InterSciencePlace
International Scientific Journal – ISSN: 1679-9844
Volume 19, Article nº 16, January/December 2024
D.O.I: http://dx.doi.org/10.6020/1679-9844/v19a16
Accepted: 21/04/2024 Published: 12/09/2024

DESENVOLVIMENTO DE MODELOS PREDITIVOS COM MACHINE LEARNING - ANÁLISE DE DADOS PARA SAÚDE DE GESTANTES E PUÉRPERAS.

DEVELOPMENT OF PREDICTIVE MODELS WITH MACHINE LEARNING - DATA ANALYSIS FOR THE HEALTH OF PREGNANT AND POSTPARTUM WOMEN.

Jackson Henrique da Silva Bezerra
Doutorando PGDRA/UFRO. Professor do Instituto Federal de Educação, Ciência e Tecnologia de Rondônia - Campus Ji-Paraná. E-mail: jackson.henrique@ifro.edu.br

Fabrício Moraes de Almeida
PhD in Physics (UFC), with post-doctorate in Scientific Regional Development (DCR/CNPq) - Specialization in Software Engineering (FUNIP). Researcher of the Doctoral and Master Program in Regional Development and Environment (PGDRA/UFRO). E-mail: dr.fabriciomoraes001@gmail.com

RESUMO

O *Machine Learning* (ML) tem um papel importante na área da saúde, fornecendo modelos preditivos criados a partir de algoritmos e grandes bases de dados. Estes modelos podem classificar pacientes para fins de diagnóstico ou prognósticos em diversas doenças. A presente pesquisa teve como objetivo o desenvolvimento de modelos preditivos de óbito por Síndrome Respiratória Aguda Grave (SRAG) em grupos populacionais vulneráveis na região Norte do Brasil. Para atingir este objetivo o estudo utilizou dados de gestantes e puérperas disponibilizados pelo Ministério da Saúde do Brasil. Como procedimento metodológico, foi realizado uma pesquisa aplicada através da metodologia CRISP-DM, que guiou todo o processo de seleção, processamento, transformação, aplicação dos algoritmos de ML e avaliação dos modelos preditivos. Os algoritmos *Random Forest*, *Regression Logistic,* K-*Nearest Neighbors* e XGBoost foram utilizados através do software Weka e biblioteca de código R, onde os modelos com *Random Forest* tiveram desempenho superior. Para garantir a confiança dos modelos foi utilizada a validação cruzada. Os modelos foram avaliados conforme as métricas de sensibilidade, especificidade, acurácia, precisão, F1-Score e AUC-ROC, sendo esta última a métrica primária de avaliação. Por fim, um protótipo de aplicação de software para uso dos modelos foi desenvolvido na

linguagem Java para que o conhecimento gerado pelo modelo chegue aos profissionais da área da saúde. Os resultados deste estudo contribuem para a redução de óbitos por SRAG no público materno da região Norte do Brasil, contribuindo para o cumprimento das metas do Brasil na redução da mortalidade materna.

Palavra-chave: *Machine Learning* (ML), Banco de dados, Síndrome Respiratória Aguda Grave (SRAG), Modelos Preditivos.

ABSTRACT

Machine Learning (ML) plays an important role in healthcare, providing predictive models created from algorithms and large databases. These models can classify patients for diagnostic or prognostic purposes in various diseases. The present research aimed to develop predictive models for death due to Severe Acute Respiratory Syndrome (SARS) in vulnerable population groups in the Northern region of Brazil. To achieve this objective, the study used data from pregnant and postpartum women made available by the Brazilian Ministry of Health. As a methodological procedure, applied research was carried out using the CRISP-DM methodology, which guided the entire process of selection, processing, transformation, application of ML algorithms and evaluation of predictive models. The Random Forest, Logistic Regression, K-Nearest Neighbors and XGBoost algorithms were used through the Weka software and R code library, where the Random Forest models had superior performance. To ensure the confidence of the models, cross-validation was used. The models were evaluated according to the metrics of sensitivity, specificity, accuracy, precision, F1-Score and AUC-ROC, the latter being the primary evaluation metric. Finally, a software application prototype for using the models was developed in the Java language so that the knowledge generated by the model reaches healthcare professionals. The results of this study contribute to the reduction of deaths from SARS in the maternal population in the Northern region of Brazil, contributing to the achievement of Brazil's goals in reducing maternal mortality.

Keywords: Machine Learning (ML), Database, Severe Acute Respiratory Syndrome (SARS), Predictive Models

1. INTRODUÇÃO

Nos últimos anos o *Machine Learning* (ML) vem se destacando como solução tecnológica importante na área da saúde, possibilitando a análise de grandes bases de dados para extração de conhecimento em tempo recorde, promovendo avanços

no aprimoramento de diagnósticos e a previsão de eventos clínicos (RAJKOMAR et al*.,* 2021). O ML fornece modelos interpretáveis que são compreensíveis para os seres humanos e que podem ser analisados, testados, verificados e/ou refutados utilizando experiências e dados reais ou outras abordagens baseadas no conhecimento. Produzem modelos baseados em regras ou árvores de decisão, que podem ser diretamente compreendidos por especialistas do domínio (tais como médicos, biólogos, epidemiologistas, decisores políticos, etc.) sem ter nenhuma formação específica em tecnologia (BOTTRIGHI et al., 2022). Segundo a OCDE (2022) o uso da IA na área da saúde ajudar a salvar vidas, melhorar o trabalho dos provedores de saúde e proteger a sociedade de emergências de saúde pública. A IA contribui de forma essencial para a saúde global, facilitando o aprendizado a partir de conjuntos de dados e populações de diferentes países, levando os sistemas de saúde a serem mais resilientes e sustentáveis, principalmente aqueles de regiões de baixa renda.

A região Norte do Brasil apresentou em 2022 uma elevada taxa de mortalidade materna (IBGE, 2022). Segundo Pazos et al. (2023) a morte materna é definida como óbito de uma mulher durante a gestação ou até 42 dias após o término da gestação, independentemente da duração e local da gravidez. A Razão de Mortalidade Materna (RMM) é o indicador utilizado para mensurar a taxa ou coeficiente de mortalidade: trata-se da relação entre o número de óbitos maternos por cada 100 mil NV. De 2012 a 2019 a Região Norte registrou, de 2012 a 2019, a média de RMM de 72,19/100 mil NVs. Em 2020, a RMM aumentou para 92,16/100 mil NVs; e, em 2021, a taxa foi de 164,17 óbitos maternos por 100 mil NVs, o que representou um crescimento expressivo da RMM, considerando os anos da pandemia de SARS-Cov-2. Assim, comparando os dados de RMM pré-pandemia, a região Norte possui números bem acima da média nacional, com 14.73 mortes maternas por 100 mil NV a mais (PAZOS et al., 2023).

A mortalidade materna é atualmente entendida como um indicador de desenvolvimento social por reunir fatores determinantes sociais complexos, por isso foi inserida em uma das metas a serem atingidas nos Objetivos de Desenvolvimento do Milênio (ODM), firmados por mais de 180 países com as Nações Unidas entre os anos de 2000 até 2015. O Brasil não conseguiu atingir a meta proposta até 2015, de reduzir em 75% a mortalidade materna e pactuou, em 2015, mediante as ODS, a

redução para 30 mortes maternas por 100 mil NVs até o ano de 2030 (PAZOS et al., 2023). Outro fator importante a ser considerado é o processo de adesão do Brasil como membro da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Segundo a OCDE (2021) no Estudo sobre a Atenção Primária à Saúde no Brasil de 2021, a saúde primária do Brasil melhorou nos últimos anos, porém o estudo aponta ações-chave que o Brasil deveria considerar nos próximos anos a fim de fortalecer o desempenho da atenção primária à saúde, como a busca de uma transformação digital, com uso de tecnologias de ponta como a IA (OCDE, 2021).

Considerando os dados de RMM da região, estudos para a criação de modelos preditivos de óbito por Síndrome Respiratória Aguda Grave (SRAG) para gestantes e puérperas com até 45 dias do parto são importantes. A SRAG é uma condição médica séria que envolve a deterioração rápida dos sintomas respiratórios, frequentemente levando a complicações graves e até mesmo risco de morte. Esta síndrome pode ser desencadeada por várias causas, incluindo infecções virais como Influenza A (H1N1) e SARS-CoV-2 (COVID-19), entre outros, bem como infecções bacterianas (LEE et al., 2024).

Para criação de modelos de ML eficientes é necessário a delimitação dos dados para que os algoritmos de ML aprendam os padrões específicos sobre gestantes e puérperas da região Norte, e então gerar conhecimento útil para auxiliar o tratamento de SRAG nestes grupos de vulneráveis. Assim, as bases de dados de SRAG de 2020 e 2021 do Ministério da Saúde disponível no portal openDataSUS foram escolhidas para aplicação do processo de engenharia de dados e ML. Os registros disponibilizados são capitados pelo Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe), que mantem o registro dos casos e óbitos por SRAG no Brasil, causada por vírus como SARS-Cov-2, Influenza A(H1N1), entre outros (BRASIL, 2024). Cabe destacar que as bases de dados SRAG do openDataSUS são publicadas nos formatos *Creative Commons Attribution (cc-by)* e *Open Data* que permite que outras pessoas compartilhem, remixem, adaptem e criem obras derivadas (BRASIL, 2024), estando de acordo com as diretrizes da Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil.

Por fim destaca-se que uma busca inicial na literatura sobre o uso do ML para predição de óbitos por SRAG nos repositórios PubMed e BVS – Biblioteca Virtual de

Saúde, demonstrou que poucos estudos foram realizados para grupos populacionais no Brasil (ARAÚJO et al., 2022), (FERNANDES et al, 2021), (SILVA & NETO, 2022). Referente a gestantes e puérperas foi encontrado apenas o trabalho dos autores Carneiro et al. (2022) que propuseram o uso de aprendizado não supervisionado com foco na identificação de sintomas, comorbidades e características hospitalares comuns de pacientes com SARS-Cov-2, ou seja, não teve como objetivo a criação e uso de modelos preditivos.

Neste contexto, considerando os índices preocupantes de RMM (PAZOS et al., 2023) na região Norte, os ODS pactuados pelo Brasil na Agenda 2030 para melhoria da saúde materna e o potencial do ML em auxiliar no tratamento de SRAG com a previsão de eventos clínicos (RAJKOMAR et al., 2021), faz necessário gerar modelos preditivos de óbito por SRAG específicos para o público materno na população brasileira, com foco na região Norte.

2. METODOLOGIA

Foi utilizada a metodologia CRISP-DM para aplicação do processo de engenharia de dados e criação dos modelos preditivos com algoritmos de ML. O CRISO-DM é um framework amplamente reconhecido e utilizado para guiar projetos de *Data Mining* (DM) e ML, contendo um ciclo de seis fases não rígidas movendo-se para frente e para trás entre diferentes fases sempre que necessário. O resultado de cada fase determina qual fase, ou atividade em particular de uma fase, deve ser executada em seguida (CHAPMAN et al., 2000). A aplicação da metodologia foi realizada conforme adaptação feita por Sena (2021), onde será guiada somente pelos objetivos e atividades de cada fase.

A primeira fase de Compreensão do Negócio (*Business Understanding*) concentrou-se em entender os objetivos e requisitos do projeto a partir de uma perspectiva do negócio, convertendo esse entendimento na definição de um problema a ser resolvido com o conhecimento adquiridos e um plano de metas para atingir estes objetivos. Nesta etapa também são avaliados os riscos e critérios técnicos para o projeto, os potenciais benefícios com projeto e por fim as metas e critérios de sucesso

para o projeto. Ferramentas para a análise, manipulação, transformação e criação dos modelos foram definidas nesta etapa.

Na segunda fase de Compreensão dos Dados (*Data Understanding*), o conjunto de dados foi examinado em profundidade, considerando todos os seus aspectos relevantes. Isso incluiu a identificação de anomalias, a verificação dos tipos de dados, a contagem de registros e outros fatores essenciais. Durante esta fase, foram realizadas atividades com foco na coleta de dados para o projeto, bem como na análise e exploração dos dados. O objetivo foi examinar suas propriedades e avaliar a qualidade dos dados, considerando a quantidade de variáveis disponíveis, o número de registros presentes e ausentes em cada variável, entre outros aspectos. Os dados utilizados neste estudo foram as bases de dados do ano de 2020 e 2021 do Base de Dados de SRAG do Ministério da Saúde (MS) do Brasil. Foram coletados dados de dois grupos de pacientes provenientes da região Norte do Brasil. No primeiro grupo, mulheres gestantes independentemente da idade. No segundo, mulheres puérpera ou parturiente que na base de dados corresponde a mulheres que pariram recentemente (no momento da coleta do dado) até 45 dias do parto. Cada base tem em média 86 atributos. Para explorar os dados foi utilizado o MySQL Workbench em conjuntos através da linguagem de programação SQL (*Structured Query Language*). Após o carregamento da base de dados no MySQL a manipulação dos registros da base de dados foi realizada através de códigos programados na linguagem SQL.

A terceira fase de Preparação dos Dados (*Data Preparation*) teve como objetivo transformar os atributos de modo a tornar o conjunto de dados adequado para aplicação dos algoritmos de ML. Durante esta fase foram desenvolvidas atividades com focos específicos, como a seleção das variáveis úteis a serem utilizadas no projeto, a limpeza nos dados para remoção de caracteres especiais como exemplo ponto e vírgula, a exclusão de variáveis inúteis para os algoritmos como datas e códigos internos, a transformação de dados para correção e/ou criação de novas variáveis a partir de outras variáveis e, por fim, a formatação dos dados para um padrão único e compreensível aos algoritmos e os aplicativos de ML. Toda a manipulação descrita nessa etapa foi realizada através da linguagem SQL diretamente na ferramenta MySQL.

Após análises e testes foram removidos atributos referentes a códigos internos de identificação e datas. Foram criados novos atributos a partir da atributos existentes como por exemplo o atributo como MES_SIN_PRI (Mês dos 1º sintomas) a partir de DT_SIN_PRI (Data dos 1º sintomas), FAIXA_IDADE_N (Faixa de idade por década) a partir de NU_IDADE_N (Idade do paciente), DIAS_UTI (Nº de dias na UTI) a partir dos atributos DT_ENTUTI (data de entrada na UTI) e DT_SAIDUTI (data de saída da UTI, entre outros. Para melhorar a interpretação dos modelos e facilitar a manipulação dos atributos e instancias na ferramenta de ML foi necessário transformar os dados da base de dados de acordo com o seu significado no dicionário de dados. Por exemplo para o atributo TOSSE o dado 1 foi transformado para Sim e o 2 para Não de acordo com o dicionário. Todas as alterações foram feitas através do script SQL. Após a manipulação dos dados no MySQL foi gerado via comando SQL um arquivo da base de dados na versão .CSV que pode ser lido pelo software Weka. Após carregamento da base de dados no Weka, a mesma foi salva no formato ARFF, padrão do Weka.

As bases de dados de 2020 e 2021 foram unificadas para facilitar a manipulação e seleção dos registros, com isso o atributo ANO foi criado para identificar o registro neste contexto. Depois disso as bases de dados foram divididas de acordo com os grupos de pacientes alvo da pesquisa e carregadas no Weka. Após este processo, o filtro *AttributeSelection* no Weka foi utilizado para selecionar os melhores atributos, onde foram utilizados os recursos *CorrelationAttributeEva* e *ClassifierAttributeEval* com o método *Ranker* que busca selecionar quais os melhores atributos de acordo com os algoritmos selecionados para o projeto. Também foi utilizado o filtro *NominalToBinary* no Weka para converter os atributos nominais em atributos numéricos binários em uma versão separada da base de dados. Essa conversão foi necessária para utilização de alguns algoritmos que não lidam com dados nominais, como o algoritmo XGBoost. Após o período de testes com os filtros foram descartados atributos que não se comportarem bem com os algoritmos escolhidos.

Na quarta fase de Modelagem (*Modeling*) os algoritmos de ML foram selecionados e aplicados nas bases de dados preparadas na etapa anterior, com a finalidade gerar modelos preditivos de acordo com os objetivos da pesquisa.

Compreende todo o processo de geração, validação, interpretação e seleção dos melhores modelos. Nesta fase foram desenvolvidas atividades com foco na escolha das técnicas de modelagem que serão utilizadas, a definição de métricas para aprovação dos modelos e a construção dos modelos com testes nos hiperparâmetros dos algoritmos. Já na quinta fase de Avaliação (*Evaluation*) os modelos são avaliados e aprovados, analisando se os conhecimentos adquiridos com estes modelos serão utilizados na etapa de implantação. Essas duas fases foram executadas de forma concomitante, uma vez que a geração e avaliação dos modelos fazem parte do mesmo processo.

A escolha dos algoritmos levou em consideração os algoritmos mais utilizados em trabalhos do gênero. Foram selecionados os algoritmos *Random Forest* (RF) que gera modelos de árvores com base em um conjunto de árvores de decisão (BENNET et al., 2021), (KIVRAK et al., 2021). O algortimo *Logistic Regression* (LR) que foi selecionado por sua capacidade de modelar a probabilidade de ocorrência de uma classe com base em variáveis independentes (HE et al., 2022), (KAR et al., 2021). E o algoritmo XGBoost (*Extreme Gradient Boosting*) que implementa algoritmos de ML na estrutura *Gradient Boosting*, proporcionando um aumento de precisão e eficiência (BÁRCENAS & FUENTES-GARCÍA, 2022). A combinação destes três algoritmos é comum literatura, sendo utilizada nos estudos de He et al. (2022), Heldt et al. (2021), Kar et al. (2021), Li J et al. (2022), Moulaei et al. (2022) e Zhao et al. (2022). Verificou-se também nos resultados da RI na Tabela 1 da página 22 que o número médio de algoritmos utilizados nos estudos é de aproximadamente quatro. Assim, também foi escolhido o algoritmo KNN (*K-Nearest Neighbors*) que busca realizar uma classificação dos dados baseado na proximidade em relação aos vizinhos mais próximos (SILVA E, 2022). A combinação entre estes quatro algoritmos também é comum, sendo utilizada nos estudos de Moulaei et al. (2022), Schöning et al. (2021) e Kivrak et al. (2021).

Por fim, na quinta e última fase de Implementação (*Deployment*) buscou-se descrever como o conhecimento adquirido será aplicado. Na metodologia CRISP-DM esta fase descreve a utilização do conhecimento gerado com projeto no âmbito de uma organização. Entretanto, como se trata de um trabalho acadêmico a primeira atividade desta fase foi adaptada para proporcionar a implantação do conhecimento

através de uma aplicação de software. A aplicação de software foi desenvolvida na linguagem Java com a ferramenta Apache Netbeans IDE 20 para o formato desktop, ou seja, instalável em qualquer dispositivo de PC. Foi utilizado a biblioteca de códigos do *weka.jar* para acesso a funcionalidades de carregamento, classificação e avaliação do modelo. A escolha da linguagem de programação Java deve-se ao fato da possiblidade de utilização da biblioteca weka.jar, além da experiência do autor com a linguagem.

Métricas de Avaliação na Fase de Modelagem e Avaliação

O processo de Validação Cruzada (*Cross-Validation*) foi usado para avaliar o desempenho e o erro geral de modelos. A validação cruzada é o procedimento de reamostragem usado para avaliar modelos de ML em uma amostra de dados. O procedimento possui um único parâmetro denominado k que expressa o número de grupos para dividir uma determinada amostra de dados. Na validação cruzada 10 vezes (nº de *folds* padrão), os modelos são treinados e testados dez vezes diferentes e, em seguida, as métricas médias de desempenho (ou seja, acurácia, precisão e assim por diante) são estimadas no final do processo (KIVRAK et al., 2021). Deve-se notar que a validação cruzada é uma técnica de validação amplamente aplicada e preferida em ML e DM devido à diferença do método convencional de instância dividida. Este método ajuda a reduzir o desvio no erro de previsão, aumenta o uso de dados tanto para treinamento quanto para validação, sem sobreajuste ou sobreposição entre os dados de teste e validação e evita que os dados sejam divididos arbitrariamente, que podem causar viés do resultado do modelo (MOULAEI et al., 2022). Para a validação cruzada de treinamento e testes dos modelos foram utilizadas 20 interações (*folds)* de acordo com achados na literatura (YU et al., 2021), (ZAREI et al., 2022), (AN et al., 2020), (SUN et al., 2021), (MAHDAVI et al., 2021). Assim, foram documentados os modelos gerados com as sementes que obtiveram o melhor desempenho.

Foram definidas métricas de avaliação dos modelos com base na literatura atual sobre o tema. A avaliação do desempenho do modelo é uma parte fundamental da construção de um modelo de ML eficaz. Para avaliar os modelos preditivos, são

aplicadas várias métricas, sendo as mais comuns a acurácia, especificidade, precisão, sensibilidade e critérios do gráfico da curva ROC (*Receiver Operating Characteristic*). Por fim, esses critérios de avaliação são comparados para determinar o modelo de predição com melhor (MOULAEI et al., 2022). Para calcular estas métricas de desempenho, é preciso obter a matriz de confusão do modelo gerado. A matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação, exibindo o número de previsões corretas e incorretas, organizadas de acordo com as classes reais e previstas. Ela permite a visualização dos acertos (VP - Verdadeiros Positivos e VN – Verdadeiros Negativos) e dos erros (FP - Falsos Positivos e FN – Falsos Negativos). Na matriz de confusão o VN corresponde ao número de resultados negativos classificados corretamente, VP é o número de resultados positivos classificados corretamente, FP é o número de resultados negativos classificados incorretamente como positivos e FN é o número de resultados positivos classificados incorretamente como negativos (BÁRCENAS et al., 2022), (BOOTH et al., 2021).

A Tabela 1 a seguir apresenta o formato da matriz de confusão e a posição dos acertos e erros.

		Valor Previsto	
		Óbito (+)	**Cura (−)**
Valor Real	**Óbito (+)**	VP	FN
	Cura (−)	FP	VN

Tabela *1* - Modelo da Matriz de Confusão. Fonte: Moulaei et al. (2022)
Nota: VP - Verdadeiros Positivos, VN – Verdadeiros Negativos, FP - Falsos Positivos e FN – Falsos Negativos

As métricas de acurácia, precisão, sensibilidade e especificidade dos modelos são calculadas a partir dos dados da matriz de confusão, conforme destacado na Tabela 2 a seguir.

Critérios de Desempenho	**Cálculo**
Acurácia	(VP + VN) / (VP + VN + FP + FN)
Precisão	VP / (VP + FP)
Sensibilidade (Recall)	VP / (VP + FN)
Especificidade	VN / (VN + FP)
F1-Score	2 * Precisão * Sensibilidade / (Precisão + Sensibilidade)

Tabela *2* - Cálculos dos Critérios de Desempenho. Fonte: Moulaei et al. (2022) e Bárcenas & Fuentes-García (2022)

Nota: VP - Verdadeiros Positivos, VN – Verdadeiros Negativos, FP - Falsos Positivos e FN – Falsos Negativos

A acurácia representa a porcentagem total de acertos de um modelo, entretanto essa métrica nem sempre é a melhor para avaliar modelos de classificação, especialmente em casos de bases de dados desbalanceadas. Em situações onde uma classe é significativamente mais frequente que outra, a acurácia pode ser enganosa, não refletindo o verdadeiro desempenho do modelo para todas as classes, induzindo o analista a acreditar que o modelo é bom ao prever corretamente a classe A, enquanto comete muitos erros ao prever a classe B. Assim, é importante considerar outras métricas além da acurácia, como a precisão, a sensibilidade e a métrica F1-score. A precisão mede a capacidade do modelo de evitar falsos positivos, indicando o percentual de acertos entre todas as instâncias classificadas como positivas. A sensibilidade, ou *recall*, mostra a capacidade do modelo de identificar corretamente todas as instâncias positivas, indicando o percentual de acertos entre todas as instâncias que são de fato positivas. A métrica F1-Score combina precisão e sensibilidade em uma média harmônica, proporcionando uma avaliação equilibrada do desempenho do modelo, especialmente em base de dados desbalanceadas (SILVA & NETO, 2022).

Outra métrica importante é curva ROC e o cálculo da AUC (*Area Under the Curve*). A curva ROC mensura a capacidade de predição do modelo por meio das taxas de sensibilidade e especificidade, representando essas métricas em um gráfico. A AUC quantifica a área total sob a curva ROC e fornece uma única métrica para o desempenho do modelo, independente do limiar de decisão específico. Essa técnica serve para visualizar, organizar e classificar o modelo com base na performance preditiva. Em termos práticos, quanto mais próxima do canto superior esquerdo do gráfico a curva estiver, melhor é o desempenho do modelo (SILVA & NETO, 2022). A AUC é o resultado da integração de todos os pontos durante o trajeto da curva, e computa simultaneamente a sensibilidade e a especificidade, sendo um estimador do comportamento da acurácia global do teste. Ela fornece uma estimativa da probabilidade de classificação correta de um sujeito ao acaso (acurácia do teste); por exemplo, uma AUC de 0,7 reflete uma chance de classificação correta de 70% do caso. De forma geral, os valores da AUC são interpretados como: 0.5-0.6 (péssimo),

0.6-0.7 (ruim), 0.7-0.8 (pobre), 0.8-0.9 (bom), > 0.9 (excelente) (POLO & MIOT, 2020). Por fim, destaca-se que se encontra na literatura diversos autores (KIVRAK et al., 2021), (SILVA & NETO, 2022), (FERNANDES et al., 2021), (VEPA et al., 2021), (BÁRCENAS et al., 2022), (SUN et al., 2021), (BENNETT et al., 2021), (ARAÚJO et al., 2022) que utilizaram as métricas de acurácia, sensibilidade, especificidade, precisão, F1-Score e a AUC-ROC na avaliação dos seus modelos de predição de óbito por SRAG. Neste contexto, conforme Bennett et al. (2021) e Moulaei et al. (2022) foi considerado a AUC-ROC como métrica primária e a sensibilidade, especificidade, acurácia, precisão e F1-Score como métricas secundárias para a avaliação e definição do melhor modelo.

Analisar a importância dos atributos em modelos de predição é essencial para compreender quais fatores influenciam mais os resultados. A avaliação da importância dos atributos em um modelo de *Random Forest* utiliza a redução do índice de Gini para determinar quais variáveis contribuem mais significativamente para a predição dos resultados, destacando os fatores mais influentes na classificação (MOSLEHI et al., 2022), (BÁRCENAS & FUENTES-GARCÍA, 2022). Para obter o índice de Gini foi necessário a utilizado na biblioteca R, utilizada através da interface Weka, uma vez que a biblioteca original do Weka não gera o índice diretamente. Para isso o script foi programado e executado:

```
1. library(randomForest)
2. data <- rdata
3. data_sem_missing <- na.omit(data)
4. modelo <- randomForest(EVOLUCAO ~ ., data = data_sem_missing)
5. importancia <- importance(modelo)
6. print(importancia)
```

A representação do índice em gráfico é comum na literatura (KUMARAN et al., 2022), (MOSLEHI et al., 2022), (BÁRCENAS & FUENTES-GARCÍA, 2022), (ZHAO et al., 2022) (AZNAR-GIMENO et al., 2021), (HELDT et al., 2021) e facilita a compreensão. Assim os índices Gini dos modelos com *Random Forest* foram demonstrados por gráficos.

Experimento de Balanceamento na Fase de Modelagem e Avaliação

Nesta fase foi identificado um desbalanceamento nos dados. Moulaei et al. (2022) destaca que uma das principais barreiras aos algoritmos de ML é o problema de dados desequilibrados. Isso ocorre quando as classes não são categorizadas igualmente. Consequentemente, os modelos treinados geralmente fornecem resultados preconceituosos em relação à classe dominante, causando uma possível tendência em categorizar novas observações para a classe majoritária. Analisando os estudos da RI verificou-se que os autores abordaram o desequilíbrio de formas distintas. Azgnar-Gimeno et al. (2021), Moulaei et al. (2022), Heldt et al. (2021), Zarei et al. (2022), Araújo et al. (2022) e Vepa et al. (2021) utilizaram a Técnica de Sobreamostragem Minoritária Sintética (SMOTE) para equilibrar o conjunto de dados, essa técnica consiste na criação de instancias sintéticas da classe minoritária com base nos padrões conhecidos dos dados da classe minoritária, na mesma proporção da classe majoritária (ARAÚJO et al., 2022), (MOULAEI et al., 2022). Já os estudos de Li J et al. (2022), Schöning et al. (2021), Booth et al. (2020), Gao et al. (2020) e An et al. (2020) utilizaram a técnica de ponderação de classes por pesos, afim de ajustar automaticamente os pesos das instâncias de forma que cada classe tenha uma importância igual durante o treinamento do modelo (BOOTH et al., 2020), (LI J et al., 2022). Por fim, autores como Woo et al. (2022), Yadaw et al. (2020), Bottrighi et al. (2022), Li Y et al. (2020), Yu L et al. (2021) e Bárcenas & Fuentes-García (2022) assumiram que os dados estavam desequilibrados e não lidaram com balanceamento. Assim, foi realizado um experimento de balanceamento com diferentes técnicas com o objetivo identificar possíveis melhorias no desempenho do modelo e as implicações práticas do balanceamento conforme literatura atual. O experimento foi realizado com a base de dados de gestantes e o algoritmo *Random Forest.*

A Tabela 3 a seguir demonstra que o resultado com o balanceamento com o SMOTE possui um desempenho superior na Sensibilidade e F1-Score em comparação aos demais modelos, porém com pouca variação no desempenho referente a AUC-ROC. Entretanto, esse ganho de desempenho se deve ao custo da criação sintética de muitas instâncias para a classe Óbito, que podem representar padrões inexistentes nos dados reais. Já o balanceamento por pesos realizado com o

filtro *ClassBalancer* do Weka apresentou um desempenho ligeiramente superior na Sensibilidade em comparação ao modelo desbalanceado, porém com desempenho inferior referente a AUC-ROC. Neste sentido, optou-se por utilizar os dados desbalanceados uma vez que não houve grandes avanços no desempenho com o balanceamento, seguindo a abordagem de Araújo et al. (2022) que indica que estudos recentes indicaram que “o desequilíbrio não é um problema em si: os métodos de correção do desequilíbrio podem causar uma calibração deficiente e até piorar o desempenho do modelo em termos do AUC-ROC”. Ademais, a métrica F1-Score avaliada nessa pesquisa fornece uma avaliação global do modelo, independentemente da quantidade de amostras em cada uma das classes.

Métricas	**Desbalanceado**	**Balanceado com SMOTE**	**Balanceado com *ClassBalancer***	**Média**	**Desvio Padrão**
Verdadeiros Positivos (TP)	230	3523	2076	1943	1651
Falsos Positivos (FP)	15	28	33	25	9
Verdadeiros Negativos (TN)	3442	3429	1827	2899	929
Falsos Negativos (FN)	114	89	336	180	136
Precisão	0.939	0.992	0.984	0.972	0.029
F1-Score	0.781	0.984	0.918	0.894	0.104
Sensibilidade	0.669	0.975	0.860	0.835	0.155
Especificidade	0.996	0.992	0.982	0.990	0.007
Acurácia	0.966	0.983	0.913	0.954	0.037
AUC-ROC	0.958	0.997	0.959	0.971	0.022

Tabela *3* - Comparativo do Experimento de Balanceamento para Desfecho Óbito Positivo. Fonte: Autor

3. CARACTERÍSTICAS DOS RESULTADOS

Fase de Compreensão do Negócio

A problemática inicial identificada foi de que a saúde pública na região Norte do Brasil carece de tecnologias para auxiliar o tratamento de SRAG. As restrições do projeto estão nos tipos de dados e número de registros nas bases de dados, que podem limitar o uso de alguns algoritmos de ML. Os riscos do projeto estão na não geração de conhecimento útil para área da saúde como o modelo preditivo. As metas para o projeto de ML são: Meta 1 - Gerar no mínimo quatro modelos com quatro

algoritmos de ML diferentes para cada grupo populacional alvo da pesquisa; Meta 2: Avaliar os modelos com métricas confiáveis e selecionar o melhor modelo para utilização do mesmo em um aplicativo de classificação de pacientes com SRAG. Como etapa crítica se destaca a etapa de modelagem, uma vez que são realizados muitos testes na geração dos modelos e a cada teste diferente, algumas tarefas da fase de preparação de dados são executadas novamente, podendo consumir um tempo maior do que o planejado no projeto.

Foi realizado a avaliação das ferramentas a serem utilizadas na geração dos modelos. Assim, conforme descrito na seção Revisão Integrativa deste estudo, os principais softwares disponíveis atualmente para uso de algoritmos de ML são Weka, R *Project for Statistical Computinge* e as bibliotecas de ML da linguagem Python como *numpy, pandas, matplotlib* e *scikit-learn*. O Weka é um software de mineração de dados desenvolvido em Java, de código aberto e desenvolvido pela Universidade de Waikato da Nova Zelândia para fomentar os estudos sobre ML, sendo uma referência mundial no assunto (WAIKATO, 2024). O Projeto R para Computação Estatística é um ambiente de software livre para computação estatística e criação de gráficos, sendo amplamente utilizado entre estatísticos e mineradores de dados para o desenvolvimento de software estatístico e análise de dados (R CORE TEAM, 2024).

O Weka foi escolhido devido a possiblidade de utilizar as bibliotecas de ML da linguagem Python e a biblioteca do software R diretamente na interface Weka, transformando o Weka em uma ferramenta completa e com interface amigável para o usuário. A escolha da ferramenta está de acordo com a literatura sobre o tema, onde o Weka foi utilizado pelos autores Bottrighi et al. (2022) e Moulaei et al. (2022) em suas pesquisas sobre modelos preditivos de óbito por SRAG.

Fase de Compreensão dos Dados - Engenharia de Dados

O conjunto de dados sobre casos de SRAG disponibilizado no portal openDataSUS do Ministério de Saúde foi analisado. A base de dados de 2020 possui 1.196.665 registros de pacientes registrados no Sistema Único de Saúde com SRAG em todo o Brasil. O atributo classe da pesquisa EVOLUCAO que informa se o paciente foi curado ou veio a óbito possui 13% de dados perdidos. Já a base de dados de 2021

possui 1.185.228 registros e o atributo EVOLUCAO possui 27% de dados perdidos. Contando os atributos principais e os complementares à base de dados de 2021 possui 166 atributos e na base de dados de 2020 são 154. No geral pode-se concluir que as bases possuem um número elevado de registros e atributos com diversas informações dos pacientes, como sintomas, fatores de riscos, dados clínicos, demográficos, vacinais, entre outros.

Fase de Preparação dos Dados

O script programado na linguagem SQL utilizado na limpeza e transformação das bases de dados de 2020 e 2021 estão disponíveis repositório de arquivos Zenodo sob DOI – *Digital Object Identifier* no link: https://doi.org/10.5281/zenodo.10850628. A lista de todos os atributos excluídos pode ser verificada no Script SQL de limpeza a partir de linha 366 identificados com o comentário *#limpeza de base de atributos não selecionados.* A base de dados unificada com registros de 2020 e 2021 possui um total de 291.775 pacientes da região Norte considerados elegíveis para a aplicação dos modelos, estando disponível no repositório de arquivos Zenodo sob link https://zenodo.org/doi/10.5281/zenodo.12636544 formato ARFF que pode ser lido pelo Weka.

Após este processo de limpeza e transformação as bases de dados foram divididas em duas partes de acordo com os grupos alvo da pesquisa e também disponibilizadas no repositório Zenodo no formato ARFF sob link https://zenodo.org/doi/10.5281/zenodo.10884313 para gestantes e https://zenodo.org/doi/10.5281/zenodo.10884309 para puérperas. Destes pacientes, 4826 de gestantes e 1529 de puérperas. Devido ao registro de dados nulos no atributo classe “evolução” 3.204 registros foram excluídos. De tal modo foram considerados para a geração dos modelos 3801 de gestantes e 1252 de puérperas. Entre as gestantes, houve 3457 casos de cura e 344 óbitos. Entre as puérperas, houve 970 casos de cura e 282 óbitos.

Fase de Modelagem e Avaliação

Os algoritmos de ML escolhidos geram modelos de classificação de acordo com um desfecho, chamado de classe no ML. No projeto, o atributo EVOLUÇÃO fornece a informação da classe, podendo ser Cura ou Óbito. Assim, os modelos foram avaliados considerando a classe Óbito como positiva, uma vez que o objetivo do modelo preditivo é realizar a previsão de óbito de pacientes por SRAG.

Referente à validação cruzada, o recurso "Random Seed for XVal" do Weka foi testado com 20 sementes diferentes para cada algoritmo e base de dados. A acurácia foi avaliada e o desvio padrão das médias foi inferior a 0,01% em todos os testes, indicando que não houve diferença estatística significativa entre eles. Os modelos documentados foram gerados com as sementes 5 e 6 no algoritmo *Random Forest*; 13 e 8 no *Regression Logistic*; 3 e 15 no KNN; e 10 e 13 XGBoost para as bases de gestantes e puérperas respectivamente.

A fim de obter modelos de alta confiabilidade capazes de prever com eficiência a classe óbito, foram realizados diversos experimentos em busca dos melhores hiperparâmetros de cada modelo de ML analisado. A partir desses experimentos chegou-se aos seguintes hiperparâmetros: No *Random Forest* o número de árvores na floresta foi configurado como igual a 110; no KNN o número de vizinhos foi definido como igual a 1 e função de distância *Euclidean Distance*; no XGBoost foi utilizado a biblioteca do R via interface Weka com a base de dados transformada para dados binários; por fim no *Regression Logistic* as configurações padrões do Weka foram utilizadas.

Avaliação do Modelos para Gestantes

A Tabela 4 a seguir apresenta os dados da matriz de confusão dos modelos gerados para gestantes.

	Predição	
Random Forest	**Óbito (+)**	**Cura (−)**
Óbito (+)	247	97
Cura (−)	14	3443
Logistic Regression	**Óbito (+)**	**Cura (−)**
Óbito (+)	153	191
Cura (−)	71	3386
KNN	**Óbito (+)**	**Cura (−)**
Óbito (+)	302	42
Cura (−)	49	3408
XGBoost	**Óbito (+)**	**Cura (−)**
Óbito (+)	155	189
Cura (−)	46	3411

Tabela *4* - Matriz de Confusão dos Modelos para Gestantes. Fonte: Adaptado de Kivrak et al. (2021)

A Tabela 5 a seguir apresenta as métricas de desempenho dos algoritmos de ML nos modelos gerados para gestantes.

Algoritmos	**Sensibilidade**	**Especificidade**	**Acurácia**	**Precisão**	**F1-Score**	**AUC-ROC**
Random Forest	0.718	0.996	0.970	0.946	0.817	0.967
Logistic Regression	0.445	0.979	0.931	0.683	0.539	0.875
KNN	0.878	0.986	0.976	0.860	0.869	0.922
XGBoost	0.451	0.987	0.938	0.771	0.569	0.844

Tabela *5* - Avaliação de Desempenho dos Algoritmos nos Modelos para Gestantes. Fonte: Adaptado de Moulaei et al. (2022)

A Figura 1 a seguir apresenta os gráficos com a Curva ROC e a AUC de cada algoritmo para fins de comparação do desempenho dos modelos gerados para gestantes, onde verifica-se uma ligeira superioridade do modelo criado com o algoritmo *Random Forest*, seguido de perto pelo algoritmo KNN.

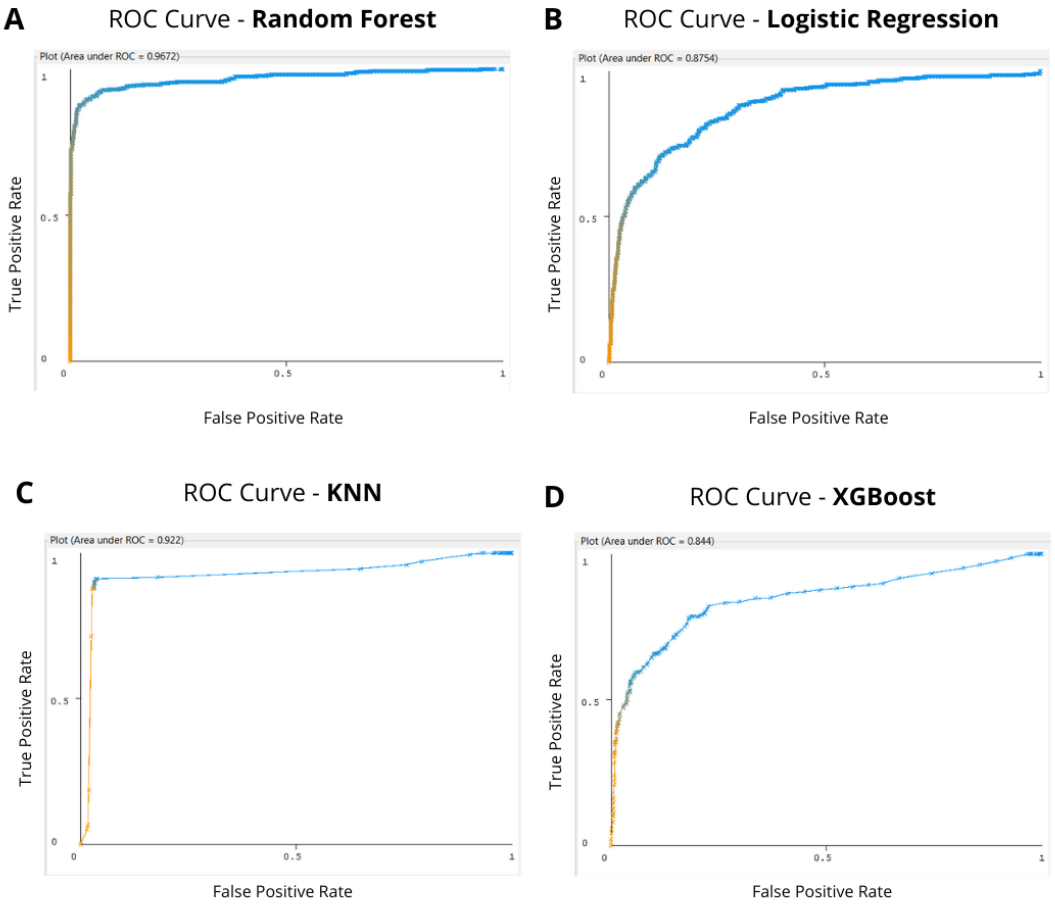

Figura 1 - AUC-ROC dos Modelos para Gestantes. *Fonte: Adaptado de Silva & Neto (2022)*

A Figura 2 a seguir apresenta o gráfico com os atributos mais importantes considerados pelo modelo com *Random Forest* para Gestantes.

Figura 2 - Gráfico com Índice Gini para Gestantes. *Fonte: Adaptado de Zhao et al. (2022)*

Avaliação do Modelos para Puérperas

A Tabela 6 a seguir apresenta os dados da matriz de confusão dos modelos gerados para puérperas.

	Predição	
Random Forest	**Óbito (+)**	**Cura (−)**
Óbito (+)	231	51
Cura (−)	24	946
Logistic Regression	**Óbito (+)**	**Cura (−)**
Óbito (+)	177	105
Cura (−)	76	894
KNN	**Óbito (+)**	**Cura (−)**
Óbito (+)	246	36
Cura (−)	47	923
XGBoost	**Óbito (+)**	**Cura (−)**
Óbito (+)	189	93
Cura (−)	67	903

Tabela *6* - Matriz de Confusão dos Modelos para Puérperas. Fonte: Adaptado de Kivrak et al. (2021)

A Tabela 7 a seguir apresenta as métricas de desempenho dos algoritmos de ML nos modelos gerados para puérperas.

Algoritmos	**Sensibilidade**	**Especificidade**	**Acurácia**	**Precisão**	**F1-Score**	**AUC-ROC**
Random Forest	0.819	0.975	0.940	0.906	0.860	0.975
Logistic Regression	0.628	0.922	0.855	0.700	0.662	0.881
KNN	0.872	0.952	0.933	0.840	0.856	0.886
XGBoost	0.670	0.931	0.872	0.738	0.703	0.890

Tabela *7* - Avaliação de Desempenho dos Algoritmos nos Modelos para Puérperas. Fonte: Adaptado de Moulaei et al. (2022)

A Figura 3 a seguir apresenta os gráficos com a Curva ROC e a AUC de cada algoritmo para fins de comparação do desempenho dos modelos gerados para puérperas, onde verifica-se o desempenho superior do modelo criado com o algoritmo *Random Forest*.

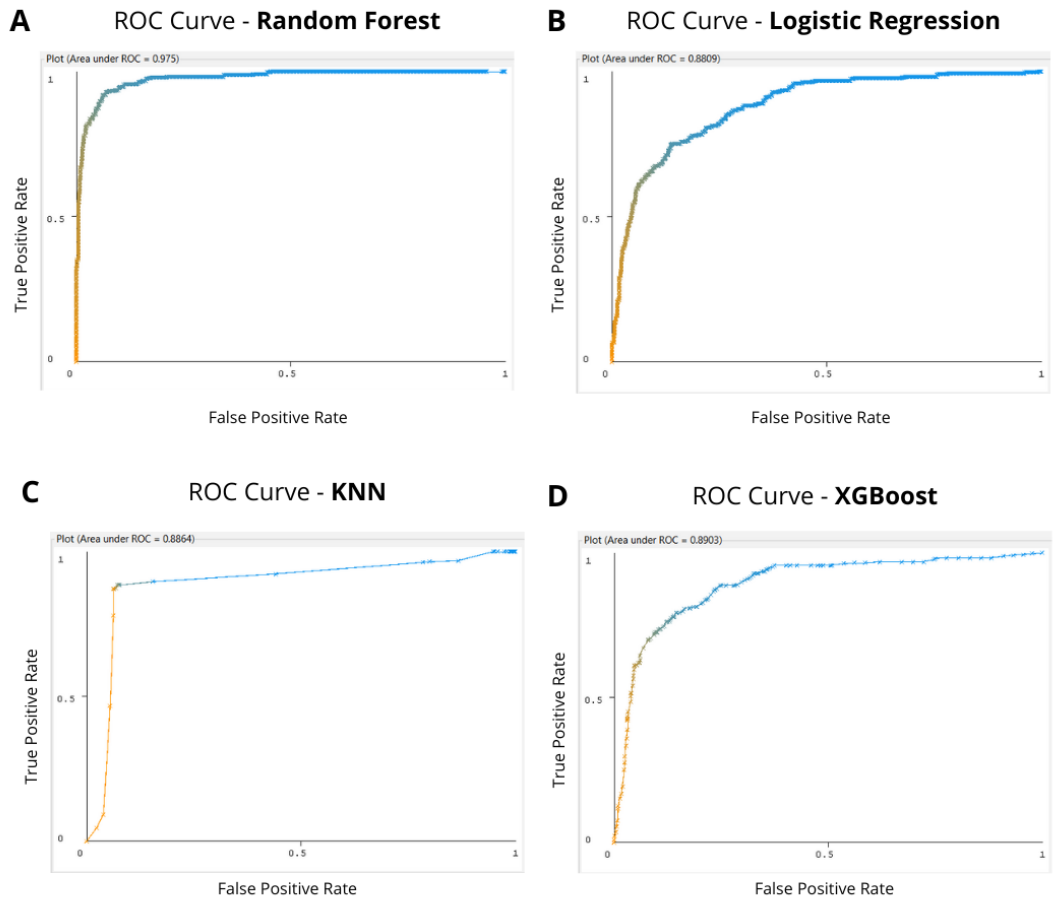

Figura 3 - AUC-ROC dos Modelos para Puérperas. Fonte: Adaptado de Silva & Neto (2022)

A Figura 4 a seguir apresenta o gráfico com os atributos mais importantes considerados pelo modelo com *Random Forest* para puérperas.

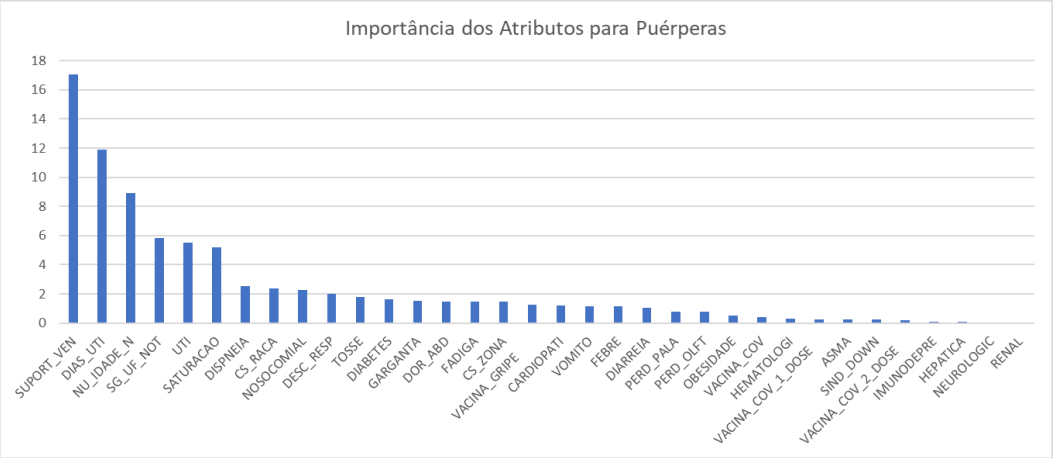

Figura 4 - Gráfico com Índice Gini para Puérperas. Fonte: Adaptado de Zhao et al. (2022)

Fase de Implementação

A seguir serão apresentadas imagens das funcionalidades da aplicação. No menu inicial o usuário poderá selecionar dois classificadores diferentes, conforme o perfil do paciente que o mesmo deseja classificar. Após a seleção do perfil Gestantes no menu inicial o sistema irá abrir o classificador para este grupo específico. O usuário então poderá informar as características do paciente a ser classificado através de caixas de combinação e então clicar no botão *Classificar*. Durante o processo de classificação o sistema exibirá uma barra de progresso enquanto a classificação ocorre. Após o fim do processo será de apresentada a chances de óbito para o paciente. A probabilidade é dada pelo modelo de predição que realiza a classificação do paciente com base no conjunto de características informadas na interface antes do clique do botão *Classificar*.

A Figura 5 a seguir apresenta a funcionalidade para classificação de Gestantes, após a ação de classificar, onde será apresentado as chances de óbito para o paciente conforme as características informadas.

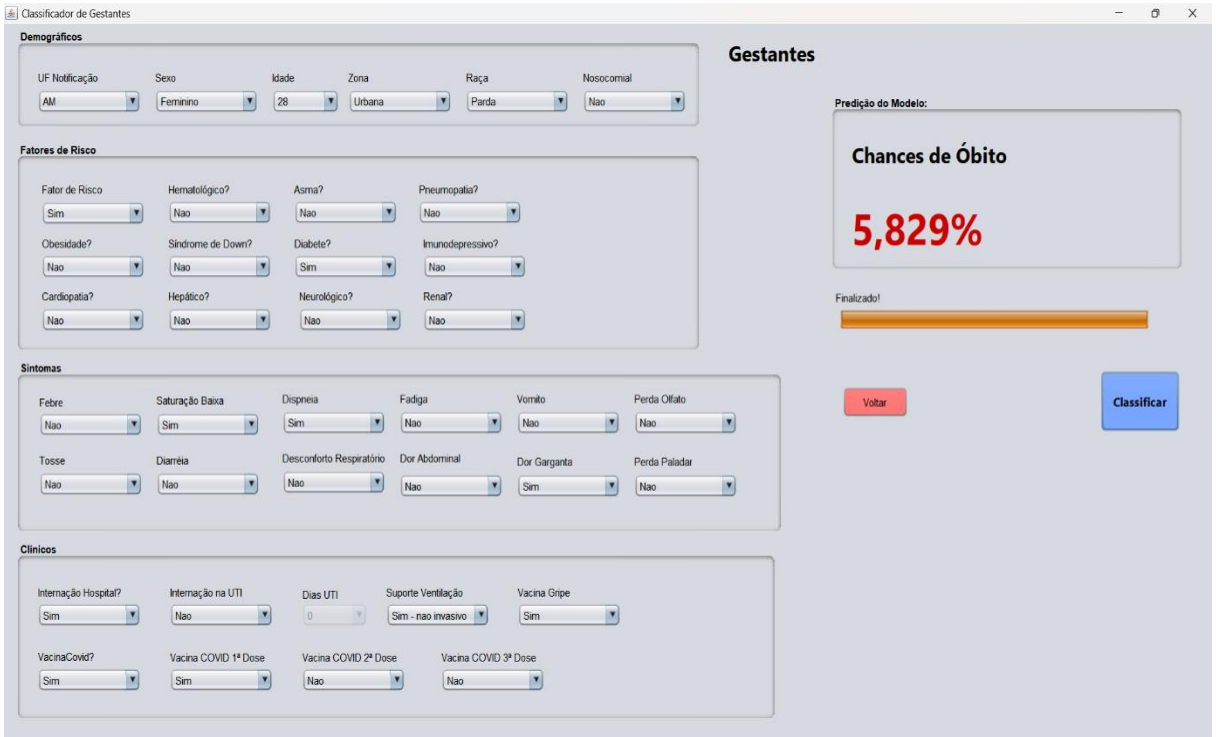

Figura 5 - Classificador de Gestantes com Resultado da Classificação. Fonte: Autores

O código fonte da aplicação está depositado no repositório de códigos GitHub e poderá ser acesso e baixado através do link: https://github.com/jacksonifro/Aplication_Tese_Doutorado.git. Para abrir a aplicação é necessário a ferramenta Apache Netbeans IDE 20. Já o setup de instalação da aplicação para ser instalado em sistemas operacionais Windows ou Linux está disponível para download no repositório Zenodo através do DOI: https://zenodo.org/doi/10.5281/zenodo.10951429.

4. ANÁLISE E DISCUSSÃO DOS RESULTADOS

Foram desenvolvidos e comparados modelos preditivos para classificação com quatros algoritmos diferentes: *Random Forest*, *Regression Logistic*, KNN e XGboost. Os modelos foram avaliados conforme as métricas de sensibilidade, especificidade, acurácia, precisão, F1-Score e AUC-ROC, sendo esta última a métrica primária de avaliação. Conforme destacado por Polo & Miot (2020), uma AUC-ROC superior a 0.90 é considerada um ótimo índice de performance de um modelo de dados quantitativos segundo sua taxa de sensibilidade (fração dos verdadeiros positivos) e a fração dos falsos positivos (1 - especificidade), segundo diferentes valores de corte do teste. Assim, as discussões a seguir consideram esse limiar para a avaliação da qualidade do modelo quanto à robustez e confiabilidade.

Para gestantes, o modelo gerado com o algoritmo *Random Forest* oferece um desempenho robusto e confiável, alcançando uma AUC-ROC de 0.967, sensibilidade de 0.718, especificidade de 0.996, acurácia de 0.970, precisão de 0.946 e F1-Score de 0.817. Embora o algoritmo KNN seja ligeiramente superior na sensibilidade com 0.878 e no F1-Score 0.869, o *Random Forest* também alcançou valores robusto na previsão da classe positiva. Verifica-se também que apesar do desequilíbrio nos dados, a diferença no F1-Score entre os modelos *Random Forest* e KNN foi de apenas 0.052 no F1-Score. Com isso, o *Random Forest* apresenta um melhor equilíbrio geral em todas as métricas, tornado sua performance na distinção entre classes superior

aos demais algoritmos analisados. Por esse motivo, o modelo gerado com *Random Forest* foi escolhido para a classificação de gestante.

Para o grupo de puérperas, o modelo gerado com o *Random Forest* também foi superior aos demais algoritmos, alcançando uma AUC-ROC de 0.975, sensibilidade de 0.819, especificidade de 0.975, acurácia de 0.940, precisão de 0.906 e F1-Score de 0.860, mostrando-se um modelo robusto e confiável na distinção entre classes. Apesar do algoritmo KNN apresentar uma ligeira vantagem na sensibilidade, com 0.872, o desempenho superior do *Random Forest* nas demais métricas o torna a melhor opção entre todos os algoritmos analisados. Por esse motivo, o modelo gerado com o *Random Forest* foi escolhido para a classificação de puérperas.

Estes resultados estão de acordo com a literatura sobre o tema. Heldt et al. (2021) avaliaram o desempenho dos algoritmos *Random Forest*, *Logistic Regression* e XGBoost usando um conjunto de dados de 619 pacientes ingleses com dados demográficos, clínicos e laboratoriais para prever a mortalidade por SARS-Cov-2. O *Random Forest* gerou o melhor modelo com AUC-ROC de 0.77, contra 0.70 e 0.76 do *Logistic Regression* e XGBoost respectivamente. Em outro estudo (MOULAEI et al., 2022), foram utilizados dados demográficos, clínicos, laboratórios e fatores de risco de 1.500 pacientes iranianos hospitalizados com SARS-Cov-2. Os resultados deste estudo mostraram que o modelo desenvolvido o algoritmo *Random Forest* apresentou o melhor desempenho, com AUC-ROC de 0.99 na previsão de morte do paciente, contra a AUC-ROC de outros algoritmos comparados como XGBoost (0.981), KNN (0.967), MLP (0.964), *Logistic Regression* (0.942), J48 (0.921) e *Naive Bayes* (0.920). Em um estudo com voltado para a população brasileira, Silva & Neto (2022) utilizou dados clínicos de 134.639 pacientes com SARS-Cov-2 registrados no Banco de Dados de SRAG do openDataSUS entre janeiro e setembro de 2021 para avaliar o desempenho dos algoritmos *Logistic Regression*, *Decision Tree* e *Random Forest* na criação de modelos preditivos de óbito. Neste estudo o *Random Forest* foi superior alcançando AUC-ROC de 0.75, acurácia de 0.77, precisão de 0.76, f1-score de 0.69 e sensibilidade de 0.63 para classe óbito. O algoritmo *Logistic Regression* alcançou uma AUC-ROC de 0.73 e o *Decision Tree* de 0.74, sendo inferiores ao *Random Forest* nessa e nas demais métricas, exceto pelo *Decision Tree* que foi ligeiramente superior na precisão com 0.78.

Com base nos índices de Gini do modelo com *Random Forest*, verificou-se que as métricas mais importantes para a predição dos modelos nos dois grupos de pacientes analisados foram os atributos SUPORT_VEN (Suporte a ventilação), DIAS_UTI (Número de dias na UTI), NU_IDADE_N (Idade do paciente), SG_UF_NOT (UF de notificação) e UTI (Internação na UTI). Essas variáveis desempenham um papel crucial na decisão do modelo, indicando que a necessidade de ventilação mecânica, a internação e o tempo na UTI, e a idade do paciente são os fatores mais determinantes. Além desses atributos, que são comuns aos dois grupos, outros se destacam e diferem entre os grupos analisados. No grupo de gestantes, os atributos OBESIDADE (Paciente obeso), FEBRE (Paciente apresentou febre), CS_RAÇA (Raça do paciente) e DISPNEIA (Paciente apresentou dispneia) são os mais relevantes. Por fim, no grupo de puérperas, os atributos SATURAÇÃO, DISPNEIA e CS_RAÇA se destacam como os mais importantes. Esta análise evidencia que, além dos fatores comuns a todos os grupos, há características específicas que influenciam a predição do modelo conforme o perfil dos pacientes.

Vários estudos identificaram características clínicas importantes como preditores de mortalidade em pacientes com SARS-Cov-2, utilizando técnicas de análise de características, como o índice de Gini obtido em algoritmos de ML como o *Random Forest*. Os atributos selecionados servem como insumos para o desenvolvimento de modelos de ML, visando prever a gravidade, deterioração e mortalidade desses pacientes (MOULAEI et al., 2022). As características preditivas mais significativas encontradas neste estudo também se destacam em outras pesquisas com dados semelhantes. Silva & Neto (2022) identificaram Suporte à Ventilação, Internação na UTI, Idade do Paciente, Desconforto Respiratório e Problemas Renais como os atributos mais importantes para o algoritmo *Random Forest* na criação de modelos preditivos de óbito em pacientes brasileiros com SRAG maiores de 18 anos, independentemente do sexo. Moulaei et al. (2022) chegaram a conclusões semelhantes, destacando Dispneia, Internação na UTI, Suporte à Ventilação e Idade como atributos mais importantes na análise de pacientes hospitalizados por SARS-Cov-2, com idade média de 57 anos e divididos entre 836 homens e 664 mulheres. Por fim, no estudo de Sena (2021), os atributos Idade, Doença Cardiovascular, Saturação de Oxigênio < 95% e Dispneia foram os mais

relevantes para o algoritmo *Random Forest com* amostra de 11.375 pacientes brasileiros da região nordeste com idade superior a 60 anos.

Os resultados demonstram que o ML pode criar modelos de predição confiáveis para auxiliar os profissionais de saúde envolvidos no tratamento de pacientes com SRAG. Entretanto, foram encontrados somente quatro estudos (AZNAR-GIMENO et al., 2021), (WOO et al., 2021), (HU et al., 2021), (KAR et al., 2021) que propuseram uma tecnologia para que o modelo preditivo de óbito por SRAG fosse utilizado na prática por profissionais de saúde. Quando os modelos são disponibilizados através de uma aplicação de software que pode ser utilizada no ambiente hospitalar, esse conhecimento tende a ser mais difundido e utilizado realmente, não ficando restrito somente a literatura. Assim, percebendo essa lacuna na literatura de estudos que busquem aplicar a teoria na prática, foi o desenvolvido um protótipo de aplicação de software de fácil utilização para que profissionais de saúde pudessem utilizar os modelos preditivos no ambiente hospitalar.

Por fim, referente as limitações deste estudo destacam-se: a dificuldade de generalização do uso dos modelos para outros grupos populacionais, como, por exemplo, idosos, uma vez que os modelos foram treinados para classificação de grupos específicos; O desequilíbrio identificado entre as classes de óbito e cura, com um número muito maior de pacientes curados do que falecidos, o que pode afetar a capacidade dos modelos em prever corretamente a classe minoritária (óbito), levando a uma tendência de superestimar a classificação da classe majoritária (cura); A falta de testes de aceitação do protótipo da aplicação pelos profissionais de saúde, uma vez a implementação bem-sucedida de uma nova tecnologia no ambiente clínico pode ser influenciada por uma série de fatores como usabilidade e a integração com sistemas existentes. Além disso, outras técnicas de ML podem ser consideradas para comparação mais abrangente, ainda que o estudo tenha utilizando os algoritmos mais utilizados em estudos do tipo.

5. CONCLUSÃO

O estudo forneceu modelos de predição de óbito baseado nos bancos de dados SRAG do Ministério da Saúde do Brasil para o público materno da região Norte do

Brasil, bem como, um software para a utilização destes modelos com finalidade de auxiliar os profissionais de saúde na identificação precoce de casos graves de SRAG. Dessa forma, considera-se o conhecimento gerado através de modelos preditivos para classificação de pacientes com SRAG tem potencial para fornecer área de saúde o conhecimento prévio acerca de prognósticos de pacientes mais graves e assim alocar melhor os recursos humanos e/ou materiais para o tratamento destes. Além disso, o estudo trouxe a perspectiva de fazer os modelos preditivos serem aplicados na prática através de software para computador. Essa ferramenta pode ser promissora para auxiliar o processo de tomada de decisões clínicas, possibilitando intervenções mais assertivas e personalizadas para prevenir óbitos relacionados à SRAG em populações vulneráveis.

REFERÊNCIAS

ALTINI, N., BRUNETTI, A., MAZZOLENI, S., MONCELLI, F., ZAGARIA, I., et al. **Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters**. *Sensors (Basel, Switzerland)*, *21*(24), 2021. Disponível em: https://doi.org/10.3390/s21248503

AN, C., LIM, H., KIM, D. W., CHANG, J. H., CHOI, Y. J., et al. **Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study**. *Scientific reports*, *10*(1), 2020. Disponível em: https://doi.org/10.1038/s41598-020-75767-2

AZNAR-GIMENO, R., ESTEBAN, L. M., LABATA-LEZAUN, G., DEL-HOYO-ALONSO, R., ABADIA-GALLEGO, D., et al. **A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning Algorithms.** *International journal of environmental research and public health*, *18*(16), 2021. Disponível em: https://doi.org/10.3390/ijerph18168677

ARAÚJO, D. C., VELOSO, A. A., BORGES, K. B. G., CARVALHO, M. D. G**. Prognosing the risk of COVID-19 death through a machine learning-based routine blood panel: A retrospective study in Brazil**. *International journal of medical informatics*. *165*, 104835, 2022. Disponível em: https://doi.org/10.1016/j.ijmedinf.2022.104835

BÁRCENAS, R., FUENTES-GARCÍA, R. **Risk assessment in COVID-19 patients: A multiclass classification approach**. *Informatics in medicine unlocked*, *32*, 101023, 2022. Disponível em: https://doi.org/10.1016/j.imu.2022.101023

BENNETT, T. D., MOFFITT, R. A., HAJAGOS, J. G., AMOR, B., ANAND, A., et al. **National COVID Cohort Collaborative (N3C) Consortium (2021). Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data from the US National COVID Cohort Collaborative**. *JAMA network open*, *4*(7), e2116901, 2021. Disponível em: https://doi.org/10.1001/jamanetworkopen.2021.16901

BOOTH, A. L., ABELS, E., & MCCAFFREY, P. **Development of a prognostic model for mortality in COVID-19 infection using machine learning**. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, *34*(3), 522–531, 2021. Disponível em: https://doi.org/10.1038/s41379-020-00700-x

BOTTRIGHI, A., PENNISI, M., ROVETA, A., MASSARINO, C., CASSINARI, A., et al. **A machine learning approach for predicting high risk hospitalized patients with COVID-19 SARS-Cov-2**. *BMC medical informatics and decision making*, *22*(1), 340, 2022. Disponível em: https://doi.org/10.1186/s12911-022-02076-1

BRASIL. Ministério da Saúde. SRAG 2021 a 2024: banco de dados de Síndrome Respiratória Aguda Grave. OpenDataSUS, 2024. Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024

CARNEIRO, I. C. R., FERONATO, S. G., SILVEIRA, G. F., CHIAVEGATTO Filho, A. D. P., SANTOS, H. G. D. **Clusters of Pregnant Women with Severe Acute Respiratory Syndrome Due to COVID-19: An Unsupervised Learning Approach**. *International journal of environmental research and public health*, 19(20), 2022. Disponível em: https://doi.org/10.3390/ijerph192013522

CARVALHO, A. L. C. **Aplicação de técnicas de aprendizagem de máquina na geração de índices para sistemas de busca**. 2012. 101 f. Tese (Doutorado em Informática) - Universidade Federal do Amazonas, Manaus, 2012. Disponível em: https://tede.ufam.edu.br/handle/tede/4517

CHAPMAN, P., KHABAZZA, T., SHEARER, C. **CRISP-DM 1.0: step by step data mining guide.** SPSS, 2000**.** Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

FERNANDES, F. T., DE OLIVEIRA, T. A., TEIXEIRA, C. E., BATISTA, A. F. M., DALLA COSTA, G., et al. **A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil**. *Scientific reports*, *11*(1), 3343, 2021. Disponível em: https://doi.org/10.1038/s41598-021-82885-y

GAO, Y., CAI, G. Y., FANG, W., LI, H. Y., WANG, S. Y., et al. **Machine learning based early warning system enables accurate mortality risk prediction for COVID-19**. *Nature communications*, *11*(1), 5033, 2020. Disponível em: https://doi.org/10.1038/s41467-020-18684-2

HU, C., LIU, Z., JIANG, Y., SHI, O., ZHANG, X., et al. **Early prediction of mortality risk among patients with severe COVID-19, using machine learning.** *International journal of epidemiology*, *49*(6), 1918–1929, 2020. Disponível em: https://doi.org/10.1093/ije/dyaa171

HELDT, F. S., VIZCAYCHIPI, M. P., PEACOCK, S., CINELLI, M., MCLACHLAN, L., et al. **Early risk assessment for COVID-19 patients from emergency department data using machine learning**. *Scientific reports*, *11*(1), 4200, 2021. Disponível em: https://doi.org/10.1038/s41598-021-83784-y

HE, F., PAGE, J. H., WEINBERG, K. R., MISHRA, A. **The Development and Validation of Simplified Machine Learning Algorithms to Predict Prognosis of Hospitalized Patients With COVID-19: Multicenter, Retrospective Study.** *Journal of medical Internet research*, *24*(1), e31549, 2022. Disponível em: https://doi.org/10.2196/31549

IBGE - Instituto Brasileiro de Geografia e Estatística. **Censo 2022.** Rio de Janeiro: IBGE, 2022.

KAR, S., CHAWLA, R., HARANATH, S. P., RAMASUBBAN, S., RAMAKRISHNAN, N., et al. **Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID).** *Scientific reports*, *11*(1), 12801, 2021. Disponível em: https://doi.org/10.1038/s41598-021-92146-7

KIVRAK, M., GULDOGAN, E., COLAK, C. **Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods**. *Computer methods and programs in biomedicine*, *201*, 105951, 2021. Disponível em: https://doi.org/10.1016/j.cmpb.2021.105951

KUMARAN, M., PHAM, T. M., WANG, K., USMAN, H., NORRIS, C. M., et al. **Predicting the Risk Factors Associated with Severe Outcomes Among COVID-19 Patients-Decision Tree Modeling Approach**. *Frontiers in public health*, *10*, 838514, 2022. Disponível em: https://doi.org/10.3389/fpubh.2022.838514

LEE, C. H.; BANOEI, M. M.; ANSARI, M.; et al. **Using a targeted metabolomics approach to explore differences in ARDS associated with COVID-19 compared to ARDS caused by H1N1 influenza and bacterial pneumonia**. Crit Care., v. 28, p. 63, 2024. doi: 10.1186/s13054-024-04843-0.

LI, Y., HOROWITZ, M. A., LIU, J., CHEW, A., LAN, H., et al. **Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods**. *Frontiers in public health*, *8*, 587937, 2020. Disponível em: https://doi.org/10.3389/fpubh.2020.587937

LI, J., LI, X., HUTCHINSON, J., ASAD, M., LIU, Y., et al. **An ensemble prediction model for COVID-19 mortality risk**. *Biology methods & protocols*, *7*(1), bpac029, 2022. Disponível em: https://doi.org/10.1093/biomethods/bpac029

MAHDAVI, M., CHOUBDAR, H., ZABEH, E., RIEDER, M., SAFAVI-NAEINI, S., et al. **A machine learning based exploration of COVID-19 mortality risk.** *PloS one*, *16*(7), e0252384, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0252384

MOSLEHI, S., MAHJUB, H., FARHADIAN, M., SOLTANIAN, A. R., MAMANI, M. **Interpretable generalized neural additive models for mortality prediction of COVID-19 hospitalized patients in Hamadan**, **Iran**. *BMC medical research methodology*, *22*(1), 339, 2022. Disponível em: https://doi.org/10.1186/s12874-022-01827-y

MOULAEI, K., SHANBEHZADEH, M., MOHAMMADI-TAGHIABAD, Z., KAZEMI-ARPANAHI, H. **Comparing machine learning algorithms for predicting COVID-19 mortality**. *BMC Med Inform Decis Mak* 22, 2., 2022. Disponível em: https://doi.org/10.1186/s12911-021-01742-0

MURRI, R., LENKOWICZ, J., MASCIOCCHI, C., IACOMINI, C., FANTONI, M., et al. **A machine-learning parsimonious multivariable predictive model of mortality risk in patients with Covid-19**. *Scientific reports*, *11*(1), 21136, 2021. Disponível em: https://doi.org/10.1038/s41598-021-99905-6

NAÇÕES UNIDAS. **Transformando o nosso mundo: A Agenda 2030 para o Desenvolvimento Sustentável.** 2015. Disponível em: https://sustainabledevelopment.un.org/post2015/transformingourworld

OCDE. **Estudo da OCDE da Atenção Primária à Saúde no Brasil**. OECD Publishing, Paris, 2021. Disponível em: https://doi.org/10.1787/9bf007f4-pt.

OCDE. **Health Data Governance for the Digital Age: Implementing the OECD Recommendation on Health Data Governance**. OECD Publishing, Paris, 2022. Disponível em: https://doi.org/10.1787/68b60796-en.

PAZOS, J. V., CASTRO, J. O., MOYSÉS, R. P. C., LOPES, F. N. B., FERREIRA, B. O. **A Evolução da Mortalidade Materna e o Impacto da COVID-19 na Região Norte do Brasil: Uma Análise de 2012 a 2021**. Saud Pesq. 16(2):e-11707, 2023. Disponível em: https://doi.org/10.17765/2176-9206.2023v16n2.e11707

POLO, T. C. F., MIOT, H. A. **Aplicações da curva ROC em estudos clínicos e experimentais**. *J Vasc Bras*. 19:e20200186, 2020. Disponível em: https://doi.org/10.1590/1677-5449.200186

R CORE TEAM. **R: A language and environment for statistical computing**. *R Foundation for Statistical Computing*, Vienna, 2024. Disponível em: https://www.R-project.org.

RAJKOMAR, A., OREN, E., CHEN, K. **Scalable and accurate deep learning with electronic health records**. NPJ *Digital Medicine*, 1(1), 1-10, 2021. Disponível em: https://doi.org/10.1038/s41746-018-0029-1

REINA, A. R., BARRERA, J. M., VALDIVIESO, B., GAS, M. E., MATÉ, A., et al. **Machine learning model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical patients**. *Scientific reports*, *12*(1), 5723, 2022. Disponível em: https://doi.org/10.1038/s41598-022-09613-y

SCHÖNING, V., LIAKONI, E., BAUMGARTNER, C., EXADAKTYLOS, A. K., HAUTZ, W. E., et al. **Development and validation of a prognostic COVID-19 severity assessment (COSA) score and machine learning models for patient triage at a tertiary hospital.** *Journal of translational medicine*, *19*(1), 56, 2021. Disponível em: https://doi.org/10.1186/s12967-021-02720-w

SENA, G. R. **Modelos Preditivos de Óbito para Pacientes com COVID-19**. Tese de doutorado apresentada ao Instituto de Medicina Integral Prof. Fernando Figueira (IMIP), 2021. Disponível em: http://higia.imip.org.br/handle/123456789/641?mode=full

SILVA, E. A. D. **Algoritmo genético assistido por surrogate para avaliar e descobrir peptídeos contra o SARS-CoV-2**. 2022. 79 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Uberlândia, Uberlândia, 2022. Disponível em: http://doi.org/10.14393/ufu.di.2022.571.

SILVA, R., SILVA NETO, D. R. DA. **Inteligência artificial e previsão de óbito por Covid-19 no Brasil: uma análise comparativa entre os algoritmos Logistic Regression, Decision Tree e Random Forest**. Saúde em Debate, 46(spe8), 118–129, 2022. Disponível em: https://doi.org/10.1590/0103-11042022E809

SUN, C., HONG, S., SONG, M., LI, H., WANG, Z. **Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning**. *BMC medical informatics and decision making*, *21*(1), 45, 2021. Disponível em: https://doi.org/10.1186/s12911-020-01359-9

VEPA, A., SALEEM, A., RAKHSHAN, K., DANESHKHAH, A., SEDIGHI, T., et al. **Using Machine Learning Algorithms to Develop a Clinical Decision-Making Tool for COVID-19 Inpatients.** *International journal of environmental research and public health*, *18*(12), 6228, 2021. Disponível em: https://doi.org/10.3390/ijerph18126228

WAIKATO. **Weka 3: Machine learning software in java**. The University of Waikato, 2024. Disponível em: https://www.cs.waikato.ac.nz/~ml/weka/index.html

WOO, S. H., RIOS-DIAZ, A. J., KUBEY, A. A., CHENEY-PETERS, D. R., ACKERMANN, L. L., et al. **Development and Validation of a Web-Based Severe COVID-19 Risk Prediction Model.** *The American journal of the medical sciences*, *362*(4), 355–362, 2021. Disponível em: https://doi.org/10.1016/j.amjms.2021.04.001

YADAW, A. S., LI, Y. C., BOSE, S., IYENGAR, R., BUNYAVANICH, S., et al. **Clinical features of COVID-19 mortality: development and validation of a clinical**

prediction model. *The Lancet. Digital health*, *2*(10), e516–e525, 2020. Disponível em: https://doi.org/10.1016/S2589-7500(20)30217-X

YU, L., HALALAU, A., DALAL, B., ABBAS, A. E., IVASCU, F., et al. **Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19**. *PloS one*, *16*(4), e0249285, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0249285

ZAREI, J., JAMSHIDNEZHAD, A., HADDADZADEH SHOUSHTARI, M., MOHAMMAD HADIANFARD, A., et al. **Machine Learning Models to Predict In-Hospital Mortality among Inpatients with COVID-19: Underestimation and Overestimation Bias Analysis in Subgroup Populations**. *Journal of healthcare engineering*, 1644910, 2022. Disponível em: https://doi.org/10.1155/2022/1644910

ZHAO, Y., ZHANG, R., ZHONG, Y., WANG, J., WENG, Z., et al. **Statistical Analysis and Machine Learning Prediction of Disease Outcomes for COVID-19 and Pneumonia Patients.** *Frontiers in cellular and infection microbiology*, *12*, 838749, 2022. Disponível em: https://doi.org/10.3389/fcimb.2022.838749