associados à probabilidade de ocorrência $P(Y = y)$, onde $y = 0, 1, 2, \ldots, m$ (BARRETO, 2011).

A definição do experimento Ω leva diretamente a uma variável aleatória Y com distribuição de probabilidade Binomial de parêmetros m e ϕ. Isso determina sua função de probabilidade, valor esperado e sua variância. (Veja as equações 2.2, 2.3 e 2.4).

$$f(Y;\phi) = P(Y_i = y) = \binom{m}{y}\phi^y(1-\phi)^{(m-y)} \tag{2.2}$$

$$E(Y_i) = m\phi \tag{2.3}$$

$$Var(Y_i) = m\phi(1-\phi) \tag{2.4}$$

Caso não haja repetições do experimento, então, ao invés de Binomial, estaremos trabalhando sob condições para distribuição de probabilidade Bernoulli, definida da mesma maneira que para uma variável Binomial, mas com $m = 1$ e a variável aleatória podendo assumir valores 1 e 0, aprensentando também uma natureza dicotômica.

Partindo para o âmbito dos Modelos Lineares Generalizados, no caso de variáveis resposta contínuas, não há restrições quanto ao domínio da resposta esperada que será estimada pelo modelo. Contudo, no caso do experimento Ω, no qual a variável aleatória Y_i pertence ao intervalo $[0, 1]$, o valor da resposta esperada será, na verdade, uma probabilidade, isto é, $P(Y_i = 1)$. Então, para uma correção desta limitação, será necessário utilizar uma transfor-

mação sobre a variável resposta Y, de modo que seu domínio se torne pertencente à reta real. A transformação mais comum para o caso de variáveis dicotômicas é a função exponencial que, por sua vez, transforma o valor esperado da variável Y na função de ligação logística. (BARRETO, 2011) (Veja a equação 2.5)

$$E(Y_i) = \phi_1 = \frac{exp(\eta_i)}{1 + exp(\eta_i)} \tag{2.5}$$

Nota-se, então, que o objetivo da transformação é plenamente atingido pela equação 2.5, uma vez que mesmo η_i podendo assumir qualquer valor real, ϕ_1, que é uma probabilidade, continuará restrito ao intervalo $[0, 1]$, como era desejado. Dessa forma o valor estimado pelo modelo linear será η_i e, para a formalização do modelo de regressão logística, ele pode ser relacionado a um modelo linear, contendo $p1$ variáveis explicativas e parâmetros, resultando em uma função conhecida como função resposta ***logit***.

$$\eta_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} \qquad i = 1, 2, \ldots, n. \tag{2.6}$$

Em que,

η_i é a resposta média *logit* para a $i - ésima$ observação;
X_i é o valor da variável preditora para a $i - ésima$ observação;
$\beta_k (k = 0, 1, \ldots, p - 1)$

Substituindo a equação 2.6 em 2.5, o modelo de regressão logística múltipla pode ser formulado da seguinte forma:

$$E(Y_i) = \phi_1 = \frac{exp(\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1})}{1 + exp(\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1})} \tag{2.7}$$

O modelo apresentado em 2.7 assume que os Y_i são variáveis aleatórias Bernoulli independentes com parâmetros $E(Y_i) = \phi$. Um modelo de regressão logística simples seria denotado por apenas uma variável explicativa e dois parâmetros a serem estimados: β_0 e β_1.

Com o modelo descrito em 2.6, o valor esperado da resposta em 2.5 significa agora que $P(Y_i) = 1$, dados os níveis para as preditoras em 2.6 e η_i se une à resposta logística esperada pela função de ligação *logit*, pois explicando em 2.5 em termos de η_i, temos:

$$\eta_i = ln\left(\frac{\phi_i}{1-\phi_i}\right) \tag{2.8}$$

O termo $\frac{\phi_i}{1-\phi_i}$ presente na equação 2.8 é conhecido como *odds* de sucesso. Se a probabilidade de sucesso equivaler a $0,5$, o *odds* vale 1; Se a probabilidade de sucesso for inferiror a $0,5$, o *odds* é inferior a 1; e se a probabilidade de sucesso for superior a $0,5$, o *odds* será maior do que 1.

A interpretação do coeficiente β em regressão logística não é tão simples e nem de direta compreensão, já que se trata de uma função de resposta não linear. Assim, no caso de regressão logística simples, um incremento unitário de X implicará em um efeito multiplicativo do *odds* estimado de sucesso, ou seja, $exp(b_1)$. De acordo com (BARRETO, 2011), para o caso de uma regressão logística múltipla, um incremento unitário X_1 ocasionaria esse mesmo efeito, mantidas constantes todas as demais variáveis do modelo $(X_2, X_3, \ldots, X_{P-1})$.

Um aspecto importante deve ser acrescentado à discussão em relação ao modelo de regressão logística é que, ao invés de formalizá-lo como descrito em 2.5 e 2.6, poderia-se tentar utilizar o modelo de regressão linear, que já é conhecido, para modelar a resposta binária, isto é:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.9}$$

Em que: $Y_i = 0$ ou $Y_i = 1$.

Contudo, se o modelo descrito em 2.9 fosse adotado, certamente ocasionaria dois problemas em relação aos pressupostos de regressão linear. O primeiro deles está relacionado com o fato de se modelar uma variável resposta binária Y_i, assumindo os valores zero ou um, o que implicaria termos de erro e, necessariamente, também assumiriam apenas dois valores diferentes: $(1 - \beta_0 + \beta_1 X_i)$ e $(-\beta_0 - \beta_1 X_i)$. Esse fato inviabiliza considerar os erros como normalmente distribuídos. Ademais, há a heterocedasticidade dos termos de erro que variam à medida que os níveis para X se alteram.

2.2.3 Árvore de decisão

Sendo um método de aprendizado supervisionado, as árvores de decisão são representações de possíveis resultados de forma simples usando métodos eficazes de construção de

modelos de classificação e regressão (GARCIA, 2003). O objetivo é criar um ajuste que preveja o valor da variável resposta, aprendendo regras de decisão simples inferidas do banco de dados (PEDREGOSA et al., 2011).

GAMA JOÃO 2009 fala que a árvore de decisão usa o que é chamado de estratégia de dividir e conquistar, ou seja, um problema complexo é dividido em problemas simples, em que o mesmo processo é usado repetidamente para cada subproblema. A capacidade de distinguir grupos da árvore de decisão vem com base em comparar possíveis ações com base em custos, probabilidades e benefícios dos Indivíduos da população. As árvores de decisão são representadas por:

- Nodos(nós): Representam um atributo;
- Ramos: Recebem os valores possíveis dos atributos;
- Folhas: Representam as classes de um conjunto de treinamento;

2.2.3.1 Árvore de decisão para classificação

Este método pode ser melhor entendido com um exemplo. Suponha que um jogador de tênis tenha que decidir se vai jogar em um determinado dia, com base no aspecto do céu, temperatura, umidade e o vento. Para montar o modelo é preciso investigar como estavam essas variáveis em dias anteriores, como feito na figura 3.

Figura 3 – Tabela para treino da árvore de decisão

Exemplos de Treino

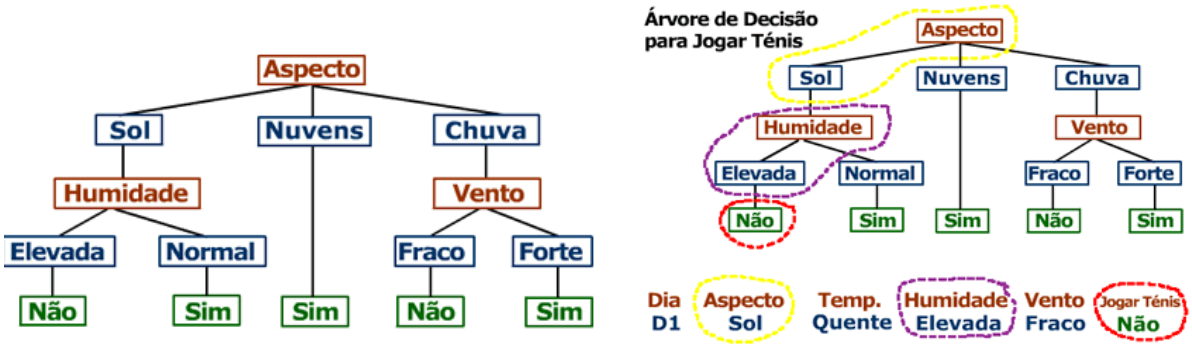
Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Fonte: Freitas (2022)

Na figura 4 à direita, é possível observar uma das formas de representar a árvore de decisão com as regras a partir da figura 3. Já a esquerda da imagem 4 tem a representação

da classificação do primeiro dia do treinamento do algoritmo, em que foi decidido não jogar.

Figura 4 – Árvore de decisão para jogar tênis



Fonte: Freitas (2022)

Se a variável resposta do modelo for de classificação existem duas medidas de impureza para validação. Dado um conjunto de entrada S as equações são dadas por:

$$Gini = \sum_{k} p_k(1 - p_k) \tag{2.10}$$

Enquanto a entropia de S é dada por:

$$Entropia(S) = -\sum_{k} p_k log(p_k) \tag{2.11}$$

Onde p_i é a proporção de dados pertencente a classe k

2.2.3.2 Árvore de decisão para Regressão

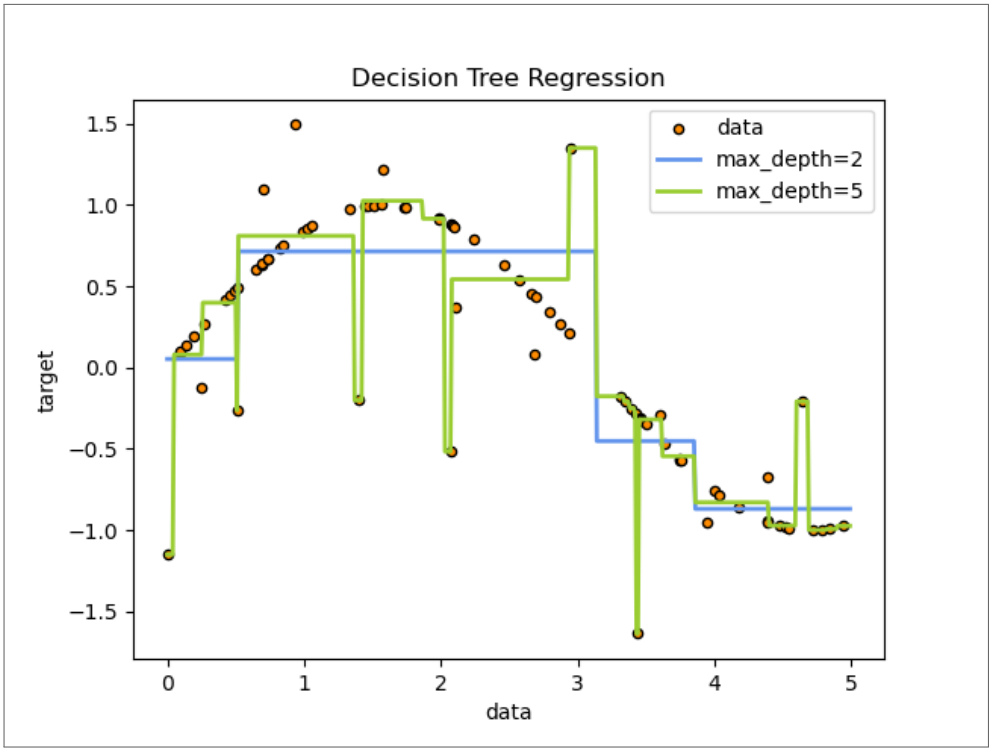
Para problemas de regressão, as árvores de decisão aprendem com os dados para aproximar uma curva senoidal, como mostrado na figura 5, com um conjunto de regras assim como em problemas de classificação.

Para os critérios de validação da Regressão usamos o Erro quadrático médio(MSE) e o desvio médio de poison, em que a matemática pode ser vista nas equações 2.12 e 2.13.

$$MSE(S) = \frac{1}{n} \sum_{y \in S} (y - \hat{y})^2 \tag{2.12}$$

$$DesvioMédiodePoisson(S) = \frac{1}{n} \sum_{y \in S} (ylog(\frac{y}{\hat{y}}) - y + \hat{y}) \tag{2.13}$$

Figura 5 – Estrutura de uma árvore de decisão para regressão



Fonte: Pedregosa et al. (2011)

Em q $\hat{y}$ é a média de aprendizado do modelo.

A árvore de decisão tem tendência de crescer de acordo com algumas aplicações, por essa razão é possível derivar regras. Elas fornecem um caminho da raiz até a folha da árvore (INGARGIOLA, 1996), essas regras acontecem por serem facilmente modeladas. Mais ainda, a árvore não assume algum modelo estatístico a priori, e a classificação de definir o espaço criado de acordo com as amostras de treinamento.

2.2.4 Floresta aleatória

Quando falamos de Métodos Ensemble, no qual encontramos uma das etapas mais interessantes do aprendizado de máquina, a ideia é que uma máquina procure lidar com a tarefa designada "a partir de diversas perspectivas", gerando respostas que, juntas, podem levar a uma generalização mais eficiente.

A aplicação dessa ideia no contexto das árvores de decisão traz alguns pontos interessantes que merecem uma discussão à parte. Eles deram origem ao conceito de *Floresta aleatória*.

Primeiro, porém, vamos começar com um ponto mais familiar: a construção de um

ensemble "convencional"de árvores. Uma maneira de construir um ensemble de árvores pode ser, por exemplo, através de *bagging*. Nesse caso, um conjunto de dados de N amostras é amostrado (com reposição), resultando em M conjuntos "novos", que por sua vez são usados para construir M árvores. As respostas dessas árvores são combinadas para gerar uma saída do ensemble. A extensão desse ensemble para noção de florestas aleatórias tipicamente envolve a construção de árvores a partir de subconjuntos de características, de tal forma a promover heterogeneidade.

Em (BREIMAN, 2001), coloca-se de modo simples a estratégia de construção de uma *Floresta aleatória* a definição de certo número de conjuntos (de mesma cardinalidade) e a geração de árvores a partir deles. Também é apresentado no estudo uma estratégica básica e sistemática:

- Se há N amostras de treinamento, gere N conjuntos de treinamento com amostragem com reposição (bootstrapping).
- Se há m atributos, escolha um número k de atributos ($k < m$) tal que, a cada nó, k atributos são escolhidos (de m) para particioná-lo. O valor k é mantido constante à medida que se constrói a floresta.
- Cada árvore cresce sem limitação externa e sem poda.

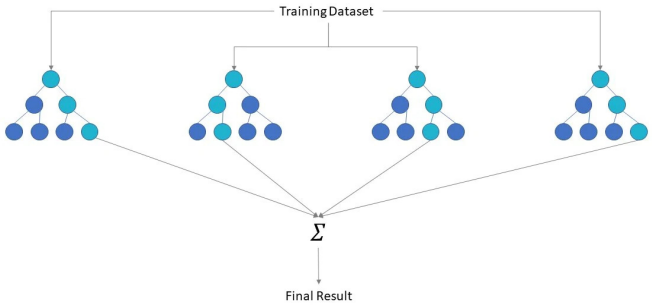
Há dois fatores cruciais nessa construção: a **capacidade** (*strength*) de cada árvore (que tem a ver com sua acurácia - árvores boas têm taxa de erro pequena) e a **correlação** entre as árvores. Um menor valor de k diminui a capacidade e a correlação. Um maior valor de k melhora a capacidade, mas aumenta também a correlação.

As ideias de construção de florestas aleatórias podem servir para medir similaridade ou dissimilaridade (SHI; HORVATH, 2006). No contexto de dados rotulados (caso supervisionado), uma possibilidade é apresentar padrões x_i e x_j a cada árvore do ensemble e, se o nó terminal (folha) obtido for o mesmo, contabiliza-se um valor "1", "0"caso contrário. Por fim, pode-se verificar o valor normalizado da soma obtida (percentual de árvores para as quais os dados tem o mesmo nó-folha). Esse índice funciona como uma medida de similaridade (ou leva a uma medida de dissimilaridade, se for o caso). No caso não-supervisionado, considera-se que se dispõe de um conjunto de dados não-rotulados D. Gera-se, então, um conjunto de dados sintéticos D' a partir de uma distribuição de referência. Desse modo, estabelecem-se duas

classes (D e D'), e passa a ser possível utilizar a medida para dados rotulados (focando no conjunto de dados "reais" D).

Em suma, o algoritmo de floresta aleatória consiste em um conjunto de árvores de decisão, e cada árvore no conjunto consiste em amostras de dados com substituições extraídas do conjunto de treinamento, chamadas de amostras bootstrap. Do exemplo de treinamento ilustrado na Figura 6, um terço é reservado para dados de teste, chamados de *out-of-bag* (oob). Outra instância de aleatoriedade é então injetada por meio do agrupamento de recursos, adicionando mais diversidade ao conjunto de dados e reduzindo a correlação entre as árvores de decisão. Dependendo do tipo de problema, a determinação das previsões irá variar. Para tarefas de regressão, as árvores de decisão individuais serão calculadas em média, enquanto para tarefas de classificação, o voto da maioria, ou seja, a variável categórica mais frequente, produzirá a classe prevista. (EDUCATION, 2020)

Figura 6 – Diagrama do Classificador de Floresta Aleatória



Fonte: Figura 2 de Education (2020)

2.2.5 Adaptive Boosting

O algoritmo foi publicado pela primeira vez por Freund e Schapire em 1995 (FREUND; SCHAPIRE et al., 1996), o qual resolveu alguns problemas e dificuldades encontradas no *Boosting* anteriormente e, atualmente, apresenta crescente número de pesquisas sobre a metodologia e suas aplicações por estudiosos de diferentes regiões do mundo, como em (BARCZAK; JOHNSON; MESSOM, 2008). Ainda segundo o inventor do algoritmo, o AdaBoost possui algumas propriedades específicas que o tornam mais prático de usar e implementar em comparação a outros algoritmos (FREUND; SCHAPIRE et al., 1996). Entre as principais diferenças deste algoritmo podemos destacar:

- **Diferenças de parâmetros utilizados:** Na análise de grandes espaços dimensionais , os valores das margens entre classes podem se apresentar muitas vezes bastante diferentes e mais precisos, se comparado a outros métodos como o SVM (FREUND; SCHAPIRE; ABE, 1999);

- **Requer baixo valor computacional**: AdaBoost corresponde a um programa linear e, logo, não necessita de componentes eletrônicos pesados para funcionar e fazer as análises (FREUND; SCHAPIRE; ABE, 1999);

O AdaBoost é utilizado junto a algoritmos base para cruzar duas ou mais informações sobre dados analisados e, ao comparar com a base de dados de treinamento, reconhecer padrões e fazer as classificações. Desse modo, suponha que duas informações diferentes xey foram coletadas sobre diversos objetos de uma amostra. Ao serem analisadas por um algoritmo já treinado previamente, o resultado final adquirido será a classificação automática de acordo com os padrões encontrados e comparados com os dados de treinamento. Este resultado pode servir de entrada para uma tomada de decisão da própria máquina ou do próprio controlador dela.

Após a primeira proposta do AdaBoost ser apresentada a cientistas e pesquisadores, muitas variações e extensões deste algoritmo foram propostas e desenvolvidas como alternativas para fornecer melhores resultados para determinados problemas. Algumas das razões para o surgimento dessas novas propostas são as aplicações que estão sendo analisadas (por exemplo, processamento de imagens) e o tipo de problema que está sendo tratado (por exemplo, binário ou multifásico).

O funcionamento do AdaBoost se dá em algumas etapas. A primeira etapa é o **treinamento**. Nesta fase, a entrada para o AdaBoost é um conjunto de exemplos na forma :

$$S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}, \;\; i = 1, 2, \ldots, m.$$

Em que,

a. Cada x_i representa um vetor de atributos do sistema, ou seja, um conjunto de dados referente aos parâmetros (ou propriedades) analisados e que representam um dado instante do sistema.

b. Dentre todos os possíveis grupos de classificação predefinidos para o sistema analisado, cada y_i representa o grupo de classificação associado ao x_i.

c. O número total de exemplos da amostra de treinamento é igual a m.

O algoritmo base escolhido para trabalhar com o AdaBoost é convocado repetidamente em uma série de iterações e, em cada uma, o AdaBoost fornece ao algoritmo base uma distribuição de peso que se refere a cada dado de treinamento. Os classificadores começam com pesos paralelos distribuídos uniformemente.

$$D_0 = \frac{1}{m} \tag{2.14}$$

Em cada ciclo de aprendizado, o algoritmo base gera uma hipótese h_t baseada nos pesos atuais para priorizar a classificação correta dos dados com maior peso relativo. Portanto, pode-se dizer que o objetivo do algoritmo base é gerar uma hipótese para reduzir o erro de treinamento ϵ_t, no qual:

$$\epsilon_t = \sum_{i:h_t(x_t)\neq y_t} D_t(i) \tag{2.15}$$

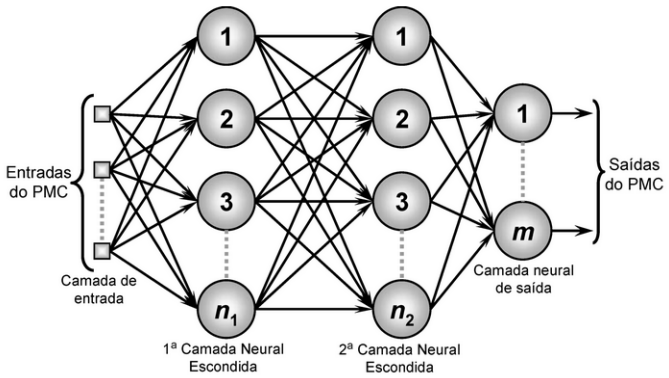
Em seguida, esses pesos são reavaliados para aumentar aqueles associados a dados classificados incorretamente e um novo ciclo é iniciado. Finalmente, quando todas as T rodadas pré-determinadas são feitas, o AdaBoost combina todas as ideias intermediárias h_t para produzir uma única hipótese final $H(X)$. De acordo com (FREUND; SCHAPIRE; ABE, 1999) o método para se obter a hipótese final $H(x)$ fornecida pelo AdaBoost é a combinação ponderada das diversas saídas das iterações. Assim, uma votação ponderada das T hipóteses fracas é realizada, no qual γ_t é a importância associada à hipótese h_t:

$$H(x) = \textit{sign}(\sum_{t=1}^{T} \gamma_y h_t(x))$$

2.2.6 Perceptron Multicamadas

Perceptron Multicamadas (MLP — Multi Layer Perceptron) é uma rede neural com uma ou mais camadas ocultas com um número infinito de neurônios. A camada oculta é assim chamada porque não é possível prever a saída desejada das camadas intermediárias. Para treinar a rede MLP, o algoritmo mais utilizado é a retropropagação (Backpropagation). O fluxo da MLP está representado na figura 7

Figura 7 – Diagrama do Perceptron Multicamadas



Fonte: Moreira (2020)

Diferentemente de algumas outras redes neurais, como Perceptron e Adaline, no qual a saída é constituída por apenas um neurônio, a MLP pode relacionar o conhecimento a mais de um neurônio de saída. De acordo com (MOREIRA, 2020) o algoritmo de aprendizado da MLP, além de ser chamado de *backpropagation*, se decompõe em 4 passos:

- **Inicialização:** Neste passo, é necessário atribuir valores para os pesos e limites; Escolher o chute inicial (influencia o comportamento da rede); Caso não haja conhecimento prévio, os pesos e limites devem ter valores iniciais aleatórios, pequenos e uniformemente distribuídos.

- **Ativação:** Os valores dos neurônios da camada oculta devem ser calculados nessa etapa, assim como os neurônios da camada de saída.

- **Treinar os pesos:** Calcular tanto os erros da camada de saída quanto os da camada oculta da rede além de calcular, também, a correção dos pesos. Após isto, atualizar os pesos dos neurônios das camadas de saída e oculta.

- **Iteração:** Por fim, repetir o processo a partir do passo 2 até que o critério de erro seja satisfeito.

Alguns aspectos práticos devem ser levados em consideração na utilização de redes neurais MLP, tais como:

- Taxa de aprendizagem: O ideal é começar com uma taxa grande e reduzir durante as iterações, geralmente valores entre 0.05 e 0.75 fornecem bons resultados.

- *Momentum*: É uma estratégia utilizada para evitar mínimos locais.

- *Online vs batch*: A diferença está no momento em que os pesos são atualizados. Na versão *On-line*, os pesos atualizados a cada padrão que a rede apresenta, na versão *batch*, todos os pesos são somados durante uma iteração (todos os padrões) e só assim os pesos são atualizados.

- Normalização: A normalização é interessante quando existem características em diversas unidades dentro do vetor de características. Nestes casos, valores muito altos podem saturar a função de ativação. Uma forma bem simples de normalizar os dados consiste em somar todas as características e dividir pela soma, outra bastante usada é a normalização Z-score.

- Inicialização dos pesos: Os pesos devem ser inicializados aleatoriamente mas de modo que fiquem na região linear da sigmoid.

- Generalização: Evitar *overfitting*. A maneira mais comum de garantir uma boa generalização é reservar uma parte da base de treinamento para **validar** a generalização.

- Tamanho da rede: Geralmente uma camada escondida é suficiente, em poucos casos se é necessário adicionar mais de uma camada oculta.

3 MATERIAIS E MÉTODOS

Este capitulo apresenta a aquisição dos materiais e a metodologia utilizada nesta pesquisa, a pesquisa aplicada, com o objetivo de empregar a aprendizagem de máquina no cenário da prematuridade.

3.1 OBTENÇÃO DOS DADOS

Com relação às bases de dados utilizadas neste trabalho, apesar de todas terem origem no Ministério da Saúde, as obtenções foram em repositórios distintos. Para os anos 2018 a 2020, foi feita na Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) mantida pela Fiocruz, e que receberam a adição de campos relacionados, como por exemplo os municípios, UF e CID 10, facilitam assim a exploração dos dados (FIOCRUZ, 2018). Os dados contêm os registros das declarações de nascidos vivos de 1996 a 2020 em território brasileiro. Para o ano de 2021 os dados foram obtidos diretamente no Sinasc (SVC, 2022). Cabe uma observação com relação a este ano, que os dados foram coletados na sua versão preliminar, logo estes podem sofrer alterações.

Apesar da abrangência territorial e temporal do Sinasc, para este projeto foram considerados somente os dados das 27 capitais brasileiras no intervalo de 2018 a 2021. Uma observação que deve ser realizada é com relação ao ano de 2020, que os dados usados foram a partir do mês de março, devido ao fato da COVID-19 ter iniciado efetivamente neste mês no Brasil (OPAS, 2020). Um outro ponto que se destaca nesta base de dados é com relação a sua cobertura e completude (SZWARCWALD et al., 2019).

3.2 TRATAMENTO DOS DADOS

Dada a obtenção dos dados, iniciou-se uma fase de análise e tratamento, sendo realizada a seleção com relação ao tipo de gravidez, dentre os três tipos (I: Única; II: Dupla; III: Tripla e mais) apresentados, por critério de inclusão foi definido o tipo I de gravidez, que corresponde a gravidez de feto único. Um processamento prévio foi realizado para o tratamento inicial dos dados com intuito de ajustar os registros com que os registros com incompletude (brancos + sem preenchimento) foram excluídos.

O direcionamento com relação às variáveis a serem utilizadas foram focadas em fontes de dados secundárias disponibilizadas no SINASC e parcialmente baseadas nos estudos de (LEE et al., 2020) e (KOIVU; SAIRANEN, 2020) que também fazem uso de dados demográficos e socio-econômicos assim como dados clínicos. Além disso, todo um processo de análise exploratória foi realizado para complementar as etapas desta seleção.

Quanto à idade gestacional, está sendo considerada de 22 à 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. Algumas variáveis de acordo com a distribuição como: idade da mãe, quantidade de gestações, filhos (mortos / vivos) e parto (cesárea / normal) sofreram um truncamento com relação a distribuição dos dados, conforme a Tabela 1.

Tabela 1 – Truncamento das variáveis

Variáveis	**Regra de truncamento**	**Valores**
Idade	Registro menor que 11 e Registro maior que 50	[11 - 50]
QtdGestAnteriores	Registro maior que 5	5
QtdFilhoMorto	Registro maior que 4	4
QtdFilhoVivo	Registro maior que 5	5
QtdPartoCesaria	Registro maior que 2	2
QtdPartoNormal	Registro maior que 6	6

Fonte: Elaborada pelo autor (2022)

Com relação a variável alvo, conforme a documentação das variáveis (PCDAS, 2018), parto_prematuro tem 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);

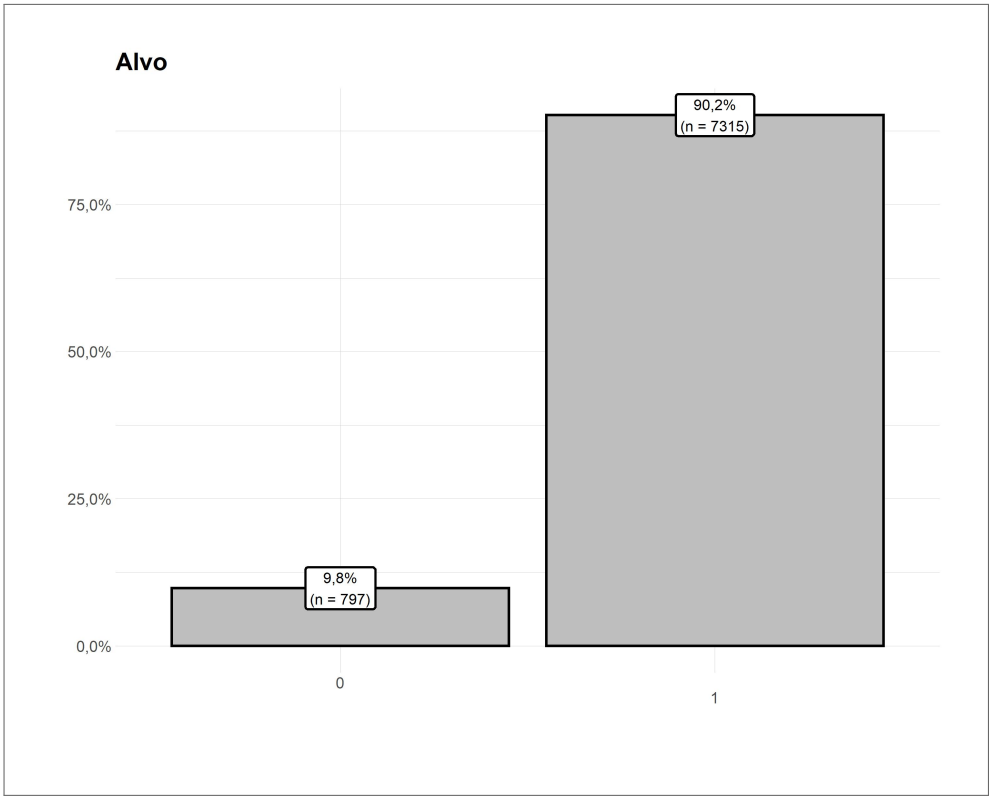
3: Pré-termo indicado pela idade gestacional e o peso ao nascer.

No entanto, foi realizada uma transformação, assumindo assim, 1 na variável alvo, os registros de parto_prematuro preenchidos com 3 e 0 caso contrário. Portanto, na nova definição temos:

0: Não Prematuro;

1: Prematuro.

Figura 8 – Alvo 2018



Fonte: Elaborada pelo autor (2022)

Também foram criadas novas variáveis como: PropQtdFilhoVivo, PropQtdFilhoMorto, SomaQtdFilhoVivoMorto e RazaoQtdFilhoVivoMorto. A definição de cada uma é dada respectivamente pelas equações 3.1, 3.2, 3.3 e 3.4 .

$$PropQtdFilhoVivo = \frac{qtdFilhoVivo}{qtdFilhoVivo + qtdFilhoMorto} \tag{3.1}$$

$$PropQtdFilhoMorto = \frac{qtdFilhoMorto}{qtdFilhoVivo + qtdFilhoMorto} \tag{3.2}$$

$$SomaQtdFilhoVivoMorto = qtdFilhoVivo + qtdFilhoMorto \tag{3.3}$$

$$RazaoQtdFilhoVivoMorto = \frac{qtdFilhoVivo}{qtdFilhoMorto} \tag{3.4}$$

Por fim, temos as bases de dados resultantes de cada ano. 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

Variáveis	Descrição	Min.	1º quart.	Mediana	Média	3º quart.	Máx.	Desvio padrão	Coef. de var.
QtdGestAnteriores	Quantidade de gestações anteriores	0	0	1	1,1730	0	5	1,377	1,1741
QtdFilhoMorto	Quantidade de filhos mortos	0	0	0	0,3627	0	4	0,711	1,9597
QtdFilhoVivo	Quantidade de filhos vivos	0	0	0	0,8432	0	5	1,168	1,3849
QtdPartoCesaria	Quantidade de parto cesáreo	0	0	0	0,2877	0	2	0,565	1,9626
QtdPartoNormal	Quantidade de parto normal	0	0	0	0,5931	0	6	1,111	1,8737
Idade	Idade da mãe	11	22	28	27,7655	22	50	7,386	0,2660
SomaQtdFilhoVivoMorto	Soma de filhos vivos e mortos	0	0	1	1,2059	0	9	1,457	1,2086
PropQtdFilhoVivo	Proporção de filhos vivos	-1	-1	0	-0,0158	-1	1	0,889	-56,3454
PropQtdFilhoMorto	Proporção de filhos mortos	-1	-1	0	-0,2457	-1	1	0,707	-2,8785
MesInicioPreNatal	Mês de início das consultas de pré-natal	-1	1	2	2,3025	1	8	1,599	0,6942

Fonte: Elaborada pelo autor (2022)

Tabela 3 – Descrição das variáveis categóricas

Variáveis	Valores	Quantidade	%
RazaoQtdFilhoVivoMorto	01.NUNCA TEVE FILHOS	3411	42,05
	02.TODOS OS FILHOS NASCERAM MORTOS	858	10,58
	03.TEVE MAIS FILHOS MORTOS QUE VIVOS	211	2,60
	04.TEVE IGUAL QTD DE FILHOS MORTOS E VIVOS	527	6,50
	05.TEVE MAIS FILHOS VIVOS QUE MORTOS	3105	38,28
Raca	0.NENHUMA	383	4,72
	01.BRANCA	2334	28,77
	02.PRETA	713	8,79
	03.AMARELA	46	0,56
	04.PARDA	4615	56,89
	05.INDIGENA	21	0,25
EstadoCivil	0.NENHUMA	62	0,76
	01.SOLTEIRA/VIUVA	4310	53,13
	02.CASADA	2275	28,04
	03.SEPARADA	111	1,37
	04.UNIAO CONSENSUAL	1354	16,69
Escolaridade	0.NENHUMA	89	1,10
	01.1 A 3 ANOS	93	1,15
	02.4 A 7 ANOS	1062	13,09
	03.8 A 11 ANOS	4785	58,99
	04.12 ANOS OU MAIS	2083	25,68
RegiaoDoPais	01.NORTE	1347	16,61
	02.NORDESTE	1909	23,53
	03.SUDESTE	3267	40,27
	04.SUL	526	6,48
	05.CENTROOESTE	1063	13,10
QtdConsultaPrenatal	0.NENHUMA	398	4,91
	01.1 A 3	2085	25,70
	02.4 A 6	3680	45,36
	03.7 OU MAIS	1949	24,03
Alvo	0	797	9,82
	1	7315	90,18

Fonte: Elaborada pelo autor (2022)

4 EXPERIMENTOS

4.1 CONFIGURAÇÃO DOS EXPERIMENTOS

No que diz respeito aos ambientes de desenvolvimento e linguagens, para a realização das análises exploratórias, dos ajustes nas bases de dados e as execuções dos modelos foram utilizadas: a linguagem R (4.1.2) e Python (3.10.4). O sistema operacional Linux, dentre as bibliotecas podemos destacar: pandas, tidyverse, scikit-learn, caret. Para a análise e manipulação de dados foi utilizado o pandas e tidyverse, já o scikit-learn e caret foram utilizados na configuração e ajustes dos modelos. 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.

Com relação aos algoritmos de aprendizado de máquina, foram utilizados: Decision Tree (DT), Adaptive Boosting (AdaBoost),Logistic Regression (LR),Random Forest (RF),Multilayer Perceptron (MLP), Linear Discriminant Analysis (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).

Passadas as etapas iniciais de tratamento dos dados, descritas nas subseções 3.0.1 e 3.0.2, para a aplicação dos algoritmos nas bases de dados, um novo processamento foi feito. Para que se iniciasse os experimentos 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 que é dada pela função 4.1.

Tabela 4 – Parâmetros usados

Parâmetros	**Descrição**	**Valores**
mtry	Número de variáveis amostradas aleatoriamente como candidatas em cada divisão	[1,60]
ntree	Número de árvores a crescer	[5,150,5]
size	Número de unidades na(s) camada(s) oculta(s)	[5,120,5]
learnFuncParams	Parâmetros para a função de aprendizagem	[0.01,0.1,0.01]
maxit	Máximo de iterações para aprender os parâmetros para a função de aprendizagem	[100,300,50]
minprob	Proporção de observações necessárias para estabelecer um nó terminal.	[0.01,0.05,0.01]
maxvar	Número máximo de variáveis que a árvore pode dividir.	[3,36]
maxdepth	Profundidade máxima da árvore.	[3,5,8,12,Inf]

Fonte: Elaborada pelo autor (2022)

$$Z = \frac{x - min(x)}{max(x) - min(x)} \tag{4.1}$$

Outra transformação necessária foi com relação às variáveis categóricas da matriz de dados, 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.

Em seguida, a divisões das bases foram 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* [1] na base de validação que tem o objetivo de encontrar uma melhor combinação dentre um conjunto de hiperparâmetros.

4.2 MÉTRICAS DE AVALIAÇÃO

Se o objetivo da modelagem é tomar decisão, faz-se necessário o uso de métricas para avaliação dos modelos. Essas estatísticas possibilitam a obtenção de um valor numérico que quantifica o quanto o modelo utilizado erra e acerta nas predições.

A princípio, é necessário entender o funcionamento de uma matriz de erro (também conhecida como matriz de confusão). A matriz de confusão é uma tabela de frequência de categorias que apresentam os erros (quando o valor predito é diferente do valor real) e acertos (quando o valor predito é igual ao valor real) de um modelo.

Tabela 5 – Matriz de Confusão

		Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: adaptada de (SALLABA, 2011)

Matriz de confusão: Apresentada na Tabela 5 sua estrutura para duas classes, um caso particular de uma matriz de confusão de n classes, como apresentada no exemplo teórico

[1] *Grid Search*: é um algoritmo de busca gulosa que tem o objetivo de encontrar em um subconjunto de hiperparâmetros especificado manualmente, a melhor combinação destes, é o melhor desempenho na base de validação (FACELI et al., 2011).

de (CONGALTON; GREEN, 2019). As linhas representam as classificações reais e as colunas representam as classificações estimadas. Para este caso, têm-se os seguintes valores:

- Verdadeiro Positivo (VP): quando há classificação correta da classe Positivo; acerto em que o valor estimado pertence à classe Positivo quando realmente pertence à classe Positivo.
- Falso Positivo (FP): erro em que o modelo estimou a classe Positivo quando, na verdade, o valor real era da classe Negativo.
- Verdadeiro Negativo (VN): quando há classificação correta da classe Negativo; acerto em que o valor estimado pertence à classe Negativo quando realmente pertence à classe Negativo.
- Falso Negativo (FN): erro em que o modelo estimou a classe Negativo quando, na verdade, o valor real era da classe Positivo;

Acurácia: Verifica o quão próximo os seus resultados obtidos no teste estão do ponto de referência, ou seja, quanto o modelo classificou corretamente. A métrica pode ser obtida pela equação 4.2.

$$AC = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.2}$$

Precisão: (MONICO et al., 2009) afirmam que a diferença entre acurácia e precisão vem da presença de erros sistemáticos, que chamam de tendência ou viés. Na matriz de confusão apresentada na Figura 5, a precisão (chamemos de P), será a proporção de classificações corretas das predições feitas pelo modelo. E o resultado dessa métrica pode ser obtido pela equação 4.3 apresentada abaixo.

$$P = \frac{VP}{VP + FP} \tag{4.3}$$

Recall ou Sensibilidade: Diferentemente da precisão, recall pode ser entendido como uma a porcentagem de classificações corretas relacionadas às quantidades reais, ou seja, o quanto dos valores reais o modelo acertou, também chamado de *True Positive Rate* (TPR). Assim, quando a precisão diminui, recall aumenta. E o cálculo dessa métrica pode ser obtido pela fórmula da equação 4.5.

$$R = \frac{VP}{VP + FN} \tag{4.4}$$

Especificidade: a proporção de casos negativos que foram identificados corretamente. Podendo ser obtida pela fórmula da equação 4.5.

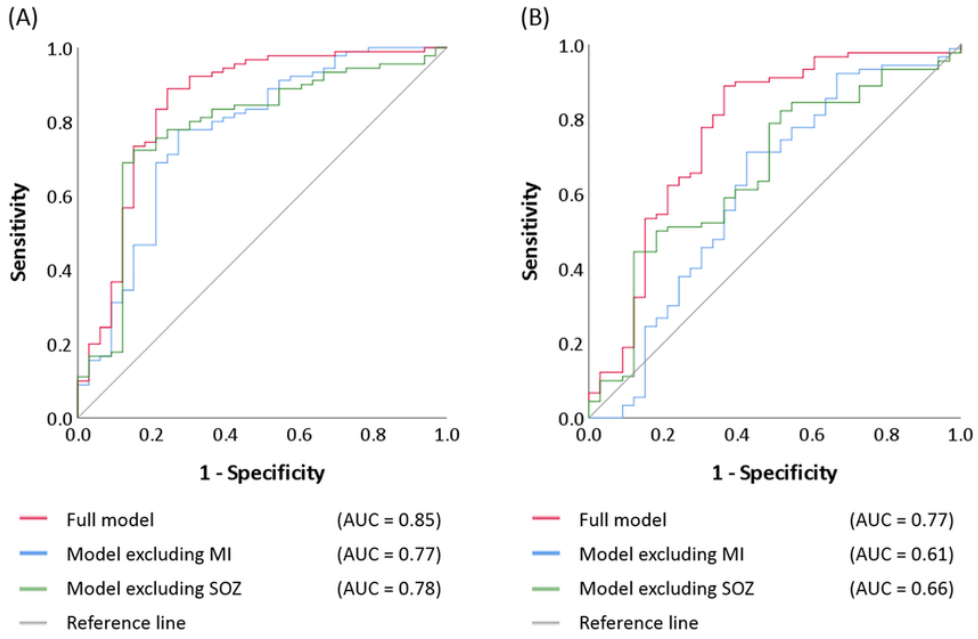
$$E = \frac{VN}{VN + FP} \tag{4.5}$$

F1-score: É a média harmônica entre precisão e recall, que possibilita observar duas métricas em 1 única métrica. Quando o resultado do F1-Score é baixo, significa que, possivelmente, a precisão ou o recall está baixo. Esse resultado pode ser encontrado pela equação 4.6.

$$F1 - Score = \frac{2 * P * RC}{P + RC} \tag{4.6}$$

ROC-AUC: *Receiver Operating Characteristic* é uma curva de probabilidade que mostra a eficiência do modelo criado para separar as classes. a curva ROC traça 2 parâmetros, TPR e FPR. Dessa forma, para simplificar a análise a AUC (*Area Under the Curve*) atribui um valor que resume a curva ROC. O valor de AUC varia de 0 até 1, quanto maior esse número for, melhor o modelo irá distinguir as classes. Além disso, deve ser adicionado um outro parâmetro limiar, que geralmente é escolhido o valor 0.5. Veja o exemplo do desempenho de um modelo na Figura 9.

Figura 9 – Curvas ROC e valores de AUC do desempenho do modelo para classificar o sucesso cirúrgico.



Fonte: (MOTOI et al., 2019)

Dado que as métricas podem ser semelhantes, o teste Kruskal-Wallis será aplicado inicialmente 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, 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).

Em caso da 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[2] continha 30 observações e foi computada 100 vezes para cada modelo e ano. Essa abordagem foi necessária pois alguns modelos utilizados não tem 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.

Devido ao desbalanceamento das bases, também será realizada a execução do experimento usando técnicas de balanceamento de dados. 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).

[2] **Amostra bootstrap**: é um conjunto de valores extraídos ao acaso com reposição da amostra original.

5 RESULTADOS

Neste capítulo será representado os resultados obtidos pelos modelos de aprendizagem de máquina aplicados e suas respectivas descrições. Na Tabela 6 são listados os melhores parâmetros aplicados de acordo com cada modelo que foi aplicada a técnica de *Grid Search*.

Tabela 6 – Melhores parâmetros (base desbalanceada)

Parâmetros (Modelo)	**Valores**
mtry (RF)	[2]
ntree (RF)	[150]
size (MLP)	[35]
maxit (MLP)	[250]
learnFuncParams (MLP)	[0.04]
maxdepth (DT)	[8]
maxvar (DT)	[4]
minprob (DT)	[0.01]

Fonte: Elaborada pelo autor (2022)

A Figura 10 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, onde 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 10 – Matriz de confusão - (A)base 2018 (B)base 2019 - (base desbalanceada)

(A)

			Real	
			0	1
RL	Predito	0	592	2221
		1	205	5094
LDA	Predito	0	612	2429
		1	185	4886
MLP	Predito	0	569	1993
		1	228	5322

			Real	
			0	1
Adaboost	Predito	0	612	2460
		1	185	4855
RF	Predito	0	558	788
		1	239	6527
DT	Predito	0	496	1745
		1	301	5570

(B)

			Real	
			0	1
RL	Predito	0	577	2070
		1	231	5056
LDA	Predito	0	602	2284
		1	206	4842
MLP	Predito	0	538	1841
		1	270	5285

			Real	
			0	1
Adaboost	Predito	0	570	2411
		1	238	4715
RF	Predito	0	371	1343
		1	437	5783
DT	Predito	0	445	1651
		1	363	5475

Fonte: Elaborada pelo autor (2022)

As matrizes de confusão permitem obter as métricas de performances, as quais são apresentadas nas tabelas 7 e 8 para, respectivamente, as bases de treinamento (2018) e validação

(2019) de forma desbalanceadas.

Tabela 7 – Métricas de performance para a base de dados de 2018 (base desbalanceada)

Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Regressão logística	0,7009	0,9613	0,6964	0,8077	0,7713
Análise Discriminante Linear	0,6778	0,9635	0,6679	0,7890	0,7714
Multilayer perceptron	0,7262	0,9589	0,7275	0,8274	0,7774
AdaBoost	0,6739	0,9633	0,6637	0,7859	0,7497
Floresta aleatória	**0,8734**	**0,9647**	**0,8923**	**0,9271**	**0,8756**
Árvore de decisão	0,7478	0,9487	0,7614	0,8448	0,7460

Fonte: Elaborada pelo autor (2022)

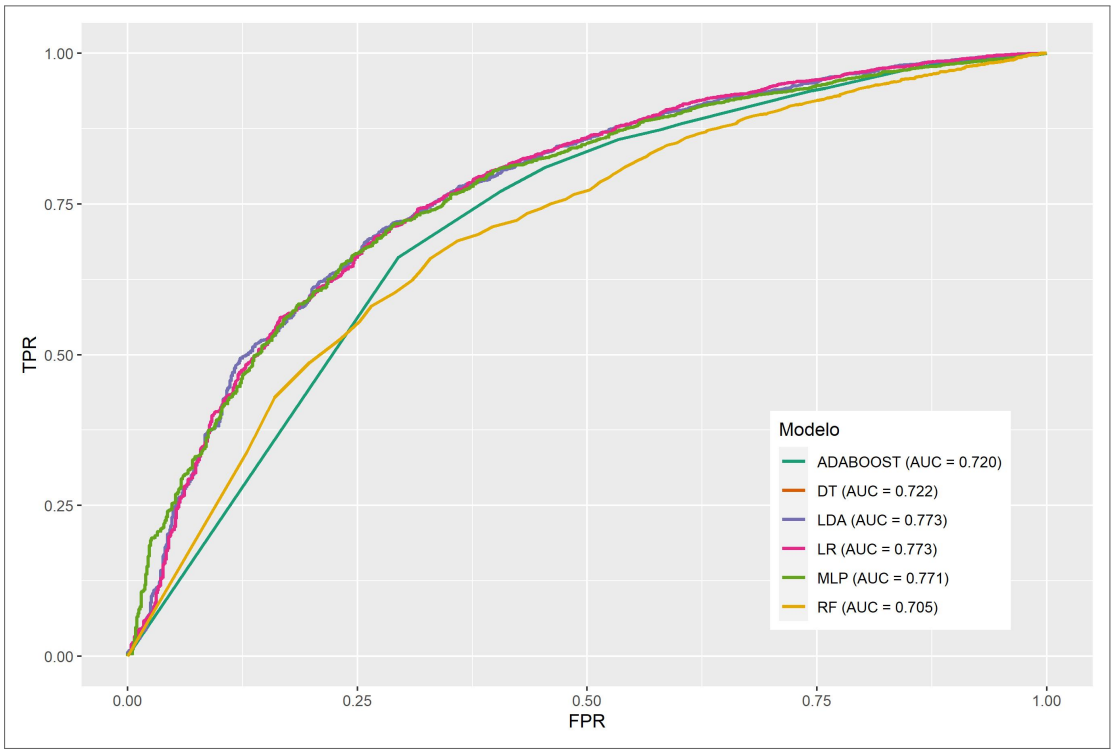
Tabela 8 – Métricas de performance para a base de dados de 2019 (base desbalanceada)

Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Regressão logística	0,7100	0,9563	0,7095	0,8146	0,7729
Análise Discriminante Linear	0,6862	0,9592	0,6795	0,7955	0,7728
Multilayer perceptron	0,7339	0,9514	0,7417	0,8335	0,7706
AdaBoost	0,6661	0,9519	0,6617	0,7807	0,7201
Floresta aleatória	**0,7756**	**0,9297**	**0,8115**	**0,8666**	**0,7052**
Árvore de decisão	0,7462	0,9378	0,7683	0,8446	0,7216

Fonte: Elaborada pelo autor (2022)

Para a base de treinamento desbalanceada (Tabela 7), a métrica acurácia teve uma variação de 0,6739 (Adaboost) à 0,8734 (Floresta aleatória), enquanto que para a base de validação desbalanceada (tabela 8), essa variação foi de 0,666 (Adaboost) à 0,7756 (Floresta aleatória). Tal comportamento é similar para as métricas Recall, F1 score e AUC, onde o modelo Floresta aleatória apresenta bastante divergência de valores entre as duas bases. A métrica precisão obteve a menor variação dentre as demais, de 0,9487 (Árvore de decisão) à 0,9647 (Floresta aleatória) na base de 2018 desbalanceada e 0,9297 (Floresta aleatória) à 0,9592 (Análise discriminante linear).

Figura 11 – AUC-2019 (base desbalanceada)

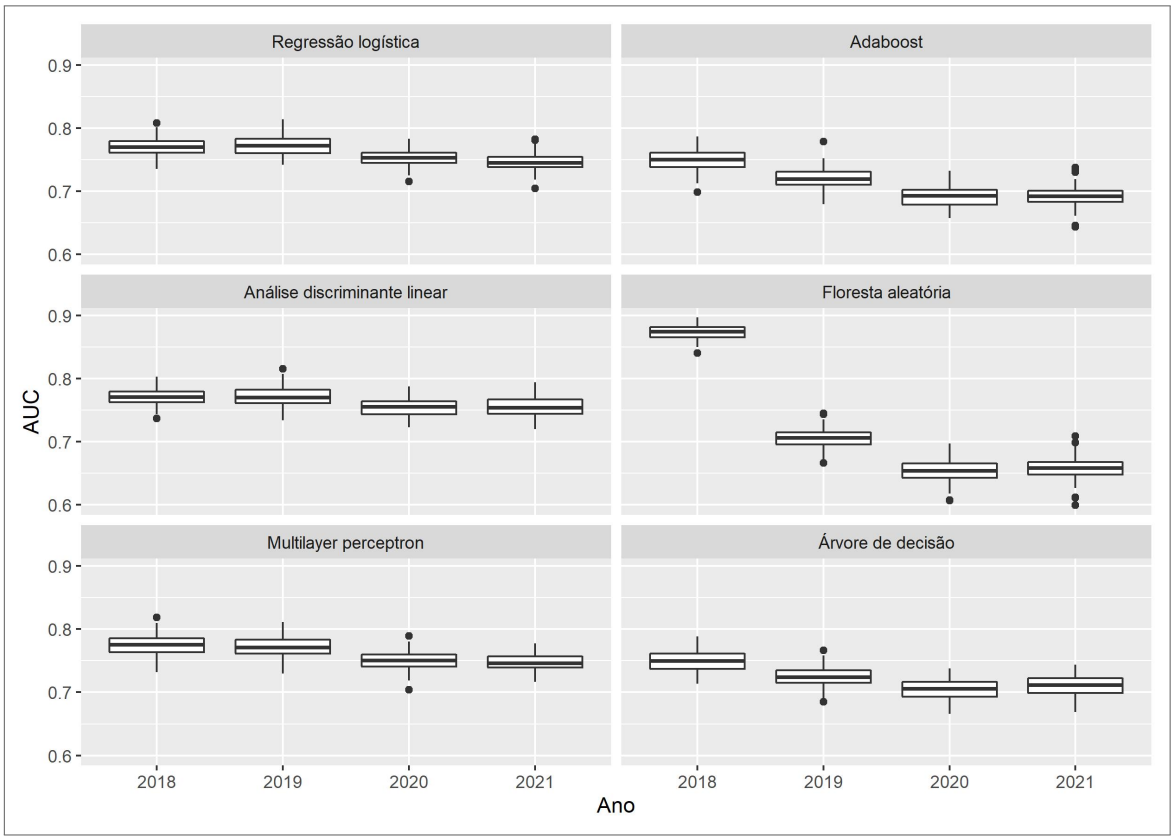


Fonte: Elaborada pelo autor (2022)

Na Figura 11, são exibidas a área sob a curva (AUC), que relaciona a taxa de falsos positivos (especificidade) e a taxa de verdadeiros positivos (sensibilidade). Nota-se dois grupos, um com valores acima de 75% como é o caso do LDA, LR e MLP e RF e um outro variando entre 70% e 74% para ADABOOST e DT.

A partir dos gráficos boxplot que contém os resultados da amostra bootstrap da métrica AUC (Figura 12) 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 em sua mediana. Ainda, o modelo RF apresenta uma queda notória também para a base de dados de 2019.

Figura 12 – Boxplot AUC (base desbalanceada)



Fonte: Elaborada pelo autor (2022)

Com o teste post-hoc de Nemenyi(Apêndice A) constatou-se para os modelos treinados com a base de dados desbalanceada que: Para os modelos de regressão logística, análise discriminante linear e multilayer perceptron, diferenças entre os anos de 2018-2020 (valor-p $<$ 0,001), 2018-2021 (valor-p $<$ 0,001), 2019-2020 (valor-p $<$ 0,001) e 2019-2021 (valor-p $<$ 0,001). A partir de 2020, a queda em relação a 2018/2019 em média foi de 2,8% para a regressão logística, 2,2% Análise discriminante linear e 3,3% para o multilayer perceptron. Já para os modelos Adaboost, Floresta aleatória e Árvore de decisão, as diferenças aparecem da seguinte maneira: 2018-2019 (valor-p $<$ 0,001), 2018-2020 (valor-p $<$ 0,001), 2018-2021 (valor-p $<$ 0,001), 2019-2020 (valor-p $<$ 0,001) e 2019-2021 (valor-p $<$ 0,001). Para o ano de 2019 em relação a 2018, os modelos Adaboost e Árvore de decisão possuem uma diminuição de respectivamente, 1,3% e 3,2% enquanto que o Floresta Aleatória 19,2%. Já para os anos de 2020/2021 em relação a 2019, os modelos demonstraram recuo médio de 3,9% (Adaboost), 7% (Floresta aleatória) e 2,3% (Árvore de decisão).

A seguir, são exibidos os resultados obtidos com a aplicação das técnicas de balanceamento de dados . Na Tabela 9 são listados os melhores parâmetros aplicados de acordo com cada

modelo que foi aplicada a técnica de *Grid Search*.

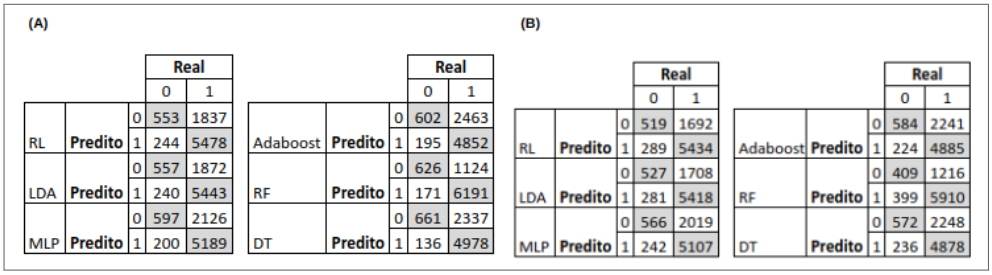
Tabela 9 – Melhores parâmetros (base balanceada)

Parâmetros (Modelo)	**Valores**
mtry (RF)	[2]
ntree (RF)	[150]
size (MLP)	[60]
maxit (MLP)	[250]
learnFuncParams (MLP)	[0.04]
maxdepth (DT)	[12]
maxvar (DT)	[9]
minprob (DT)	[0.01]

Fonte: Elaborada pelo autor (2022)

Na Figura 13 as matrizes de confusão são exibidas, onde na diagonal principal (destacada em cor cinza) representa as classificações corretas (parto prematuro e não prematuro) e os valores fora dessa diagonal os erros de classificação.

Figura 13 – Matriz de confusão - (A)base 2018 (B)base 2019 - (base balanceada)



Fonte: Elaborada pelo autor (2022)

Essas matrizes, permitem a obtenção das métricas de performance abordadas, as quais são apresentadas nas Tabelas 10 e 11 para, respectivamente, as bases de treinamento (2018) e validação (2019) de forma balanceadas.

Tabela 10 – Métricas de performance para a base de dados de 2018 (base balanceada)

Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Regressão logística	0,7435	0,9574	0,7489	0,8404	0,7737
Análise Discriminante Linear	0,7396	0,9578	0,7441	0,8375	0,7735
Multilayer perceptron	0,7133	0,9629	0,7094	0,8169	0,7895
AdaBoost	0,6723	0,9614	0,6633	0,7850	0,7694
Floresta aleatória	**0,8404**	0,9731	**0,8463**	**0,9053**	**0,8891**
Árvore de decisão	0,6951	**0,9734**	0,6805	0,8010	0,8258

Fonte: Elaborada pelo autor (2022)

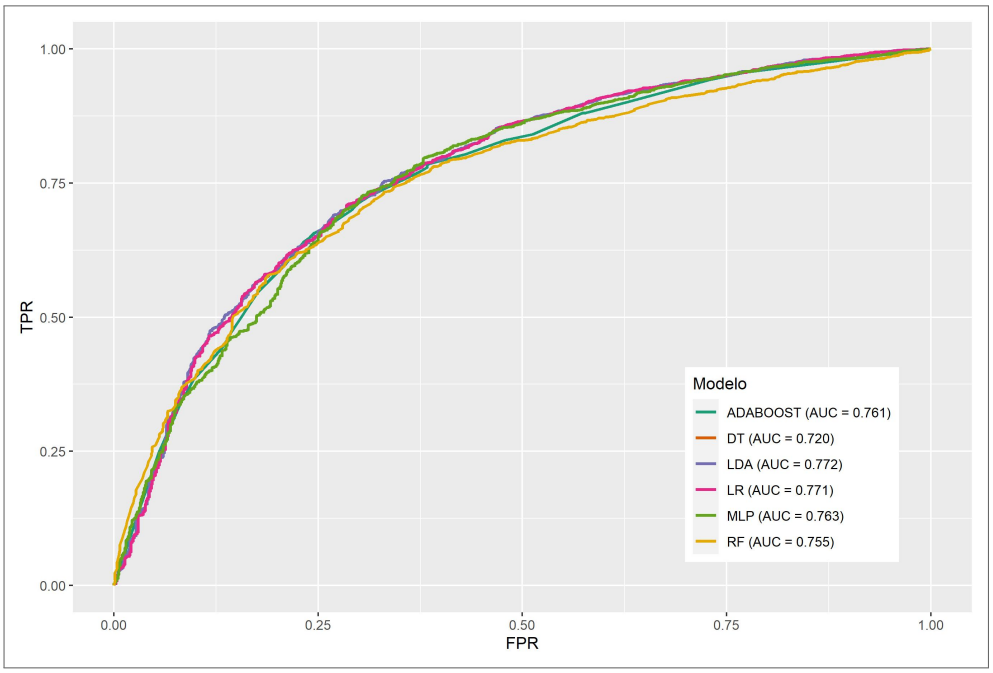
Tabela 11 – Métricas de performance para a base de dados de 2019 (base balanceada)

Modelo	Acurácia	Precisão	Recall	F1 score	AUC
Regressão logística	0,7503	0,9495	0,7626	0,8458	0,7705
Análise Discriminante Linear	0,7493	0,9507	0,7603	0,8449	**0,7717**
Multilayer perceptron	0,7150	0,9548	0,7167	0,8188	0,7634
AdaBoost	0,6893	**0,9562**	0,6855	0,7985	0,7608
Floresta aleatória	**0,7964**	0,9368	**0,8294**	**0,8798**	0,7548
Árvore de decisão	0,6869	0,9539	0,6845	0,7971	0,7199

Fonte: Elaborada pelo autor (2022)

Para a base de treinamento balanceada (Tabela 10), a métrica acurácia teve uma variação de 0,6723 (Adaboost) à 0,8404 (Floresta aleatória), enquanto que para a base de validação balanceada (tabela 11), essa variação foi de 0,6869 (Árvore de decisão) à 0,7964 (Floresta aleatória). Tal comportamento é similar para as métricas Recall, F1 score e AUC, onde o modelo Floresta aleatória apresenta bastante divergência de valores entre as duas bases. A precisão foi a métrica que obteve a menor variação dentre as demais, de 0,9575 (Regressão logística) à 0,9734 (Árvore de decisão) na base de 2018 balanceada e 0,9731 (Floresta aleatória) à 0,9578 (Análise discriminante linear).

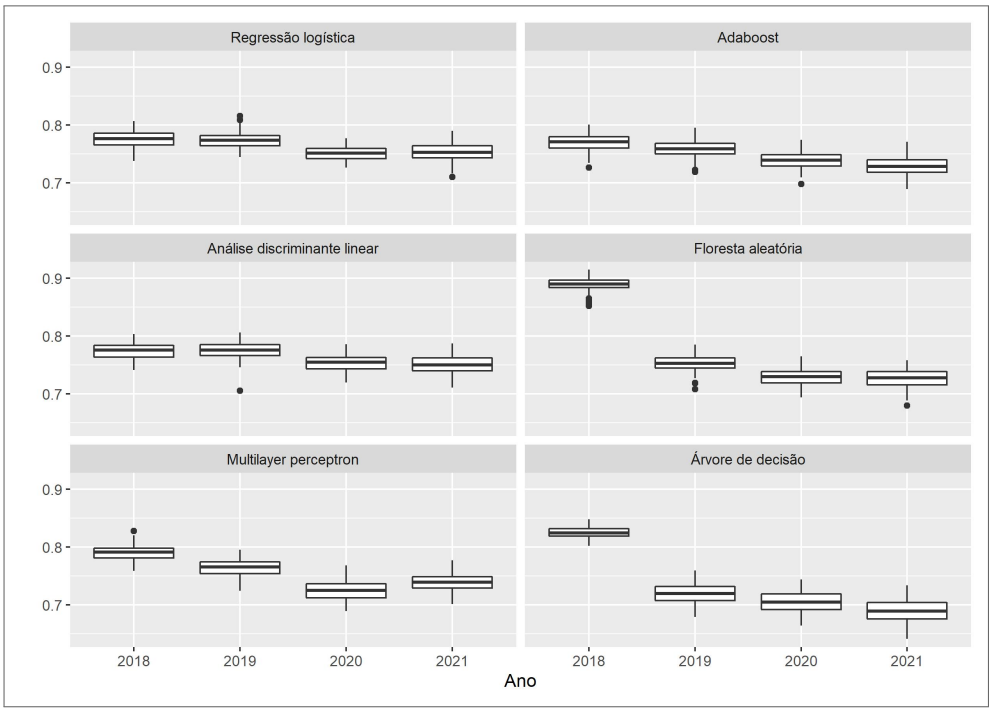
Figura 14 – AUC-2019 (base balanceada)



Fonte: Elaborada pelo autor (2022)

Na Figura 14, são exibidas a área sob a curva (AUC), que relaciona a taxa de falsos positivos (especificidade) e a taxa de verdadeiros positivos (sensibilidade). Nota-se dois grupos, um com valores acima de 75% como é o caso do LDA, LR e MLP e RF e um outro variando entre 70% e 72% para ADABOOST e DT.

Figura 15 – Boxplot AUC (base balanceada)



Fonte: Elaborada pelo autor (2022)

A Figura 15 contém os resultados da amostra bootstrap da métrica AUC 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 em sua mediana. Ainda, o modelo RF apresenta uma queda notória também para a base de dados de 2019.

Para complementar a análise da figura 15 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, conforme o Apêndice B.

A partir do teste post-hoc de Nemenyi (Apêndice B), evidenciou-se que para os modelos treinados com a base de dados balanceada: Os modelos de regressão logística, análise discriminante linear e multilayer perceptron, diferenças entre os anos de 2018-2020 (valor-p $<$ 0,001), 2018-2021 (valor-p $<$ 0,001), 2019-2020 (valor-p $<$ 0,001) e 2019-2021 (valor-p $<$ 0,001). A partir de 2020, a queda em relação a 2018/2019 em média foi de 3% para a regressão logística/análise discriminante linear e 6% para o multilayer perceptron. Já para os modelos Adaboost, Floresta aleatória e Árvore de decisão, as diferenças surgem da seguinte maneira: 2018-2019 (valor-p $<$ 0,001), 2018-2020 (valor-p $<$ 0,001), 2018-2021 (valor-p $<$ 0,001), 2019-2020 (valor-p $<$ 0,001) e 2019-2021 (valor-p $<$ 0,001). Para o ano de 2019 em relação a 2018, os modelos Árvore de decisão e Floresta Aleatória possuem uma diminuição de respec-

tivamente, 12,8% e 15,2% enquanto que o Adaboost 1,2%. Já para os anos de 2020/2021 em relação a 2019, os modelos demonstraram recuo médio de 4,1% (Adaboost), 3,2% (Floresta aleatória) e 3,2% (Árvore de decisão).

6 CONSIDERAÇÕES FINAIS

Esta pesquisa teve como objetivo a aplicação de algoritmos de aprendizado de máquina e averiguação das suas predições relacionadas a parto prematuro em gestantes únicas e verificar se nos dois primeiros anos da pandemia Covid-19 trouxeram impactos significativos para as distribuições das variáveis contidas na base de dados, em comparação ao que foi utilizado para treinamento dos modelos.

Os resultados demonstraram que em geral os modelos testados obtiveram performance preditiva competitiva em relação a estudos recentes (DRESSE, 2022). Além disso, a métrica AUC teve em alguns casos resultados acima do que foi observado em outros trabalhos da área como dados similares (dados não clínicos). Nos trabalhos relacionados, a AUC variou de 0,54 a 0,76, enquanto que nos achados deste trabalho, alguns modelos obtiveram AUC acima de 0,77, como a Regressão logística com 0,7729.

Quanto ao impacto da pandemia de COVID-19 sobre os classificadores, nota-se que as evidências apontam para um impacto sobre os classificadores mais estáveis: Regressão logística, Análise discriminante linear e Multilayer perceptron. Os demais classificadores apresentaram diferenças significativas de AUC entre 2019 e 2018 (teste post-hoc de Nemenyi), eliminando a hipótese de que a COVID-19 os afetou.

Ainda, os classificadores que apresentaram diferença da AUC na base de 2019 são todos baseados em árvore, o que pode ser um indicio de que esse tipo de classificador não seja o mais ideal para atacar este tipo de problema. Os modelos que tiveram estabilidade em 2019 em relação a 2018 são bastante consolidados na literatura, ótimas qualidades para modelos matemáticos de inferência computacional.

Pode-se concluir que os modelos baseados em árvore devem ser analisados com cautela, visto que os mesmos não são aderentes a base de treinamento. Já quanto aos modelos estáveis(Regressão logística, Análise discriminante linear e Multilayer perceptron), há indícios de que a COVID-19 afetou a estabilidade dos mesmos.

Portanto, em soluções elaboradas com bases de dados antes da COVID-19 devem ser monitoradas com cautela, e caso necessário, a realização de seu re-treinamento deve ser considerada.

Como trabalhos futuros, podem ser considerados: maior cobertura da idade gestacional; acrescentar mais variáveis (sociodemográficas e/ou clínicas) e considerar outros de modelos.

Além disso, como o Brasil é um país continental e por sua vez heterogêneo, seria interessante treinar os modelos de acordo com algum critério espacial, como a região do país por exemplo.

REFERÊNCIAS

ARNAEZ, J.; OCHOA-SANGRADOR, C.; CASERÍO, S.; GUTIÉRREZ, E. P.; JIMÉNEZ, M. d. P.; CASTAÑÓN, L.; BENITO, M.; PEÑA, A.; HERNÁNDEZ, N.; HORTELANO, M. et al. Lack of changes in preterm delivery and stillbirths during covid-19 lockdown in a european region. *European journal of pediatrics*, Springer, v. 180, n. 6, p. 1997–2002, 2021. Disponível em: <https://link.springer.com/article/10.1007/s00431-021-03984-6>.

BARANAUSKAS, J. A.; NETTO, O. P.; NOZAWA, S. R.; MACEDO, A. A. A tree-based algorithm for attribute selection. *Applied Intelligence*, Springer, v. 48, n. 4, p. 821–833, 2018. Disponível em: <https://doi.org/10.1007/s10489-017-1008-y>.

BARCZAK, A. L.; JOHNSON, M. J.; MESSOM, C. H. Empirical evaluation of a new structure for adaboost. In: *Proceedings of the 2008 ACM symposium on Applied computing*. [S.l.: s.n.], 2008. p. 1764–1765.

BARRETO, A. S. Modelos de regressão: Teoria e aplicação com o programa estatístico r. *Edição do Autor*, v. 1, p. 109–114, 2011.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

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

CASELLA, G.; BERGER, R. L. Transformations and expectations. *Statistical inference*, Wadsworth, v. 2, p. 47–55, 2002.

CHERMONT, A. G.; SILVA, E. F. A. da; VIEIRA, C. C.; FILHO, L. E. C. de S.; MATSUMURA, E. S. de S.; CUNHA, K. da C. Fatores de risco associados à prematuridade e baixo peso ao nascer nos extremos da vida reprodutiva em uma maternidade privada. *Revista Eletrônica Acervo Saúde*, n. 39, p. e2110–e2110, 2020.

CONGALTON, R. G.; GREEN, K. *Assessing the accuracy of remotely sensed data: principles and practices*. [S.l.]: CRC press, 2019.

DRESSE, Z. Predicting preterm birth with machine learning methods. New York, NY, USA, 2022. Disponível em: <http://lup.lub.lu.se/student-papers/record/9083000>.

EDUCATION, I. C. *Random Forest*. 2020. Disponível em: <https://www.ibm.com/cloud/learn/random-forest>. Acesso em: 5 aug. 2022.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011.

FERREIRA, A. P. A.; ALBUQUERQUE, R. C.; RABELO, A. R. de M.; FARIAS, F. C. de; CORREIA, R. C. de B.; GAGLIARDO, H. G. R. G.; SOUZA, A. C. V. M. de et al. Comportamento visual e desenvolvimento motor de recém-nascidos prematuros no primeiro mês de vida. *Journal of Human Growth and Development*, v. 21, n. 2, p. 335–343, 2011.

FERREIRA, D. F. Análise multivariada. *Lavras: UFLA*, v. 22, p. 394, 1996.

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">.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156.

GAMA JOÃO, a. Árvores de decisão. *Machine Learning, v. 55, p. 219–250, 2004. 22, 31, 32,50*, 2009.

GARCIA, S. C. O uso de árvore de decisão na descoberta de conhecimento na área da saúde. 2003. *Bacharelado em Estatística*, v. 26, 2003.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. [S.l.]: Alta Books, 2019.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009.

HEDERMANN, G.; HEDLEY, P. L.; BÆKVAD-HANSEN, M.; HJALGRIM, H.; ROSTGAARD, K.; POORISRISAK, P.; BREINDAHL, M.; MELBYE, M.; HOUGAARD, D. M.; CHRISTIANSEN, M. et al. Danish premature birth rates during the covid-19 lockdown. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, BMJ Publishing Group, v. 106, n. 1, p. 93–95, 2021.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. *Nonparametric statistical methods*. [S.l.]: John Wiley & Sons, 2013.

HOWSON. Born too soon: the global action report on preterm birth. 2012.

HUG, L.; SHARROW, D.; YOU, D. *Levels & Trends in Child Mortality: Report. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation*. [S.l.]: United Nations Inter-Agency Group for Child Mortality Estimation, New York . . . , 2017.

HUSEYNOVA, R.; MAHMOUD, L. B.; ABDELRAHIM, A.; HEMAID, M. A.; ALMUHAINI, M. S.; JAGANATHAN, P. P.; HUSEYNOV, O. Prevalence of preterm birth rate during covid-19 lockdown in a tertiary care hospital, riyadh. *Cureus*, Cureus, v. 13, n. 3, 2021.

INGARGIOLA, G. Building classification models: Id3 and c4. 5. *Disponível por WWW em: http://www. cis. temple. edu/˜ ingargio/cis587/readings/id3-c45. html*, 1996.

ITALO, J. *Tipos de aprendizado*. 2018. Disponível em: <https://medium.com/brasil-ai/tipos-de-aprendizagem-1c1339f73bdf>.

KOIVU, A.; SAIRANEN, M. Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health information science and systems*, Springer, v. 8, n. 1, p. 1–12, 2020.

KRAMER, M. S.; PAPAGEORGHIOU, A.; CULHANE, J.; BHUTTA, Z.; GOLDENBERG, R. L.; GRAVETT, M.; IAMS, J. D.; CONDE-AGUDELO, A.; WALLER, S.; BARROS, F. et al. Challenges in defining and classifying the preterm birth syndrome. *American journal of obstetrics and gynecology*, Elsevier, v. 206, n. 2, p. 108–112, 2012.

LEE, K.-S.; AHN, K. H. Application of artificial intelligence in early diagnosis of spontaneous preterm labor and birth. *Diagnostics*, Multidisciplinary Digital Publishing Institute, v. 10, n. 9, p. 733, 2020.

LEE, K.-S.; SONG, I.-S.; KIM, E.-S.; AHN, K. H. Determinants of spontaneous preterm labor and birth including gastroesophageal reflux disease and periodontitis. *Journal of Korean medical science*, Korean Academy of Medical Sciences, v. 35, n. 14, 2020.

LUNARDON, N.; MENARDI, G.; TORELLI, N. Rose: A package for binary imbalanced learning. *R journal*, v. 6, n. 1, 2014. Disponível em: <https://journal.r-project.org/archive/2014-1/menardi-lunardon-torelli.pdf>.

MARTINELLI, K. G.; DIAS, B.; LEAL, M. L.; BELOTTI, L.; GARCIA, É. M.; NETO, E. T. d. S. et al. Prematuridade no brasil entre 2012 e 2019: dados do sistema de informações sobre nascidos vivos. *Revista Brasileira de Estudos de População*, SciELO Brasil, v. 38, 2021.

MAS-CABO, J.; PRATS-BOLUDA, G.; GARCIA-CASADO, J.; ALBEROLA-RUBIO, J.; PERALES, A.; YE-LIN, Y. Design and assessment of a robust and generalizable ann-based classifier for the prediction of premature birth by means of multichannel electrohysterographic records. *Journal of Sensors*, Hindawi, v. 2019, 2019.

MENARDI, G.; TORELLI, N. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, Springer, v. 28, n. 1, p. 92–122, 2014.

MINGOTI, S. A. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2007. p. 295–295.

MONICO, J. F. G.; POZ, A. P. D.; GALO, M.; SANTOS, M. C. D.; OLIVEIRA, L. C. D. Acurácia e precisão: revendo os conceitos de forma acurada. *Boletim de Ciências Geodésicas*, Universidade Federal do Paraná, v. 15, n. 3, p. 469–483, 2009.

MOREIRA, S. *Rede Neural Perceptron Multicamadas*. 2020. Disponível em: <https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9>. Acesso em: 5 aug. 2022.

MOTOI, H.; JEONG, J.-W.; JUHASZ, C.; MIYAKOSHI, M.; NAKAI, Y.; SUGIURA, A.; LUAT, A.; SOOD, S.; ASANO, E. Quantitative analysis of intracranial electrocorticography signals using the concept of statistical parametric mapping. *Scientific Reports*, v. 9, p. 17385, 11 2019.

MWANIKI, M. K.; ATIENO, M.; LAWN, J. E.; NEWTON, C. R. Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: a systematic review. *The Lancet*, Elsevier, v. 379, n. 9814, p. 445–452, 2012.

OBERMEYER, Z.; LEE, T. H. Lost in thought: the limits of the human mind and the future of medicine. *The New England journal of medicine*, NIH Public Access, v. 377, n. 13, p. 1209, 2017.

OPAS. Folha informativa–covid-19: doença causada pelo novo coronavírus. 2020.

ORACLE, C. S. *O que é inteligência artificial (IA)?* 2014. Disponível em: <https://www.oracle.com/br/artificial-intelligence/what-is-ai/>.

PATBERG, E. T.; ADAMS, T.; REKAWEK, P.; VAHANIAN, S. A.; AKERMAN, M.; HERNANDEZ, A.; RAPKIEWICZ, A. V.; RAGOLIA, L.; SICURANZA, G.; CHAVEZ, M. R. et al. Coronavirus disease 2019 infection and placental histopathology in women delivering at term. *American journal of obstetrics and gynecology*, Elsevier, v. 224, n. 4, p. 382–e1, 2021.

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.

REZAEIAN, A.; REZAEIAN, M.; KHATAMI, S. F.; KHORASHADIZADEH, F.; MOGHADDAM, F. P. Prediction of mortality of premature neonates using neural network and logistic regression. *Journal of Ambient Intelligence and Humanized Computing*, Springer, p. 1–9, 2020.

SALLABA, F. *The potential of support vector machine classification of land use and land cover using seasonality from MODIS satellite data.* Tese (Doutorado), 06 2011.

SENTILHES, L.; MARCILLAC, F. D.; JOUFFRIEAU, C. et al. Covid-19 in pregnancy was associated with maternal morbidity and preterm birth [published online ahead of print, 2020 jun 15]. *Am J Obstet Gynecol*, p. 30639–6, 2020.

SHI, T.; HORVATH, S. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 15, n. 1, p. 118–138, 2006.

SOUZA, L. D. A. d. Uso de regressão logística para identificar os fatores de risco associados à ocorrência de anomalias congênitas em recém-nascidos. Estatística, 2014.

SVC, e. d. V. e. S. *Sistema de Informações de Nascidos Vivos - SINASC*. 2022. Disponível em: <"url:https://svs.aids.gov.br/daent/cgiae/sinasc/apresentacao/">.

SZWARCWALD, C. L.; LEAL, M. d. C.; ESTEVES-PEREIRA, A. P.; ALMEIDA, W. d. S. d.; FRIAS, P. G. d.; DAMACENA, G. N.; SOUZA, P. R. B. d.; ROCHA, N. M.; MULLACHERY, P. M. H. Avaliação das informações do sistema de informações sobre nascidos vivos (sinasc), brasil. *Cadernos de Saúde Pública*, SciELO Brasil, v. 35, 2019.

VOGEL, J. P.; CHAWANPAIBOON, S.; MOLLER, A.-B.; WATANANIRUN, K.; BONET, M.; LUMBIGANON, P. The global epidemiology of preterm birth. *Best Practice & Research Clinical Obstetrics & Gynaecology*, Elsevier, v. 52, p. 3–12, 2018.

WASTNEDGE, E. A.; REYNOLDS, R. M.; BOECKEL, S. R. V.; STOCK, S. J.; DENISON, F. C.; MAYBIN, J. A.; CRITCHLEY, H. O. Pregnancy and covid-19. *Physiological reviews*, American Physiological Society Bethesda, MD, v. 101, n. 1, p. 303–318, 2021.

WEI, S. Q.; BILODEAU-BERTRAND, M.; LIU, S.; AUGER, N. Increased pregnancy problems with covid-19–meta-analysis and letter to editor–april 2021. *CMAJ*, v. 193, n. 16, p. E540–E548, 2021.

APÊNDICE A – RESULTADO TESTE KW E NEMENYI (BASE DESBALANCEADA)

Tabela 9 – Resultado teste Kruskal-Wallis e Nemenyi (base desbalanceada)

Ano	Media AUC		Coef. Var.	Kruskal-Wallis	Dif18-All	Dif19-All	Dif20-All	Modelo
2018	0,7698	0,7699	0,0145	0,0189	1,73E-32			
2019	0,7732	0,7725	0,0155	0,0200	0,0033\n(p-valor 0,7264)			
2020	0,7529	0,7534	0,0135	0,0179	-0,0169 (p-valor 5,5E-11)	-0,0202 (p-valor 8,2E-14)	-	Regressão logística
2021	0,7468	0,7453	0,0149	0,0200	-0,0229 (p-valor 4,9E-14)	-0,0263 (p-valor 3,7E-14)	-0,0060 (p-valor 0,0941)	
2018	0,7501	0,7505	0,0161	0,0214	6,46E-59			
2019	0,7203	0,7193	0,0168	0,0234	-0,0297 (p-value 6,0e-09)			
2020	0,6922	0,6932	0,0155	0,0225	-0,0578 (p-value 2e-16)	-0,0281 (p-value 8,2e-14)		Adaboost
2021	0,6917	0,6922	0,0166	0,0239	-0,0583 (p-valor 2e-16)	-0,0286 (p-valor 8,1e-14)	-0,0005 (p-valor 0,9994)	
2018	0,7716	0,7708	0,0133	0,0172	2,32E-21			
2019	0,7714	0,7704	0,0154	0,0199	-0,0001 (p-value 0,9908)			
2020	0,7545	0,7557	0,0142	0,0189	-0,0171 (p-valor 1,2e-12)	-0,0169 (p-valor 1,0e-11)		Análise Discriminante Linear
2021	0,7548	0,7541	0,0165	0,0218	-0,0168 (p-valor 1,4e-11)	-0,0166 (p-valor 1,1e-10)	0,0002 (p-valor 0,9868)	
2018	0,8738	0,8745	0,0115	0,0132	2,21E-71			
2019	0,7053	0,7061	0,0150	0,0213	-0,1684 (p-valor 1,7e-09)			
2020	0,6540	0,6541	0,0170	0,0259	-0,2198 (p-valor 2e-16)	-0,0513 (p-valor 3,7e-14)		Floresta aleatória
2021	0,6584	0,6588	0,0174	0,0264	-0,2153 (p-valor 2e-16)	-0,0468 (p-valor 4,2e-14)	0,0044 (p-valor 0,725)	
2018	0,7749	0,7756	0,0162	0,0209	1,62E-39			
2019	0,7725	0,7713	0,0152	0,0197	-0,0024 (p-valor 0,8963)			
2020	0,7495	0,7502	0,0141	0,0189	-0,0254 (p-valor 3,8e-14)	-0,0230 (p-valor 3,3e-14)		Multilayer perceptron
2021	0,7469	0,7458	0,0140	0,0188	-0,0280 (p-valor 3,5e-14)	-0,0255 (p-valor 3,3e-14)	-0,0025 (p-valor 0,7519)	
2018	0,7492	0,7492	0,0169	0,0225	6,60E-45			
2019	0,7249	0,7238	0,0167	0,0230	-0,0243 (p-valor 1,5e-10)			
2020	0,7053	0,7059	0,0149	0,0211	-0,0439 (p-valor 2e-16)	-0,0195 (p-valor 3,5e-10)		Árvore de decisão
2021	0,7108	0,7112	0,0157	0,0221	-0,0384 (p-valor 2e-16)	-0,0141 (p-valor 1,6e-05)	0,0054 (p-valor 0,2458)	

Obs: As colunas com prefixo Dif **Ano**-All representam a diferença entre as médias da AUC do respectivo ano e os demais.

APÊNDICE B – RESULTADO TESTE KW E NEMENYI (BASE BALANCEADA)

Tabela 13 – Resultado teste Kruskal-Wallis e Nemenyi (base balanceada)

Ano	Media AUC		Coef. Var.	Kruskal-Wallis	Dif18-All	Dif19-All	Dif20-All	Modelo
2018	0,7763	0,7760	0,0137	0,0177	1,54E-37	-		
2019	0,7740	0,7735	0,0144	0,0186		-0,0023 (p-valor 0,7533)	-	
2020	0,7509	0,7512	0,0120	0,0160		-0,0254 (p-valor 3,5e-14)	-0,0230 (p-valor 3,7e-14)	- Regressão logística
2021	0,7533	0,7529	0,0152	0,0202		-0,0230 (p-valor 4,9e-14)	-0,0207 (p-valor 1e-14)	0,0023 (p-valor 0,598)
2018	0,7694	0,7710	0,0148	0,0192	1,64E-48	-		
2019	0,7599	0,7591	0,0156	0,0205		-0,0095 (p-valor 0,0129)	-	
2020	0,7381	0,7393	0,0151	0,0205		-0,0313 (p-valor 3,5e-14)	-0,0218 (p-valor 1,2e-12)	- Adaboost
2021	0,7292	0,7282	0,0152	0,0209		-0,0401 (p-valor 2e-16)	-0,0306 (p-valor 4,7e-14)	-0,0088 (p-valor 0,0354)
2018	0,7744	0,7757	0,0136	0,0175	2,00E-33	-		
2019	0,7750	0,7754	0,0147	0,0190		0,0006 (p-valor 0,9847)	-	
2020	0,7533	0,7546	0,0143	0,0190		-0,0210 (p-valor 4,3e-14)	-0,0217 (p-valor 3,2e-14)	- Análise Discriminante Linear
2021	0,7510	0,7501	0,0170	0,0226		-0,0233 (p-valor 2,9e-14)	-0,0239 (p-valor 4,4e-14)	-0,0022 (p-valor 0,9537)
2018	0,8890	0,8900	0,0115	0,0130	2,14E-63	-		
2019	0,7536	0,7529	0,0148	0,0196		-0,1353 (p-valor 6,1e-13)	-	
2020	0,7300	0,7301	0,0139	0,0190		-0,1590 (p-valor 2e-16)	-0,0236 (p-valor 7,0e-11)	- Floresta aleatória
2021	0,7262	0,7277	0,0162	0,0223		-0,1628 (p-valor 2e-16)	-0,0274 (p-valor 2,3e-13)	-0,0038 (p-valor 0,8561)
2018	0,7897	0,7912	0,0130	0,0164	7,57E-64	-		
2019	0,7639	0,7653	0,0146	0,0191		-0,0258 (p-valor 9,6e-09)	-	
2020	0,7244	0,7248	0,0179	0,0247		-0,0652 (p-valor 2e-16)	-0,0394 (p-valor 4,3e-14)	- Multilayer perceptron
2021	0,7391	0,7391	0,0158	0,0214		-0,0506 (p-valor 2e-16)	-0,0248 (p-valor 5,1e-10)	0,0146 (p-valor 0,0056)
2018	0,8252	0,8243	0,0104	0,0127	1,20E-59	-		
2019	0,7193	0,7198	0,0171	0,0238		-0,1058 (p-valor 3,4e-14)	-	
2020	0,7054	0,7050	0,0180	0,0255		-0,1197 (p-valor 2e-16)	-0,0138 (p-valor 0,0027)	- Árvore de decisão
2021	0,6899	0,6891	0,0189	0,0274		-0,1352 (p-valor 2e-16)	-0,0294 (p-valor 3,3e-12)	-0,0155 (p-valor 0,0011)

Obs: As colunas com prefixo Dif**Ano**-All representam a diferença entre as médias da AUC do respectivo ano e os demais.