CAPÍTULO 1

MACHINE LEARNING APLICADA À SAÚDE - ANÁLISE DE DADOS PARA SAÚDE DE CRIANÇAS DE 0 A 3 ANOS

Data de submissão: 24/09/2024 *Data de aceite: 01/10/2024*

Jackson Henrique da Silva Bezerra
Doutorando PGDRA/UFRO. Professor do Instituto Federal de Educação, Ciência e Tecnologia de Rondônia -Campus Ji-Paraná

Fabrício Moraes de Almeida
PhD in Physics (UFC), withpost-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)

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 um modelo preditivo de óbito por Síndrome Respiratória Aguda Grave (SRAG) para crianças de 0 a 3 anos da região Norte do Brasil, através de dados disponibilizados pelo Ministério da Saúde do Brasil. Uma pesquisa aplicada foi realizada através da metodologia CRISP-DM que guiou todo o processo de seleção, processamento, transformação, aplicação dos algoritmos de ML e avaliação do modelo. Os algoritmos *Random Forest*, *Regression Logistic,* K-*Nearest Neighbors* e XGBoost foram utilizados através do software Weka, onde o modelo com o *Random Forest* teve desempenho superior. O modelo foi gerado com validação cruzada e avaliado 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 do modelo foi desenvolvido na linguagem Java para que o conhecimento gerado pelo modelo chegue aos profissionais da área da saúde.

PALAVRA-CHAVE: *Machine Learning* (ML), Banco de dados, Síndrome Respiratória Aguda Grave (SRAG), Modelo Preditivo.

MACHINE LEARNING APPLIED TO HEALTH - DATA ANALYSIS FOR THE HEALTH OF CHILDREN AGED 0 TO 3 YEARS

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. This research aimed to develop a predictive model for death due to Severe Acute Respiratory Syndrome (SARS) for children aged 0 to 3 years in the North region of Brazil, using data provided by the Brazilian Ministry of Health. An applied research was carried out using the CRISP-DM methodology that guided the entire process of selection, processing, transformation, application of ML algorithms and evaluation of the model. The Random Forest, Logistic Regression, K-Nearest Neighbors and XGBoost algorithms were used through the Weka software, where the model with Random Forest had superior performance. The model was generated with cross-validation and 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 model was developed in the Java language so that the knowledge generated by the model reaches healthcare professionals.
KEYWORDS: Machine Learning (ML), Database, Severe Acute Respiratory Syndrome (SARS), Predictive Models

INTRODUÇÃO

O *Machine Learning* (ML) é um conjunto de regras utilizadas para ensinar computadores a "aprenderem" de forma automática padrões e comportamentos a partir de dados de treinamento (SILVA E, 2022), (SENA, 2021). O objetivo principal de um modelo de ML é construir um sistema de computador que aprenda com um banco de dados pré-definido e gere, ao final, um modelo de previsão, classificação ou detecção (PAIXÃO et al., 2022). A aplicação de ML na prática é voltada principalmente para o uso de bases de dados consolidadas com informações heterogêneas, para as quais há uma limitação do uso de técnicas de estatística derivadas (PAIXÃO et al., 2022). Os algoritmos de ML já estão difundidos em diversas áreas, como sistemas bancários para detecção de fraudes (LOPES, 2019), mecanismos de busca na internet (CARVALHO, 2012), sistemas de vigilância em vídeo (MOITINHO & BENICASA, 2023), segurança de dados (HENKE et al., 2018), robótica (RYBCZAK et al., 2024) e, na medicina, para diagnóstico e prognóstico (GROSSARTH et al, 2023). Na área médica, com a digitalização dos prontuários médicos, exames laboratoriais e de imagem, houve um crescimento dos bancos de dados, que são fontes para a aplicação de técnicas de ML, visando a prevenção, diagnóstico precoce e o tratamento das doenças (PAIXÃO et al., 2022).

Os algoritmos de ML podem ser divididos basicamente em duas modalidades: supervisionado e não supervisionado. No aprendizado não supervisionado, o modelo de ML extrai as características dos dados e construiu uma representação sem o conhecimento prévio dos rótulos de cada dado, ou seja, identifica o padrão das informações de classe

heuristicamente. Essa falta de supervisão para o algoritmo pode ser vantajosa, pois permite que o algoritmo analise os padrões que não foram considerados anteriormente (SENA, 2021), (PAIXÃO et al., 2021). No aprendizado supervisionado o modelo de ML tem o conhecimento do rótulo dos dados, ou seja, as amostras estão corretamente definidas. O treinamento é baseado na comparação entre os resultados previstos pelo modelo e os valores reais. Esse processo é repetido até obter um erro mínimo (PAIXÃO et al., 2021). Assim, se o resultado da previsão de um modelo de ML supervisionado for uma categoria, então a tarefa é chamada de classificação, como por exemplo, a predição do conceito de um aluno em uma disciplina em uma das categorias A, B, C, D e E. No entanto, se a predição for um valor numérico específico, então a tarefa é denominada de regressão, como por exemplo, a predição do valor da nota de um aluno em uma disciplina. Algoritmos de ML podem aprender por alterações de parâmetros (como pesos lineares) ou estruturas de aprendizagem (como árvores) (SILVA E, 2022).

Nos últimos anos o 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, como em casos de Síndrome Respiratória Aguda Grave (SRAG) (BEZERRA & ALMEIDA, 2024). 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).

Neste sentido, modelos preditivos desenvolvidos com ML podem identificar pacientes que apresentam maior risco de mortalidade por SRAG, fornecendo suporte para intervenções que visam à redução de mortes (MOULAEI et al., 2022). O conhecimento gerado através do ML pode auxiliar no prognóstico por SRAG, ajudando profissionais da saúde a alocar melhor os recursos materiais e humanos no tratamento de pacientes com maior chance de óbito. O ML ajuda a prever a gravidade e a progressão de doenças como a SRAG, ao analisar grandes conjuntos de dados de registros eletrônicos de saúde, avaliações clínicas e imagens. Esses modelos apoiam a tomada de decisões em várias etapas, desde a triagem até a alta hospitalar, garantindo que recursos como leitos de UTI, ventiladores e equipe médica sejam utilizados de maneira eficiente para priorizar os pacientes mais necessitados e melhorar os resultados gerais dos pacientes (DEBNATH et al., 2020), (VAN DER SCHAAR et al., 2021). Além da geração dos modelos, a criação de mecanismos para disponibilizar os modelos para os profissionais de saúde é importante, conforme verificado nos estudos de Aznar-gimeno et al. (2021), Woo et al. (2021), Hu et al., (2021) e Kar et al. (2021).

Neste contexto, o objetivo do presente trabalho é demonstrar a aplicação prática do ML na área da saúde, através da geração de modelos preditivos de óbito e cura para pacientes com SRAG registros nas bases de dados de SRAG de 2020 e 2021 do Ministério da Saúde disponível no portal openDataSUS. Mantida pela Secretaria de Vigilância em Saúde (SVS), essas bases de dados destacam-se como um importante repositório de dados de pacientes hospitalizados por SRAG. 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). Por fim, 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). Outro fator importante é que todos os registros disponíveis são anonimizados de acordo com as diretrizes da Lei Nº 13.709 de 14 de agosto de 2018 que trata da Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil. Assim, nenhum indivíduo registrado na base de dados pode ser identificado.

METODOLOGIA

A metodologia CRISP-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 CRISP-DM 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. 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. Foram coletados dados de crianças de 0 a 3 anos de idade. A base de dados dispõe de 86 atributos, subdivididos em outros atributos complementares ao atributo original, como por exemplo o atributo “41-Data da vacinação” que possui mais 6 campos adicionais como “Se < 6 meses: a mãe recebeu a vacina” e “Se sim, data”. O openDataSUS disponibiliza os dados na extensão .CSV. Assim, para explorar os dados foi utilizado o MySQL Workbench em conjuntos através da linguagem de programação SQL (*Structured Query Language*).

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. Após análises e testes foram removidos 100 atributos da base de dados de 2021 e 95 da base de 2020, sendo na grande maioria atributos referentes a códigos internos de identificação e datas, pois não são de interesses da pesquisa. Foram criados novos atributos a partir da atributos existentes como por exemplo o atributo NU_IDADE_N (Idade do paciente), DIAS_UTI (Nº de dias na 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. 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. Vale destacar que o Weka também permite a manipulação de atributos e instâncias. 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 dividas 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 estudados 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.

Os algoritmos *Random Forest* (RF), *Logistic Regression* (LR), XGBoost (*Extreme Gradient Boosting*) e KNN (*K-Nearest Neighbors*) foram escolhidos, devido a combinação comum entre estes quatro algoritmos em estudos do gênero, como nos estudos de Moulaei et al. (2022), Schöning et al. (2021) e Kivrak et al. (2021).

Random Forest (Floresta Aleatória)

O *Random Forest* (RF), ou Floresta Aleatória, consiste em um classificador composto por múltiplas árvores, ou seja, uma floresta de decisão (SENA, 2021). Neste algoritmo as árvores de decisão são construídas e representadas através de dois elementos: nós e os ramos que conectam nós. Para tomada de uma decisão, o fluxo começa no nó raiz, navega através dos ramos até chegar a um nó folha. Cada nó da árvore denota um teste de um atributo, e os ramos denotam os possíveis valores que o nó pode assumir. Durante o processo de formação da árvore, também conhecido por treinamento ou aprendizado, leva-se em consideração a homogeneidade das classes para cada divisão do nó. Basicamente, o algoritmo avalia o ganho de informação dos atributos para separação das amostras presentes no conjunto de dados destinado ao treinamento (LIMA et al., 2021). Por exemplo, durante a construção do modelo, três classificadores (árvores) serão construídos e uma nova instância será rotulada por cada classificador. Se os três classificadores cometerem erros distintos, quando o primeiro estiver errado, é possível que o segundo e terceiro sejam correto, de modo que a combinação das hipóteses por votação possa classificar corretamente. Essa técnica é conhecida como *bagging* também conhecida como Agregação de *Bootstrap* e é uma abordagem utilizada em modelo de regressão ou classificação para melhorar a estabilidade e a precisão dos modelos (HU et al., 2021), (SILVA & NETO, 2022)

Uma das principais qualidades da *Random Forest* é a facilidade de medir a importância relativa de cada atributo para a previsão, calculando esse valor automaticamente para cada atributo após o treinamento, quanto maior o valor, mais importante é o atributo. Para isso o algoritmo utiliza a Impureza de Gini (GI) que é um índice para avaliação de atributos na separação de amostras com o mesmo rótulo, ou seja, busca-se a homogeneidade das classes para compor um nó. O índice avalia todos os preditores selecionados aleatoriamente para construir a árvore e escolherá aquele com maior grau de homogeneidade entre as amostras (LIMA et al., 2021). A Figura 1 demonstra o funcionamento de uma floresta aleatória no processo de classificação. Destaca-se que resultado final é obtido pela média (no caso de regressão) ou pela maioria dos votos (no caso de classificação) das previsões de todas as árvores.

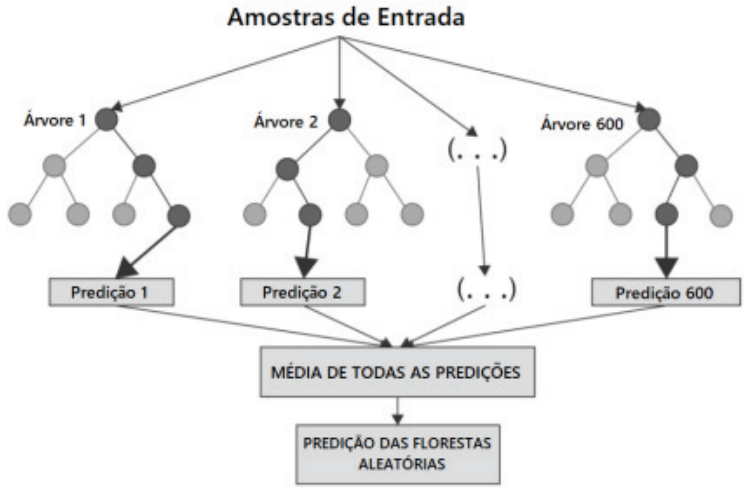

Figura 1 - Exemplo do Esquema da Floresta Aleatória

Fonte: SILVA E (2022)

Logistic Regression (Regressão Logística)

O *Logistic Regression* (RL), ou Regressão Logística, é um modelo linear para classificação. Também é conhecida na literatura como regressão *logit*, classificação de entropia máxima ou classificador log-linear. A regressão logística binária representa os casos de regressão logística em que a variável dependente é binária ou dicotômica, isto é, assume apenas dois valores (SILVA & NETO, 2022). A regressão logística é usada para estimar a associação de uma ou mais variáveis independentes (preditoras) com uma variável dependente binária (resultado). Uma variável binária (ou dicotômica) é uma variável categórica que só pode assumir dois valores ou níveis diferentes, como "morto" ou "vivo" por exemplo. A regressão logística pode ser usada para estimar a probabilidade (ou risco) de um resultado específico, de acordo com os valores das variáveis independentes. Destaca-se que esta probabilidade é dada por valores entre 0 a 1, ou seja, 1 para "vivo" e 0 para "morto" no exemplo citado (SCHOBER e VETTER, 2021).

A fórmula geral para a regressão logística aplica a função sigmoide à combinação linear das variáveis independentes, o que permite transformar a saída linear em uma probabilidade entre 0 e 1. A Equação 1 a seguir apresenta a fórmula da regressão logística para problemas de classificação binária:

$$P(Y=1) = \frac{1}{1+e^{(\beta^0+\beta^1 X^1+\beta^2 X^2+\cdots+\beta_K X_K)}} \qquad (1)$$

onde P ($Y=1$) representa a probalidade do evento de interesse ocorrer, β^0 é o intercepto, β^1, β^2,..., β^k são os coeficientes das variáveis independentes X^1, X^2,..., X^k, e X^1 e e é a base do logaritmo natural, também chamado de número de Euler que corresponde ao número de 2.71 (HOSMER et al., 2013). A Figura 2 apresenta um exemplo de classificação regressão logística.

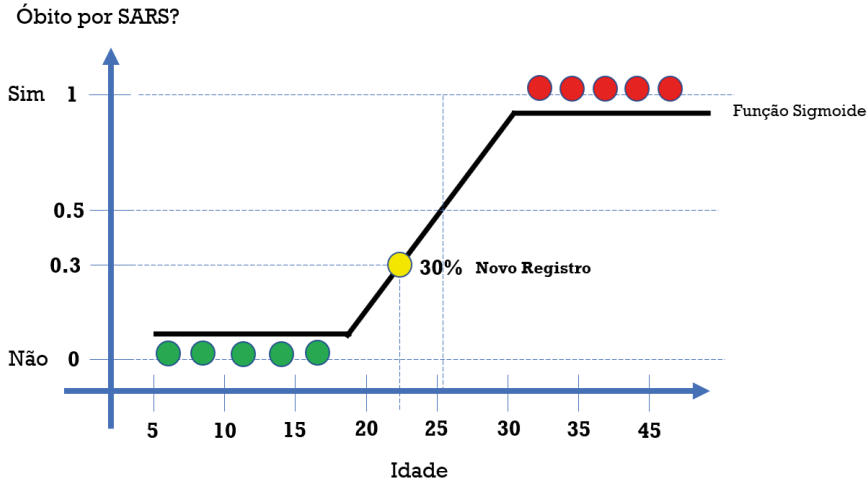

Figura 2 - Exemplo de Modelo de Classificação com Regressão Logística

Fonte: Autor

KNN – *K-Nearest Neighbors* (K Vizinhos Mais Próximos)

O algoritmo *K-Nearest Neighbors* (KNN), ou K-Vizinhos Mais Próximos, é um método de ML amplamente utilizado em problemas de classificação binária devido à sua simplicidade e eficácia, esse algoritmo também suporte classificação não binária e a regressão. O princípio fundamental do KNN é a determinação da classe de um ponto de dados com base nas classes dos seus vizinhos mais próximos em um espaço multidimensional. Em um problema de classificação binária, cada ponto de dados é rotulado com uma das duas classes possíveis, e o objetivo do KNN é prever a classe de novos pontos de dados com base nas observações anteriores. Para implementar o KNN, primeiro é necessário escolher um valor para K, que representa o número de vizinhos a serem considerados (MLADENOVA & VALOVA, 2023). Em seguida, o algoritmo calcula a distância entre o ponto de dados a ser classificado e todos os pontos de dados no conjunto de treinamento. A distância pode ser qualquer medida métrica, como a distância *manhattan*, distância de *minkowski* e a distância euclidiana, esta última sendo uma das mais comuns (SILVA E, 2022). Uma vez calculadas as distâncias, o KNN identifica os K pontos de dados mais próximos e determina a classe predominante entre esses vizinhos (MLADENOVA & VALOVA, 2023). A Figura 3 a seguir exemplifica este processo:

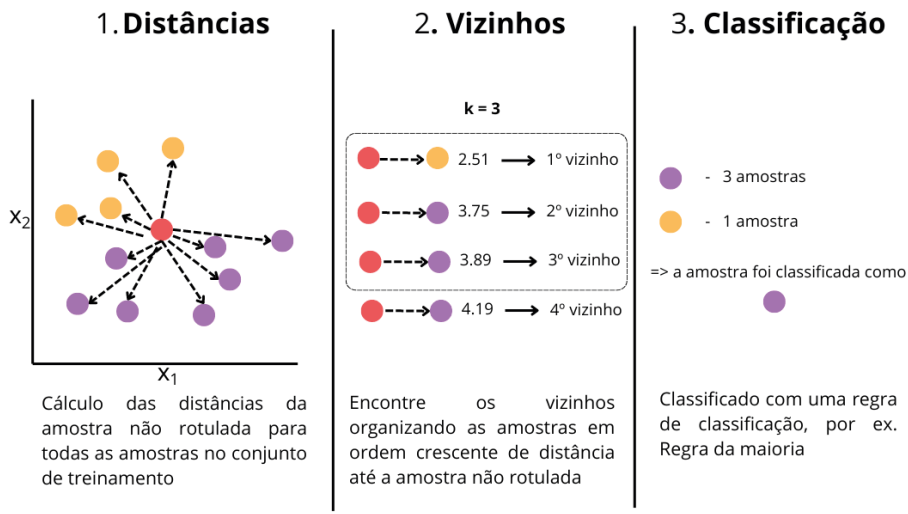

Figura 3 - Etapas da Classificação com o KNN

Fonte: Mladenova & Valova, 2023 (traduzido)

XGBoost – *Extreme Gradient Boosting* (Aumento de Gradiente Extremo)

O XGBoost - *Extreme Gradient Boosting*, ou Aumento de Gradiente Extremo, é um algoritmo de ML baseado em árvores de decisão, projetado para ser altamente eficiente e escalável. Ele utiliza uma abordagem de *boosting*, onde múltiplas árvores são construídas de forma sequencial, cada uma corrigindo os erros da anterior (CHEN & GUESTRIN, 2016). O XGBoost é um algoritmo de árvore de decisão iterativo com múltiplas árvores de decisão. Cada árvore está aprendendo com os resíduos de todas as árvores anteriores. Em vez de adotar a maioria dos resultados de saída de votação no algoritmo *Random Forest*, a saída prevista do XGBoost é a soma de todos os resultados (WANG et al., 2019). O XGBoost cria um modelo que é a soma de várias árvores de decisão a partir da Fórmula 2 a seguir:

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), f_k \in F \quad (2)$$

onde F significa o espaço de árvores de regressão, f_k corresponde a uma árvore, então $f_k(x_i)$ é o resultado da árvore k, e, $\hat{y}_i$ é o valor previsto da i-ésima instância (WANG et al., 2019). O principal objetivo do XGBoost é minimizar uma função de custo regularizada que inclui tanto a função de perda, que mede a discrepância entre as previsões do modelo e os valores reais, quanto termos de regularização que penalizam a complexidade do modelo para evitar o *overfitting*. Essa regularização adicional diferencia o XGBoost de outros algoritmos de *boosting*, tornando-o mais robusto e capaz de generalizar melhor para novos dados (WANG et al., 2019). A função objetivo é dada pela Equação 3 a seguir:

$$Obj(\theta) = L(\theta) + \Omega(\theta)(3)$$

onde L(θ) é a função de perda que mede a diferença entre as previsões (*ŷl i*) e o valores reais (WANG et al., 2019). Por fim, para a classificação binária a função de perda comum é a *log-loss* dada pela Equação 4 a seguir, onde *l*(*yi, ŷi*) é a perda logarítmica entre o valor real e a previsão (*ŷl i*) (WANG et al., 2019).

$$L(\theta) = \sum_{i=1}^{n} l(yi, \hat{y}_i)(4)$$

A Figura 4 a seguir ilustra o processo de funcionamento do algoritmo XGBoost, destacando como ele combina múltiplas árvores de decisão para formar um modelo robusto e preciso.

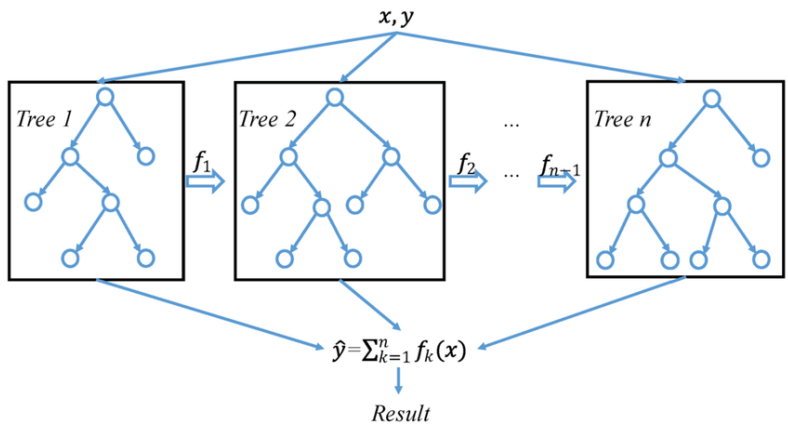

Figura 4 - Arquitetura do XGBoost

Fonte: Wang et al. (2019)

No topo da imagem, as variáveis de entrada *x* (características) e *y* (rótulos) são fornecidas ao modelo. O processo começa com a construção da primeira arvore de decisão (*Tree* 1). A função de predição desta árvore é denotada por f_1. Em seguida a segunda árvore (*Tree* 2) é construída. Ela se baseia nos resíduos ou erros das previsões da primeira árvore e sua função de predição é f_2. Esse processo continua sucessivamente, com cada arvore tentando corrigir os erros das previsões das árvores anteriores. As previsões de todas as árvores são combinadas para formar uma predição final. A fórmula que representa essa combinação é dada pela Equação 1 e a predição final $\hat{y}_i$ é a soma das predições de todas as árvores (WANG et al., 2019).

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

Nota: VP - Verdadeiros Positivos, VN – Verdadeiros Negativos,

FP - Falsos Positivos e FN – Falsos Negativos

Tabela 1 - Modelo da Matriz de Confusão

Fonte: Moulaei et al. (2022)

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)

Nota: VP - Verdadeiros Positivos, VN – Verdadeiros Negativos,

FP - Falsos Positivos e FN – Falsos Negativos

Tabela 2 - Cálculos dos Critérios de Desempenho

Fonte: Moulaei et al. (2022) e Bárcenas & Fuentes-García (2022)

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

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 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)	253	6665	3523	3480	3206
Falsos Positivos (FP)	11	35	113	53	53
Verdadeiros Negativos (TN)	7066	7042	3620	5909	1983
Falsos Negativos (FN)	239	223	1213	558	567
Precisão	0.958	0.995	0.969	0.974	0.019
F1-Score	0.669	0.981	0.842	0.831	0.156
Sensibilidade	0.514	0.968	0.744	0.742	0.227
Especificidade	0.998	0.995	0.970	0.988	0.015
Acurácia	0.966	0.981	0.843	0.930	0.076
AUC-ROC	0.950	0.996	0.946	0.964	0.028

Tabela 3 - Comparativo do Experimento de Balanceamento para Desfecho Óbito Positivo

Fonte: Autor

Na metodologia CRISP-DM a fase de implantação 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.

RESULTADOS

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. Para realizar a manipulação e transformação dos registros da base de dados, além das ferramentas mencionadas anteriormente, foi utilizada a ferramenta MySQL Workbench da Oracle. Esta ferramenta foi selecionada devido à capacidade de programação de scripts SQL, permitindo a automação do processo. Outro ponto considerado foi a capacidade do MySQL lidar com grandes volumes de 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 a base de dados foi disponibilizada no repositório Zenodo no formato ARFF sob link https://zenodo.org/doi/10.5281/zenodo.10879240, contendo 9471 registros. Devido ao registro de dados nulos no atributo classe "evolução" 3.204 registros foram excluídos, assim foram considerados para a geração dos modelos 7569 registros, destes, houve 7077 casos de cura e 492 óbitos. Assim, a versão final das bases de dados definida para a fase de modelagem contou com 40 atributos.

O atributo EVOLUÇÃO foi defino como a classe, com valores de Cura ou Óbito como possíveis. 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 a semente 8 no algoritmo RF, 11 no RL, 8 no KNN e 20 XGBoost. 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.

Métricas do Modelo Preditivo

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

	Predição	
RandomForest	**Óbito (+)**	**Cura (–)**
Óbito (+)	261	231
Cura (–)	9	7068
Logistic Regression	**Óbito (+)**	**Cura (–)**
Óbito (+)	106	386
Cura (–)	83	6994
KNN	**Óbito (+)**	**Cura (–)**
Óbito (+)	327	165
Cura (–)	98	6979
XGBoost	Óbito (+)	Cura (–)
Óbito (+)	139	353
Cura (–)	70	7007

Tabela 4 - Matriz de Confusão

Fonte: Adaptado de Kivrak et al. (2021)

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

Algoritmos	Sensibilidade	Especificidade	Acurácia	Precisão	F1-Score	AUC-ROC
Random Forest	0.530	0.999	0.968	0.967	0.685	0.951
Logistic Regression	0.215	0.988	0.938	0.561	0.311	0.861
KNN	0.665	0.986	0.965	0.769	0.713	0.843
XGBoost	0.283	0.990	0.944	0.665	0.397	0.837

Tabela 5 - Avaliação de Desempenho dos Algoritmos

Fonte: Adaptado de Moulaei et al. (2022)

A Figura 5 a seguir apresenta os gráficos com a Curva ROC e a AUC de cada algoritmo para fins de comparação do desempenho dos modelos gerados, onde verifica-se o desempenho superior do modelo criado com o algoritmo *Random Forest*.

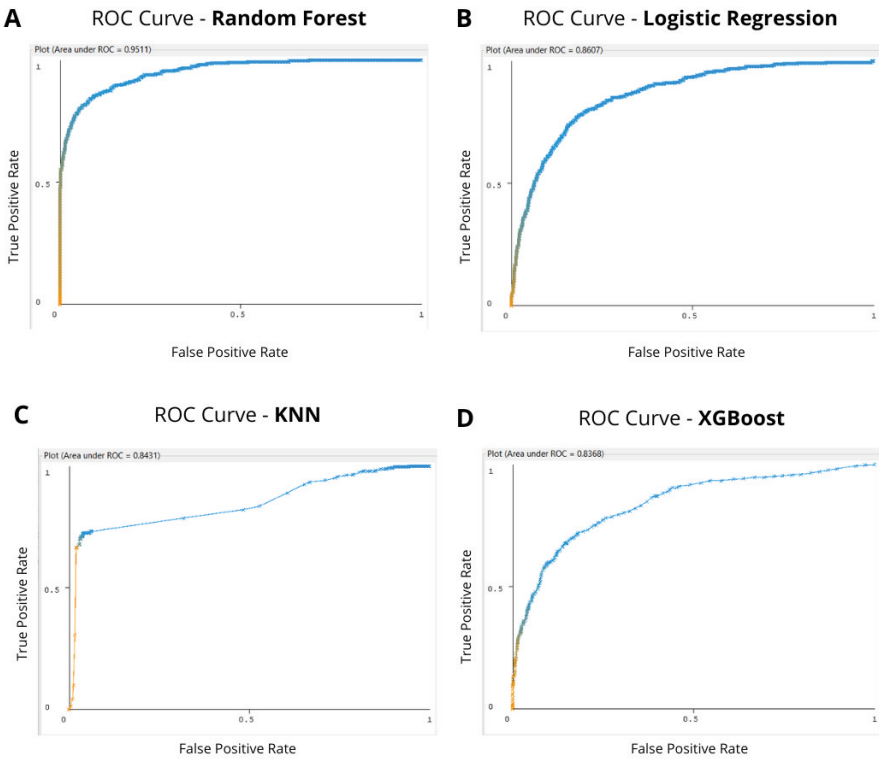

Figura 5 - AUC-ROC dos Modelos

Fonte: Adaptado de Silva & Neto (2022)

A Figura 6 a seguir apresenta o gráfico com os atributos mais importantes considerados pelo modelo com *Random Forest*.

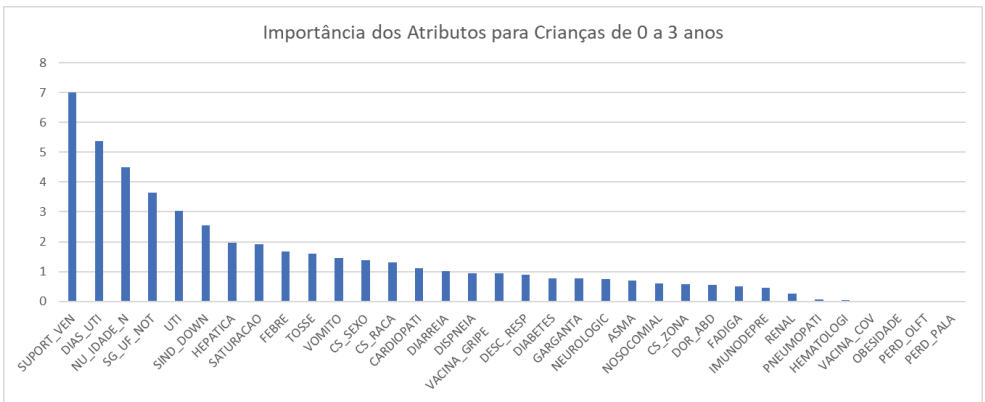

Figura 6 - Gráfico com Índice Gini

Fonte: Adaptado de Zhao et al. (2022)

Simulação de Cenário com o Protótipo da Aplicação

A seguir serão apresentadas imagens das funcionalidades da aplicação, que poderá ser baixada e utilizado por qualquer usuário com acesso a um computador desktop. A Figura 7 a seguir apresenta o menu inicial da aplicação com as opções.

Figura 7 - Menu Inicial da Aplicação

Fonte: Autor

No menu inicial o usuário poderá selecionar três classificadores diferentes, conforme o perfil do paciente que o mesmo deseja classificar. Destaca-se que os públicos foram definidos de acordo com os objetivos da pesquisa. Após a seleção do perfil desejado no menu inicial o sistema irá abrir o classificador para o grupo. A Figura 8 a seguir apresenta a funcionalidade de classificação sem o preenchimento das características do paciente.

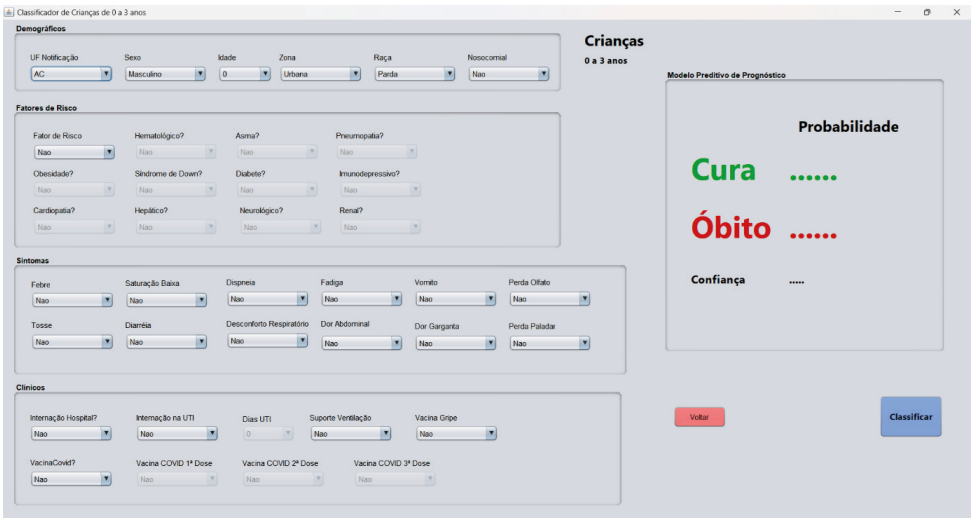

Figura 8 - Classificador

Fonte: Autor

O usuário então poderá informar as características do paciente a ser classificado através de caixas de combinação e então clicar no botão *Classificar*. Ou poderá voltar para o Menu Inicial ao clicar no botão Voltar. 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 e cura para o paciente. A confiança mostrada é a porcentagem de acerto do modelo com base na base de dados de teste, ou seja, a acurácia do modelo. Este valor é calculado pelo aplicativo ao testar o modelo com todos os registros da base de dados, após o treinamento do modelo utilizando validação cruzada.

A Figura 9 a seguir apresenta uma simulação de cenário, com a mudança de estado da funcionalidade após a ação de classificar, onde é apresentado as probabilidades de óbito e cura preditas pelo modelo para a criança conforme as características informadas. Observa-se que neste cenário a criança não possui fatores de risco, sintomas ou foi internada. Assim, a probabilidade de cura predita pelo modelo é alta.

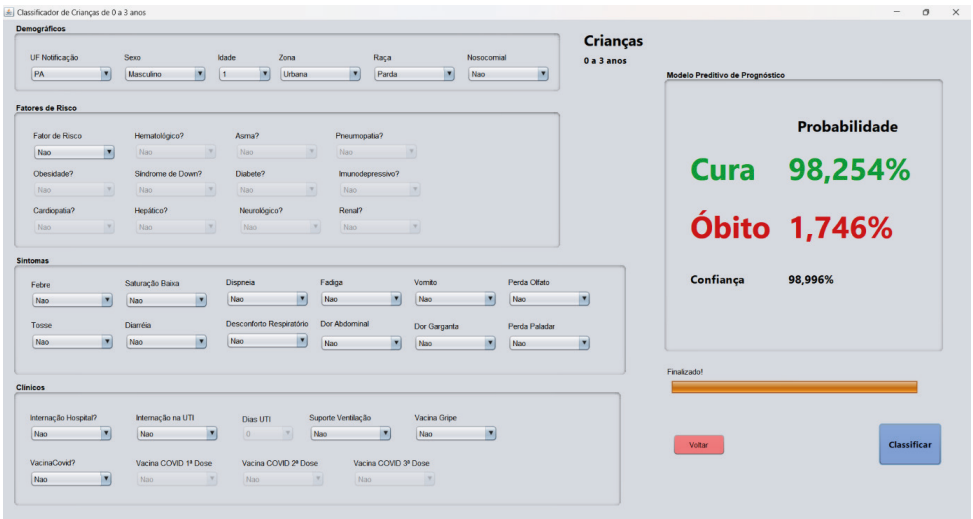

Figura 9 - Classificador de Crianças com Resultado da Simulação de Cenário I

Fonte: Autor

A probabilidade é dada pelo modelo de predição após classificar o paciente com base no conjunto de características informadas na interface antes do clique do botão *Classificar*. No cenário simulado, as probabilidades de Cura e Óbito mudam conforme as características do paciente mudam. A Figura 10 a seguir apresenta a classificação para a criança na simulação de um segundo cenário, onde foi características sobre fatores de riscos e sintomas comuns a SRAG foram inseridos.

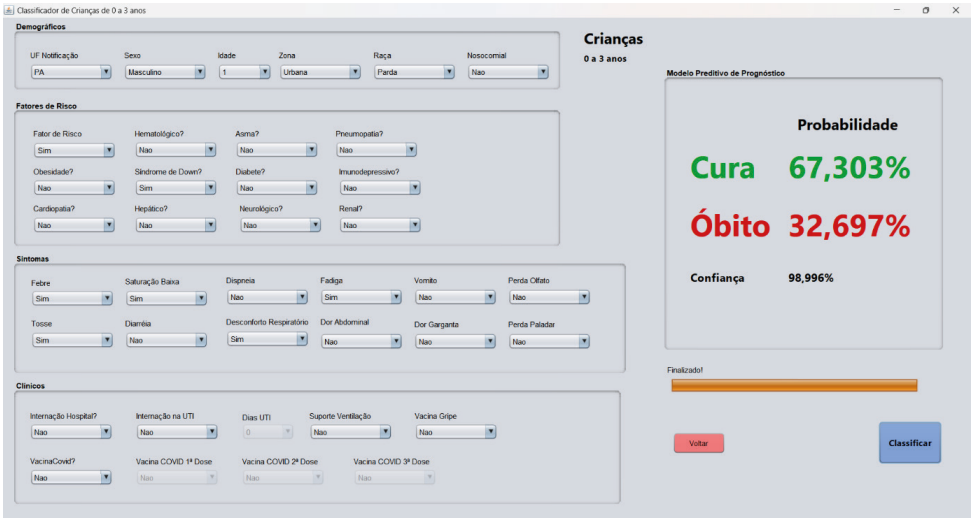

Figura 10 - Classificador de Crianças com Resultado da Simulação de Cenário II

Fonte: Autor

A simulação de cenários apresentados anteriormente demonstra a redução das chances de cura e o aumento das chances de óbito à medida que as características do paciente mudam, evidenciando a capacidade do modelo preditivo de lidar com os fatores relacionados à degradação da saúde do paciente. Por fim, o código fonte da aplicação está depositado no repositório de códigos GitHub (acesso privado) e poderá ser acesso e baixado através do link: https://github.com/jacksonifro/Aplication_Tese_Doutorado.git mediante a solicitação. 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.

DISCUSSÃO

Este estudo representa um avanço importante na criação de modelos de classificação capazes de identificar pacientes com maior risco de óbito por SRAG em grupos de populações vulneráveis da região Norte. 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.

O modelo gerado com o algoritmo *Random Forest* oferece um desempenho robusto e confiável, alcançando uma AUC-ROC de 0.951, sensibilidade de 0.530, especificidade de 0.999, acurácia de 0.968, precisão de 0.967 e F1-Score de 0.685. Esses resultados indicam uma excelente capacidade de distinção entre classes. Embora tenha sido superado pelo KNN na sensibilidade e F1-Score com 0.665 e 0.713 respectivamente, o equilíbrio geral das outras métricas torna sua performance superior. Vale mencionar que apesar da vantagem do KNN na sensibilidade, seu desempenho é inferior ao do *Random Forest* e *Logistic Regression* na AUC-ROC, onde alcançou apenas 0.843. Outro ponto importante é que apesar do desequilibro das classes, verifica-se que não houve grande vantagem do KNN sobre o *Random Forest*, com uma diferença de apenas 0.028 de F1-Score. Neste contexto, o algoritmo *Random Forest* obteve o melhor desempenho geral, sendo o modelo gerado por ele o escolhido para classificação de crianças de 0 a 3 anos no aplicativo.

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.

O algoritmo KNN obteve bom desempenho neste estudo, alcançando AUC-ROC superior a 0.84 nos modelos. Bottrighi et al. (2022) obteve uma AUC-ROC de 0.81 com o algoritmo KNN em um estudo com 824 pacientes italianos utilizando dados demográficos, comorbidades e sintomas, sendo superado pelo algoritmo JRIP. Já os autores Altini et al. (2021) utilizou o algoritmo KNN na comparação outros algoritmos utilizando dados demográficos, clínicos e laboratoriais de 303 pacientes italianos, onde o algoritmo alcançou uma AUC-ROC de 0.778, sendo superado pelo algoritmo *Decision Tree* com AUC-ROC de 0.896.

O algoritmo *Logistic Regression* também alcançou um bom desempenho nos modelos analisados neste estudo, alcançando um AUC-ROC superior a 0.86. Estes resultados se aproximam dos encontrados por outros estudos de modelagem preditiva de óbito, como os achados pelos autores Hu et al. (2021) e Reina et al. (2022) que obtiveram um desempenho na AUC-ROC de 0.895 e 0.871 respectivamente, sendo superior na comparação com outros algoritmos como *Random Forest*, SVM, KNN e MLP. Os autores Murri et al. (2021) e Woo et al. (2021) também chegaram a desempenho superior 0.87 e 0.81 respectivamente na AUC-ROC com o *Logistic Regression*, porém estes autores trabalharam somente com um algoritmo, não comparando com outros estudos.

Também cabe destacar que o algoritmo XGBoost também alcançou um bom desempenho nos modelos analisados neste estudo, com um AUC-ROC superior a 0.83. Estes resultados estão de acordo com os achados de Aznar-Gimeno et al (2021) que obteve uma AUC-ROC de 0.821 com o algoritmo XGBoost em um estudo com 3.623 pacientes espanhóis, superando o algoritmo *Random Forest*. Também de Bárcenas & Fuentes-García (2022) que alcançou uma AUC-ROC de 0.899 com o XGboost no estudo com 220.657 pacientes mexicanos, utilizando dados demográficos, clínicos, sintomas e comorbidades, superando também o *Random Forest*. Assim, como Kar et al. (2021), onde o XGBoost superou o *Random Forest* e o *Regression Logistic* em um estudo com 2.370 pacientes indianos com dados clínicos e laboratoriais. Destaca-se que todos os estudos citados com XGBoost tiveram como objetivo central a criação e comparação de modelos preditivos de óbitos.

Com base nos índices de Gini do modelo com *Random Forest*, verificou-se que as métricas mais importantes para a predição dos modelos nos dados analisados foram os atributos SIND_DOWN (Possui síndrome de Down), HEPATICA (Possui doença hepática) e SATURAÇÃO (Saturação abaixo de 95%), 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.

Por fim, destaca-se que 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, diante da necessidade de 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.

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; E por fim o fato de outras técnicas de ML não terem sido consideradas para uma comparação mais abrangente, ainda que o estudo tenha utilizando os algoritmos mais utilizados em estudos do tipo.

CONCLUSÃO

O estudo forneceu modelos de predição de óbito baseado nos bancos de dados SRAG do Ministério da Saúde do Brasil para o público infantil da região Norte do Brasil, bem como, um software para a utilização destes modelos, afim de auxiliar os profissionais de saúde na identificação precoce de casos graves de SRAG. Considera-se que o conhecimento gerado tem potencial para fornecer aos agentes de saúde conhecimento prévio acerca de prognósticos de pacientes mais graves e assim alocar melhor os recursos humanos e/ou materiais para o tratamento destes. Esta alocação mais eficaz de recursos é importante em regiões de baixa e média renda, onde estes recursos são escassos e periodicamente registram aumento dos índices de casos de SRAG, como por exemplo o período de queimadas na região Norte.

REFERÊNCIAS

ALTINI, N., BRUNETTI, A., MAZZOLENI, S., MONCELLI, F., ZAGARIA, 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

BEZERRA, J. H. S., ALMEIDA, F. M. **DESENVOLVIMENTO DE MODELOS PREDITIVOS COM MACHINE LEARNING - ANÁLISE DE DADOS PARA SAÚDE DE GESTANTES E PUÉRPERAS**. InterSciencePlace, 19, 2024. Disponível em: https://www.interscienceplace.org/index.php/isp/article/view/763

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

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

CHEN, T., GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** *In Proceedings of the* 22nd ACM SIG *KDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). *Association for Computing Machinery*, New York, NY, USA, 785–794, 2016. Disponível em: https://doi.org/10.1145/2939672.2939785

DEBNATH, S., BARNABY, D. P., COPPA, K., MAKHNEVICH, A., KIM, E. J., et al. **Machine learning to assist clinical decision-making during the COVID-19 pandemic**. *Bioelectronic Medicine*, v. 6, p. 14, 2020. Disponível em: https://doi.org/10.1186/s42234-020-00050-8

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

GROSSARTH, S., MOSLEY, D., MADDEN, C., IKE, J., SMITH, I., et al. **Recent Advances in Melanoma Diagnosis and Prognosis Using Machine Learning Methods**. Current Oncology Reports, v. 25, p. 635–645, 2023. Disponível em: https://doi.org/10.1007/s11912-023-01407-3

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, 2021. Disponível em: https://doi.org/10.1093/ije/dyaa171

HENKE, M., SANTOS, C., NUNAN, E., FEITOSA, E., SANTOS, E., et al. **Aprendizagem de Máquina para Segurança em Redes de Computadores: Métodos e Aplicações**. In: XXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC). Anais... Manaus: Universidade Federal do Amazonas, 2018. p. 53-74. Disponível em: https://books-sol.sbc.org.br/index.php/sbc/catalog/download/95/419/690?inline=1

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

HOSMER, D. W., LEMESHOW, S., STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. Wiley, 2013.

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

LIMA, T. P. F., SENA, G. R., NEVES, C. S., VIDAL, S. A., LIMA, J. T. O., et al. **Death risk and the importance of clinical features in elderly people with COVID-19 using the Random Forest Algorithm**. Revista Brasileira de Saúde Materno Infantil, 21(suppl 2), 445–451, 2021. Disponível em: https://doi.org/10.1590/1806-9304202100s200007

LOPES, M. A. **Aplicação de aprendizado de máquina na detecção de fraudes públicas**. 2019. Dissertação (Mestrado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2019. Disponível em: https://doi.org/10.11606/D.12.2020.tde-10022020-174317

MLADENOVA, T., VALOVA, I. **Classification with K-Nearest Neighbors Algorithm: Comparative Analysis between the Manual and Automatic Methods for K-Selection**. *International Journal of Advanced Computer Science and Applications* (IJACSA), 14(4), 2023. Disponível em: http://dx.doi.org/10.14569/IJACSA.2023.0140444

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

MOITINHO, L. C. C., BENICASA, A. X. **Aprendizado de Máquina para o Auxílio à Localização de Pessoas em Ambientes Indoor Monitorados por Câmeras**. In: Concurso de trabalhos de conclusão de curso em sistemas de informação - simpósio brasileiro de sistemas de informação (SBSI), 19, 2023, Maceió/AL. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 71-80. Disponível em: https://doi.org/10.5753/sbsi_estendido.2023.229347

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

PAIXÃO, G. M. M., SANTOS, B. C., ARAÚJO, R. M., RIBEIRO, M.H., MORAES J. L., RIBEIRO, A. L. **Machine Learning in Medicine: Review and Applicability**. Arq Bras Cardiol. Jan;118(1):95-102, 2022. Disponível em: https://doi.org/10.36660/abc.20200596

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

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

RYBCZAK, M., POPOWNIAK N., LAZAROWSKA A. **A Survey of Machine Learning Approaches for Mobile Robot Control**. *Robotics*. 2024; 13(1):12. Disponível em: https://doi.org/10.3390/robotics13010012

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

SCHOBER, P., VETTER, T. R. **Logistic Regression in Medical Research**. *Anesthesia and analgesia*, 132(2), 365–366, 2021. Disponível em: https://doi.org/10.1213/ANE.0000000000005247

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

VAN DER SCHAAR, M., ALAA, A. M., FLOTO, A., GIMSON, A., SCHOLTES, S., et al**. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19**. *Mach Learn*, v. 110, p. 1–14, 2021. Disponível em: https://doi.org/10.1007/s10994-020-05928-x

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

WANG, Y. PAN, Z., ZHENG, J., QIAN, L., MINGTAO, Li. **A hybrid ensemble method for pulsar candidate classification**. *Astrophysics and Space Science*. 364. 2019. Disponível em: https://doi.org/10.1007/s10509-019-3602-4

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., SHOUSHTARI, H. M., HADIANFARD, M. A., CHERAGHI, M., 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