UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

DANIEL MATHEUS KUHN

Aprendizado de Máquina em Tarefas Prognósticas de COVID-19: Avaliação de Algoritmos de Classificação

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Viviane P. Moreira

Porto Alegre
2023

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Kuhn, Daniel Matheus

Aprendizado de Máquina em Tarefas Prognósticas de COVID-19: Avaliação de Algoritmos de Classificação / Daniel Matheus Kuhn. – Porto Alegre: PPGC da UFRGS, 2023.

122 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2023. Orientador: Viviane P. Moreira.

1. Aprendizado de Máquina. 2. Classificação. 3. COVID-19. 4. Informações Admissionais. I. Moreira, Viviane P.. II. Título.

AGRADECIMENTOS

Agradeço à Universidade Federal do Rio Grande do Sul (UFRGS) e ao Programa de Pós-Graduação em Ciência da Computação (PPGC) pela oportunidade de realizar o curso de mestrado. Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

Agradeço à minha orientadora, profa. Dra. Viviane P. Moreira, que sempre prestativa, orientou-me de forma muito atenciosa ao longo da jornada de pós-graduação.

Agradeço também ao prof. Dr. João Luiz Dihl Comba, à profa. Dra. Mariana Recamonde Mendoza e à Ma. Melina Silva de Loreto, que participaram do projeto que originou este trabalho e que contribuíram diretamente para o desenvolvimento do mesmo.

Estendo os meus agradecimentos à todos os meus professores, desde o ensino fundamental até o ensino superior, que me guiaram pelo caminho do conhecimento e foram fundamentais nesta jornada. Agradeço aos meus orientadores de percurso, profa. Dra. Lisiane Cézar de Oliveira, prof. Dr. Cristiano Roberto Cervi (*in memorian*) e ao prof. Dr. Edimar Manica. Ao prof. Edimar, meu agradecimento especial pelo incentivo ímpar.

Agradeço imensamente aos meus pais Adilson Jaime Kuhn e Janete Maier Kuhn, que sempre me deram todo o apoio e suporte, os quais foram de suma importância para a realização dos meus estudos. Também agradeço à Caroline Antunes do Nascimento, minha namorada, que esteve ao meu lado me incentivando em todos os momentos e me auxiliando nas horas de dificuldade.

RESUMO

Modelos preditivos na área da saúde têm sido investigados por inúmeros trabalhos visando o prognóstico e diagnóstico de pacientes. O cenário emergencial de saúde estabelecido pela pandemia da COVID-19 acentuou o interesse em utilizar modelos preditivos para apoiar a tomada de decisão no contexto clínico hospitalar. Esses modelos podem ser empregados nos mais variados desafios enfrentados pelos profissionais de saúde, promovendo um melhor atendimento, otimizando processos de gestão clínica e alocação de recursos. Este trabalho tem como principal objetivo avaliar algoritmos de Aprendizado de Máquina em três tarefas prognósticas a partir de exames disponíveis na admissão hospitalar. As tarefas avaliadas foram: (i) predição de mortalidade; (ii) predição de necessidade de internação em CTI; e (iii) predição de necessidade de recursos de ventilação mecânica invasiva (VMI). Para subsidiar o estudo, foram utilizados registros de 3795 pacientes internados em dois hospitais brasileiros. Avaliamos seis algoritmos de classificação nas três tarefas supracitadas e aplicamos técnicas de visualização de dados, bem como abordagens de explicabilidade para auxiliar na compreensão dos atributos levados em consideração pelos classificadores durante a predição. Além disso, desenvolvemos uma técnica de visualização baseada na abordagem de explicabilidade SHAP, com o intuito de extrair *insights* sobre a relação entre os atributos consideradas relevantes pelos modelos preditivos e suas previsões. Os resultados nas tarefas de classificação para os conjuntos de dados utilizados neste trabalho foram promissores. Os maiores escores de sensibilidade foram atingidos pelo algoritmo de regressão logística. As investigações acerca dos fatores levados em consideração pelos classificadores apontaram, recorrentemente, a idade avançada dos pacientes como o principal fator relacionado à mortalidade. Para a predição de VMI e CTI, atributos relacionados à função respiratória dos pacientes, como baixos índices de saturação de oxigênio e altos índices de pressão parcial de CO2, também foram elencados como relevantes durante a predição. Por fim, a avaliação cruzada utilizando pacientes de diferentes CTI mostrou que os classificadores são sensíveis às características das populações com as quais foram treinados, podendo não generalizar para diferentes unidades hospitalares.

Palavras-chave: Aprendizado de Máquina. Classificação. COVID-19. Informações Admissionais.

Machine Learning in COVID-19 Prognostic Tasks: Evaluation of Classification Algorithms

ABSTRACT

Predictive models in the health area have been investigated by numerous studies aimed at the prognosis and diagnosis of patients. The emergency health scenario established by the COVID-19 pandemic has heightened the interest in using predictive models to support decision-making in the hospital clinical context. These models can be used in the most varied challenges faced by health professionals, promoting better care, optimizing clinical management processes and resource allocation. The main objective of this work is to evaluate Machine Learning algorithms in three prognostic tasks based on exams available at patient's admission. The tasks evaluated were: (i) prediction of hospitalization outcome; (ii) prediction of need for ICU admission; and (iii) prediction of need for invasive mechanical ventilation (IMV). To support the study, records of 3795 patients admitted to two Brazilian hospitals were used. We evaluated six classification algorithms in the three aforementioned tasks and applied data visualization techniques, as well as explicability approaches to assist in understanding the attributes taken into account by the classifiers during prediction. In addition, we developed a visualization technique based on the SHAP explainability approach in order to extract insights into the relationship between the variables considered by the predictive models and their predictions. The results in the classification tasks for the datasets used in this work were promising. The highest sensitivity scores were achieved by the logistic regression algorithm. Investigations into the factors taken into account by the classifiers have repeatedly pointed to the advanced age of patients as the main factor related to mortality. For the prediction of IMV and ICU, attributes related to the respiratory function of patients, such as low levels of oxygen saturation and high levels of CO2 partial pressure, were also listed as relevant during the prediction. Finally, the cross-assessment using patients from different ICUs showed that the classifiers are sensitive to the characteristics of the populations with which they were trained and may not generalize to different hospital units.

Keywords: Machine Learning, Classification, COVID-19, Admission Information.

LISTA DE ABREVIATURAS E SIGLAS

AB *AdaBoost*

AM Aprendizado de Máquina

AUPRC *Area Under the Precision-Recall Curve*

AUROC *Area Under the Receiver Operating Characteristic Curve*

CMS *Centers for Medicare & Medicaid Services*

CNN Redes Neurais Convolucionais

CTI Centro de Terapia Intensiva

DL *Deep Learning*

DPOC Doença Pulmonar Obstrutiva Crônica

Esp Especificidade

FN Falso Negativo

FP Falso Positivo

F1 Medida-F1

HCPA Hospital de Clínicas de Porto Alegre

HMV Hospital Moinhos de Vento

KNN *K-Nearest Neighbors*

LDH Lactato Desidrogenase

LIME *Local Interpretable Model-Agnostic Explanations*

LR *Logistic Regression*

MEWS *Modified Early Warning Score*

MICE *Multivariate Imputation By Chained Equations*

MLP *Multi Layer Perceptron*

NEWS *National Early Warning Score*

PCR Proteína C-reativa

PR *Precision-Recall*

PROBAST *Prediction Model Risk Of Bias Assessment Tool*

REMS *Rapid Acute Physiology Score*

ReLU *Rectified Linear Unit*

RF *Random Forest*

ROC *Receiver Operating Characteristic*

RT-PCR *Reverse Transcription Polymerase Chain Reaction*

SHAP *SHapley Additive exPlanations*

SA Seleção de Atributo

Sen Sensibilidade

SOFA *Sequential Organ Failure Assessment*

SVM *Support Vector Machine*

VMI Ventilação Mecânica Invasiva

VN Verdadeiro Negativo

VP Verdadeiro Positivo

VPN Valor Preditivo Negativo

VPP Valor Preditivo Positivo

XGB *XGBoost*

4C *4C Mortality Score*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento de modelos preditivos na área da saúde para o diagnóstico e prognóstico de pacientes tem sido recentemente investigado por inúmeros trabalhos. O aumento no interesse em modelos clínicos preditivos pode ser explicado por fatores como a crescente capacidade de coleta de dados, o impacto positivo na tomada de decisão e no atendimento clínico, bem como, os grandes avanços em Aprendizado de Máquina (AM). Modelos preditivos visam auxiliar nos diversos desafios enfrentados pelos profissionais de saúde no tratamento de pacientes, promovendo um melhor atendimento, otimizando processos de gestão clínica e alocação de recursos hospitalares.

A pandemia da COVID-19 foi uma emergência crítica de saúde pública em todo o mundo, dada a necessidade repentina de um grande número de leitos hospitalares e o aumento da demanda por equipamentos médicos e profissionais da saúde. Esse cenário intensificou o interesse em utilizar modelos preditivos para apoiar a tomada de decisão para diagnóstico e prognóstico no contexto da COVID-19. Conforme discutido em trabalhos anteriores (ALBALLA; AL-TURAIKI, 2021), um grande número de artigos relacionados à COVID-19 foi publicado em um curto período, incluindo artigos propondo aplicações de AM para previsão de prognóstico da COVID-19.

Dentre os modelos de previsão prognóstica, podemos citar modelos preditivos que visam antecipar o desfecho do paciente (por exemplo, alta hospitalar ou óbito), a gravidade da doença, o tempo de internação, a evolução do quadro clínico do paciente durante a internação, a necessidade de internação em Centros de Terapia Intensiva (CTI) ou intervenções como intubação e o desenvolvimento de complicações (por exemplo, problemas cardíacos, trombose, síndrome respiratória aguda, entre outros) (WYNANTS et al., 2020; SUBUDHI; VERMA; PATEL, 2020).

Especialmente no que compete as demandas ocasionadas pelo COVID-19, três modelos de predição prognóstica de acentuada importância relacionam-se com a predição de mortalidade, necessidade de leitos de CTI e predição de necessidade de VMI. Neste trabalho abordamos estas três tarefas de predição. A predição de VMI é de acentuada importância no contexto da COVID-19, pois a doença compromete diretamente o sistema respiratório dos infectados. Da mesma forma, as predições de CTI e mortalidade são cruciais para o gerenciamento de recursos, visto que a disponibilidade de leitos de CTI é limitada.

Embora estudos anteriores tenham mostrado que algoritmos de AM alcançaram

resultados promissores nas tarefas supracitadas (KNIGHT et al., 2020; ALBALLA; AL-TURAIKI, 2021), a falta de explicabilidade quanto aos atributos levadas em conta pelos modelos preditivos tem sido apontada como um fator limitante para sua aplicação na rotina clínica (KNIGHT et al., 2020).

Neste trabalho, avaliamos algoritmos de AM nas tarefas de predição de mortalidade, necessidade de internação em CTI e, necessidade de VMI para pacientes com COVID-19 e exploramos abordagens de explicabilidade para auxiliar na compreensão dos fatores considerados relevantes pelos modelos preditivos durante as predições. Além disso, desenvolvemos uma técnica de visualização baseada em valores SHAP (LUNDBERG; LEE, 2017) que permite extrair *insights* sobre a relação entre os atributos considerados relevantes pelos modelos preditivos e suas previsões corretas e incorretas. Para a avaliação dos algoritmos de AM, realizamos um estudo retrospectivo de pacientes diagnosticados com COVID-19, analisando dados coletados durante a internação em dois hospitais brasileiros. Nossas principais contribuições podem ser resumidas como:

- avaliação de seis algoritmos de AM para as tarefas de predição de mortalidade, predição de internação em CTI e predição de necessidade de VMI;
- proposta de uma técnica de visualização baseada em valores SHAP para auxiliar na explicabilidade; e
- experimentos com dados reais de dois hospitais.

Nossos achados mostraram que, em geral, os pacientes mais velhos foram mais suscetíveis a um prognóstico desfavorável, principalmente quando considerada a tarefa de predição de mortalidade. Para as tarefas de predição de CTI e VMI, os atributos recorrentemente elencados como de maior contribuição para as predições, estavam relacionados à função respiratória do paciente. Histórico de doença cardíaca também foi um fator relacionado ao prognóstico desfavorável. Entre os algoritmos avaliados, *logistic regression* atingiu os maiores índices de sensibilidade, atingindo escores de até 0,84, 0,66 e 0,89 para as tarefas de predição de mortalidade, internação em CTI e necessidade de VMI, respectivamente. Ainda, classificadores treinados para a predição de mortalidade e CTI utilizando apenas a idade, bem como os classificadores treinados apenas com o atributo *saturação de O2* para predição de necessidade de VMI, atingiram níveis de sensibilidade semelhantes, ou até mesmo superiores aos classificadores treinados com múltiplos atributos. Entretanto, esses níveis de sensibilidade foram geralmente atingidos em detrimento dos níveis de especificidade.

O restante deste trabalho está organizado como segue: O Capítulo 2 apresenta os principais conceitos relacionados ao nosso trabalho. O Capítulo 3 discorre acerca de trabalhos que abordam as tarefas de predição de mortalidade, VMI e CTI. O Capítulo 4 apresenta a metodologia adotada neste trabalho para o desenvolvimento e avaliação dos preditores nas tarefas supracitadas. O Capítulo 5 contempla os experimentos realizados, a configuração experimental e apresenta os resultados. Por fim, as conclusões de nosso trabalho são apresentadas no Capítulo 6.

2 BACKGROUND

Este capítulo aborda os principais conceitos necessários ao entendimento deste trabalho.

2.1 Terminologias

A seguir são apresentadas terminologias adotadas nesse trabalho e os seus significados.

- **Valores dos atributos** - durante a discussão dos resultados, utilizamos expressões como *valores baixos*, *valores medianos* e *valores altos* para nos referirmos aos valores dos atributos. É importante ressaltar que essas expressões são empregadas para caracterizar e/ou sumarizar as informações com base nos valores observados para os atributos e não se baseiam em análises relativas aos limiares de referência laboratoriais.
- **Hospitalização índice e readmissão** - Na área da saúde, a *hospitalização índice* é a denominação dada para a primeira vez, entre uma série de hospitalizações, que um paciente é admitido em um hospital ou conjunto de hospitais devido a uma condição específica ou diagnóstico. Se o paciente for admitido novamente devido ao mesmo diagnóstico, as próximas hospitalizações são denominadas readmissões.
  Entre as etapas de pré-processamento aplicadas no presente trabalho, está contemplado o processamento dos registros de pacientes readmitidos. A Seção 4.1.1 apresenta mais detalhes sobre como esta etapa de pré-processamento foi aplicada neste trabalho.

2.2 Algoritmos de Classificação

Os algoritmos de classificação figuram entre os maiores esforços dentro da área de AM. A tarefa de classificação pode ser definida como o processo de atribuir instâncias (representadas por um conjunto de atributos) a classes predeterminadas. Através deste processo, cria-se um modelo que consegue prever a classe da instância a partir de seus atributos. As tarefas de classificação apresentam-se nos mais variados domínios, como saúde, finanças, agricultura, educação, geologia, *etc*.

Em nosso trabalho, avaliamos seis algoritmos de classificação. Cada algoritmo baseia-se em um conjunto de pressupostos para a indução do processo que é conhecido como aprendizado. Esta seção tem por objetivo introduzir os principais conceitos sobre os quais baseiam-se os algoritmos avaliados em nosso trabalho. É importante ressaltar que embora a maioria dos algoritmos de AM mencionados tenham sido projetados tanto para tarefas de regressão quanto de classificação, vamos apresentá-los sob a perspectiva das tarefas de classificação.

2.2.1 *Logistic Regression*

Logistic Regression (LR) é um dos algoritmos mais utilizados para tarefas de classificação quando a variável dependente (variável de interesse) é discreta. Usualmente, a variável dependente assume dois valores possíveis (regressão logística binomial), mas extensões do algoritmo permitem trabalhar com a variável dependente assumindo mais de dois valores (regressão logística multinomial) (WRIGHT, 1995). Na área de AM, as versões binomial e multinomial são destinadas respectivamente para tarefas de classificação binária e multi-classe.

Assim como outros modelos lineares, os modelos LR tornaram-se imprescindíveis em análises de dados onde há preocupação em descrever a relação entre as variáveis independentes e a variável dependente.

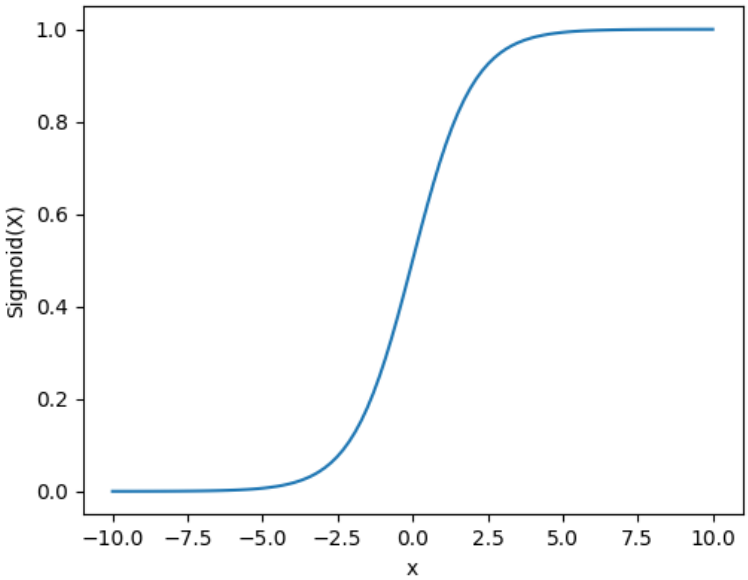


Figura 2.1 – Função logística
Fonte: O Autor

LR faz uso da função logística para o processo de modelagem. A Figura 2.1 apresenta um exemplo da curva obtida a partir da função logística. É possível observar que a curva varia entre o intervalo $[0 - 1]$. Esses valores referem-se à probabilidade de ocorrência do evento dadas as variáveis independentes.

Em termos matemáticos, a função logística é dada por:

$$y' = \frac{1}{1 + e^{-z}}, \tag{2.1}$$

onde y' é a saída do modelo de regressão logística para uma determinada instância. Ou seja, um número entre 0 e 1 referente à probabilidade predita; z é o resultado do modelo linear ($z = b + w_1x_1 + w_2x_2 +, ..., +w_Nx_N$), sendo que b refere-se ao intercepto e w são os pesos – ambos parâmetros aprendidos durante o procedimento de modelagem – e x são os valores de cada um dos atributos da instância.

O processo de modelagem é realizado a partir da maximização de verossimilhança. Devido às suas propriedades interpretativas, os modelos de LR têm sido amplamente utilizados em áreas sensíveis, como é o caso da área da saúde.

2.2.2 Árvore de decisão

Árvore de decisão é um algoritmo amplamente utilizado para tarefas de AM. Uma série de algoritmos têm como base o algoritmo de árvore de decisão, como é o caso das florestas aleatórias (*Random Forest - RF*), *Gradient Boosting Tree*, *AdaBoost*, entre outros.

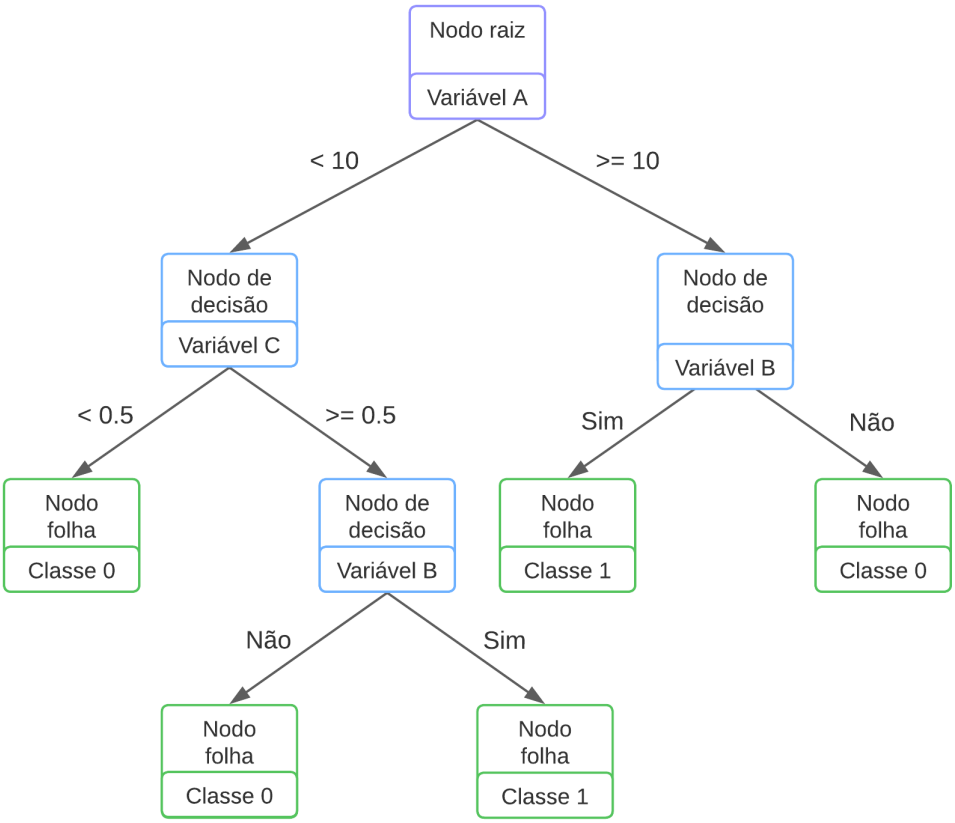


Figura 2.2 – Árvore de decisão
Fonte: O Autor

Árvores de decisão são constituídas por três componentes básicos: (i) nodos raiz, (ii) nodos de decisão e (iii) nodos folha. O nodo raiz é o primeiro nodo de decisão da árvore. Os nodos internos são os nodos de decisão. São os nodos de decisão que especificam o caminho a ser percorrido com base em critérios definidos durante o processo de geração da árvore. Por sua vez, os nodos folha representam as classes a serem preditas.

A Figura 2.2 apresenta uma árvore de decisão para classificação binária, composta por três variáveis: A, B e C. As variáveis A e C são numéricas, enquanto que a variável B é categórica. Após o processo de treinamento, que consiste na construção da árvore, uma nova instância I pode ser classificada com base nos valores das variáveis A, B, C. A classificação é realizada seguindo a estrutura da árvore com base nos critérios estabelecidos para cada variável, até que um nodo folha (classe) seja alcançado.

O treinamento de uma árvore de decisão consiste na construção da árvore com base em um conjunto de dados de treinamento. É durante esse procedimento que as variáveis são selecionadas para dar origem aos nodos de decisão. O critério de seleção das variáveis se dá a partir de uma métrica de pureza da variável em prever as classes. Existem diferentes métricas de pureza, sendo que as mais populares são *Gini index* e *Information Gain*.

Uma característica das árvores de decisão é que suas classificações são facilmente interpretáveis, visto que a classificação se dá a partir da transição entre os nodos da árvore, com base nos valores das variáveis. Entre os algoritmos avaliados em nosso trabalho o J48 é um exemplo de algoritmo de árvore de decisão.

2.2.3 *Random Forest*

Compreendido entre os algoritmos mais populares de AM, o *random forest* é um algoritmo do tipo *ensemble*. Uma das vantagens do *random forest* frente às árvores de decisão refere-se a baixa variância durante as predições (BREIMAN, 2001). Para isso, o algoritmo gera um número A de árvores a partir do conjunto de dados de treino. A Figura 2.3 apresenta o esquema de geração de uma *random forest*. As árvores são geradas com base na técnica estatística de amostragem *bootstrap*. Assim, conjuntos de instâncias diferentes, amostrados a partir do conjunto de dados original, são utilizados para a geração das diferentes árvores.

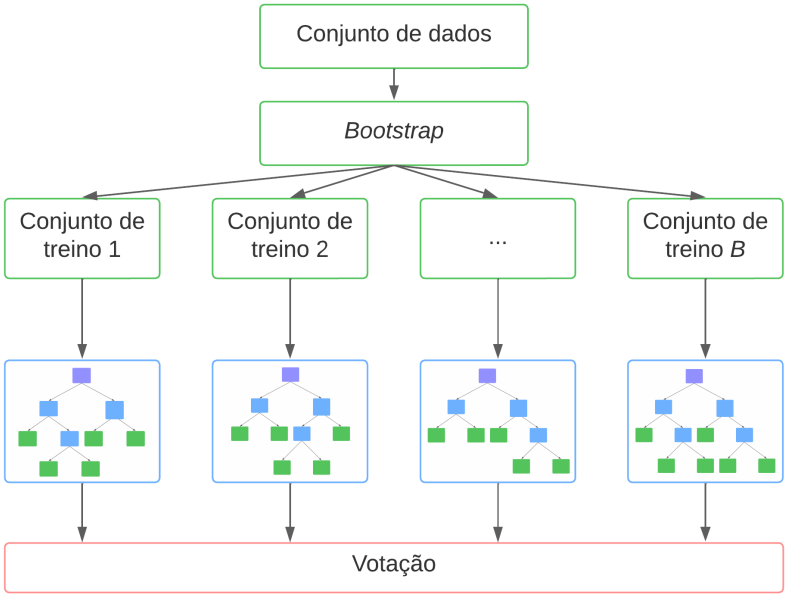


Figura 2.3 – *Random Forest*
Fonte: O Autor

No momento da predição, cada árvore retorna a classe predita de forma isolada e então, as predições retornadas pelas árvores são combinadas. Embora existam diferentes formas de combinar as predições, como votação majoritária ponderada e contagem de borda, uma forma recorrentemente utilizada para combinar as predições consiste na vota-

ção majoritária, onde a classe apontada pela maioria das árvores é escolhida como a classe predita. Neste processo de votação, cada árvore exerce o mesmo poder de voto. Dessa forma, a classe predita contempla a opinião majoritária entre um conjunto de modelos preditores. Esta etapa é representada na Figura 2.3 pelo retângulo denominado *votação*.

2.2.4 *Multi Layer Perceptron*

Multi Layer Perceptron (MLP) é uma das arquiteturas de redes neurais mais populares. A MLP é constituída por múltiplas camadas de *perceptrons* (BISHOP et al., 1995).

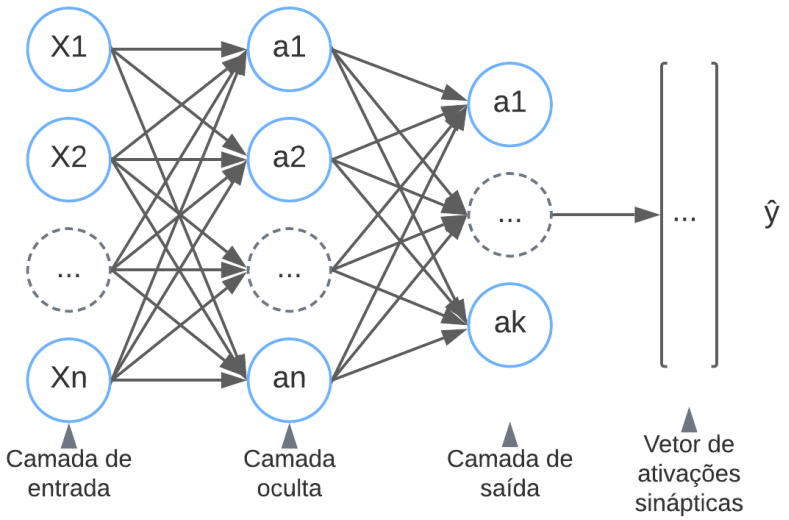


Figura 2.4 – Rede neural *Multi Layer Perceptron*
Fonte: O Autor

A Figura 2.4 apresenta a estrutura de uma rede MLP com uma camada oculta. A primeira camada refere-se às variáveis de entrada que são propagadas para todos os neurônios da camada oculta. Cada neurônio da camada oculta recebe como entrada os valores da camada anterior. Os valores recebidos como entrada são representados pelas setas (arestas).

Cada neurônio realiza a ponderação dos valores de entrada através de uma combinação linear. Os parâmetros de ponderação da combinação linear fazem parte do conjunto de parâmetros aprendidos durante o processo de treinamento. Na sequência, cada neurônio possui uma função não linear de ativação que inibe ou excita a saída do neurônio. Algumas das funções de ativação mais recorrentes são *sigmoid*, *hyperbolic tangent*, e *Rectified Linear Unit* (ReLU). Por fim, a última camada usualmente é composta por uma

função de ativação *sigmoid* ou *softmax* (BISHOP et al., 1995).

O treinamento é realizado a partir de um processo iterativo, onde busca-se minimizar o erro calculado por uma função de perda - que calcula a diferença do vetor estimado pela rede e o vetor com a distribuição empírica. A minimização do erro é alcançada através do uso do algoritmo *backpropagation* que calcula o gradiente da função de perda e realiza a atualização dos parâmetros da rede neural. Devido às funções de ativação não lineares, as redes MLP são capazes de criar fronteiras de decisão complexas, podendo identificar classes que *a priori* não são linearmente separáveis (BISHOP et al., 1995).

2.2.5 *Boosting*

Boosting é um método *ensemble* que combina um conjunto de modelos preditivos *fracos* para compor um modelo preditivo *forte*. Por modelos preditivos fracos, deve-se entender aqueles modelos que podem ter poder preditivo não muito superior à predição aleatória (POLIKAR, 2006). Originalmente o *boosting* foi proposto para utilizar como base classificadores com alto viés (os ditos classificadores fracos), como é o caso do *AdaBoost*, um dos primeiros métodos de *boosting* e de reconhecido sucesso. Com a evolução da área, novos métodos *ensemble* baseados em *boosting* foram propostos, como o *XGBoost*.

As próximas seções apresentam as principais características dos algoritmos de *boosting* avaliados nesse trabalho.

2.2.5.1 AdaBoost

Proposto por Freund and Schapire (1997), *AdaBoost* estabeleceu-se entre os algoritmos de *boosting* mais conhecidos. *AdaBoost* treina iterativamente uma sequência de modelos preditivos fracos com o objetivo de compor um modelo preditivo forte que apresente bom desempenho. Cada modelo da sequência enfatiza os erros cometidos pelo modelo antecessor. Para enfatizar os erros cometidos pelo modelo antecessor, um peso é atribuído para cada instância.

Na primeira iteração, todas as instâncias recebem o mesmo peso para o treinamento do primeiro modelo preditivo. Após o treinamento, as instâncias têm seus pesos atualizados. Instâncias preditas incorretamente têm o seu peso acrescido, enquanto que as instâncias preditas corretamente têm seu peso reduzido. O maior peso atribuído às

instâncias classificadas incorretamente força o próximo modelo preditivo a focar nessas instâncias mais difíceis. Após a atualização dos pesos das instâncias, inicia-se a segunda iteração, onde um próximo modelo preditivo é treinado e uma nova atualização é realizada com base nos seus acertos e erros (FREUND; SCHAPIRE; ABE, 1999). O procedimento se repete até que o limite previamente estabelecido de iterações seja atingido, ou até que atinja-se um critério relacionado à redução do erro.

Um peso de votação é atribuído para cada modelo preditivo com base no seu desempenho de classificação obtido durante a etapa de treinamento (não confundir com o peso atribuído às instâncias). Na etapa de predição, cada modelo atribui a classe que julga mais adequada para a instância a ser classificada. A classe final a ser retornada é eleita levando-se em consideração o peso de votação atribuído a cada modelo preditivo. Assim, os modelos com maior peso, exercem maior influência durante a votação da classe final a ser predita.

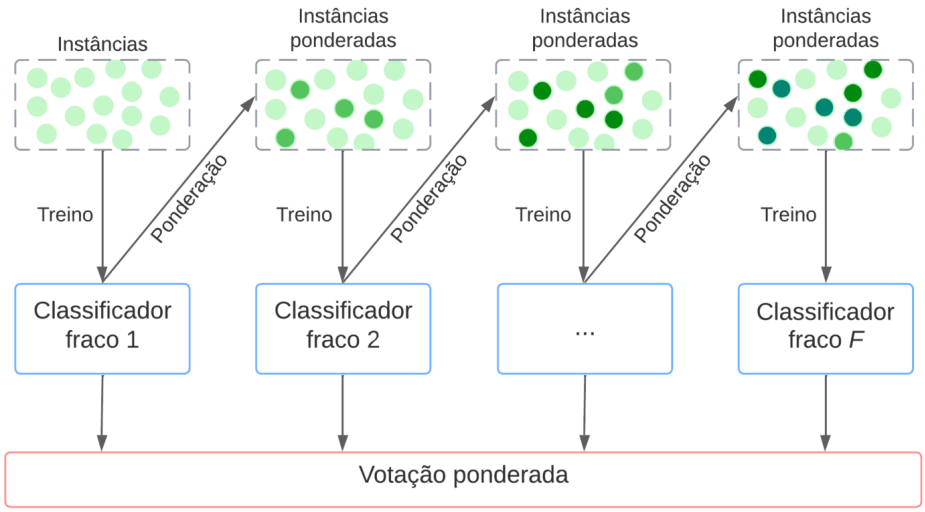


Figura 2.5 – Estrutura sequencial de treinamento do *AdaBoost*
Fonte: O Autor

A Figura 2.5 apresenta a estrutura sequencial do algoritmo *AdaBoost*, bem como o seu esquema de votação ponderada. É importante observar que embora outros algoritmos possam ser utilizados, usualmente os modelos *fracos* utilizados consistem em árvores com um único nodo de decisão, denominadas *decision-stumps*. Neste caso, cada *decision-stump* é gerada a partir de uma única variável entre as variáveis disponíveis que compõem as instâncias.

2.2.5.2 XGBoost

XGBoost (CHEN; GUESTRIN, 2016) é uma implementação que deriva da técnica *Gradient Boosting Tree* (GBT). Assim como no algoritmo *AdaBoost*, nos algoritmos de GBT, os modelos preditivos são construídos de forma iterativa levando-se em conta os erros do preditor antecessor. Em suma, os modelos preditivos são árvores de decisão construídas de forma a minimizar os resíduos das predições realizadas pela árvore antecessora. O processo de treinamento ocorre até que o número predeterminado de árvores construídas seja atingido ou até que não ocorra significativa redução do erro.

XGBoost têm sido extensivamente adotado devido a sua eficiência e desempenho atingido em diversas tarefas de AM nos mais variados domínios.

2.3 Desbalanceamento de classes

Em abordagens de classificação supervisionadas, o desbalanceamento de classes refere-se à desigualdade das proporções de instâncias de cada uma das classes (JAPKOWICZ; STEPHEN, 2002). O desbalanceamento de classes é um fenômeno amplamente conhecido em AM, sendo um objeto de estudo extensivamente abordado (HE; GARCIA, 2009; KRAWCZYK, 2016; JOHNSON; KHOSHGOFTAAR, 2019). Em cenários do mundo real, várias tarefas envolvem o trabalho com classes desbalanceadas (LONGADGE; DONGRE, 2013; JAPKOWICZ; STEPHEN, 2002). Esse comportamento também é observado em tarefas de diagnóstico e prognóstico de pacientes. Por exemplo, em tarefas de diagnóstico de pacientes, a ocorrência do evento (existência da doença) é normalmente menos frequente do que a não ocorrência do evento (ausência da doença).

Entretanto, a maioria dos algoritmos de classificação assumem que os eventos são uniformemente distribuídos entre as classes. Como consequência, ao serem treinados com conjuntos de dados desbalanceados sem os cuidados necessários, estes algoritmos tendem a apresentar resultados com *viés*, priorizando a predição da classe majoritária (HE; GARCIA, 2009; THAI-NGHE; GANTNER; SCHMIDT-THIEME, 2010).

Diversas abordagens têm sido propostas para aprimorar a eficácia dos classificadores em cenários de dados desbalanceados. Entre as principais abordagens, pode-se citar técnicas de *undersampling*, *oversampling* e a aplicação de matrizes de custo.

A aplicação de matrizes de custo tem como objetivo ponderar os diferentes tipos de erros cometidos pelos modelos preditivos (falso positivo e falso negativo). A Tabela 2.1

Tabela 2.1 – Exemplo de matriz de custo

		Predito	
Real		0	8
		1	0

Fonte: O Autor

apresenta uma matriz de custo de penalização 1:8, a qual penaliza os erros do tipo falso negativo 8 vezes mais do que os erros falsos positivos. Além de ser amplamente utilizada para atribuir pesos aos erros, a matriz de custo mostrou-se eficaz para compensar o desbalanceamento das classes minoritárias, evitando que o classificador priorize a classificação da classe majoritária.

Em nossos experimentos, adotamos a matriz de custo para lidar com os dados desbalanceados. Os detalhes sobre a aplicação da matriz de custo e do procedimento adotado para a definição da penalização são especificados na Seção 4.4.

2.4 Redução de dimensionalidade

A alta dimensionalidade de dados é um desafio amplamente conhecido na área de AM e de acentuada importância. No que se refere aos algoritmos de AM, têm-se conhecimento de que a alta dimensionalidade é prejudicial para o processo de aprendizado de vários algoritmos (HASTIE et al., 2009). Além disso, a alta dimensionalidade impõe desafios naturais no que tange a visualização dos dados, visto que estamos limitados a visualizar confortavelmente 3, ou, em alguns casos e com algum esforço, 4 dimensões. Neste contexto, inúmeras técnicas de redução de dimensionalidade têm emergido com o intuito de atenuar esses desafios.

A redução de dimensionalidade consiste na transformação de dados de alta dimensão em uma representação de dimensionalidade reduzida que preserve, tanto quanto possível, a estrutura dos dados de alta dimensionalidade (MAATEN; HINTON, 2008; MAATEN et al., 2009).

t-SNE é uma técnica de redução de dimensionalidade para visualização de dados. A técnica é capaz de capturar tanto a estrutura local dos dados de alta dimensão, quanto preservar a estrutura global, como *clusters* (MAATEN et al., 2009). A Figura 2.6 apresenta uma projeção obtida com t-SNE ao submeter ativações sinápticas de 1024 dimensões retornadas por uma rede convolucional, referentes à instâncias de quatro classes do conjunto de dados BRCars (KUHN; MOREIRA, 2021). Cada classe é representada

por uma cor. Pode-se observar que, de modo geral, a projeção preserva tanto a estrutura local (arranjo entre as instâncias pertencentes a mesma classe), quanto a estrutura global (*clusters* que separam instâncias de diferentes classes).

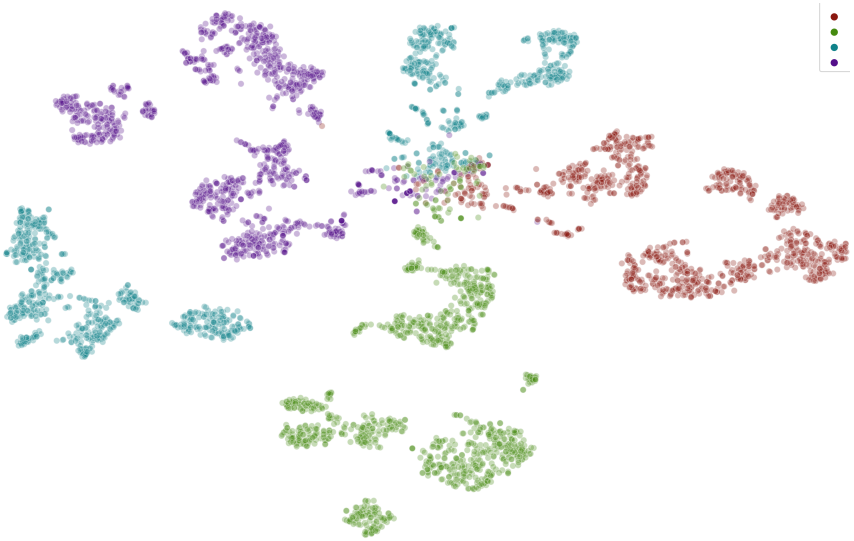


Figura 2.6 – Visualização 2D de ativações sinápticas 1024D retornadas por uma rede convolucional, referentes à instâncias pertencentes a quatro classes do conjunto de dados BRCars
Fonte: O Autor

Em nosso trabalho, a técnica t-SNE foi aplicada em conjunto com a abordagem de explicabilidade SHAP para compor uma técnica de visualização que permite extrair *insights* sobre a relação entre as variáveis consideradas pelos modelos preditivos e suas previsões corretas e incorretas. A Seção 4.6 apresenta mais detalhes acerca da aplicação da técnica t-SNE.

2.5 Seleção de atributos

Ao considerar conjuntos de dados compostos por dados provenientes de cenários reais, muitas variáveis podem ser irrelevantes, redundantes ou ruidosas. O processo de seleção de um subconjunto de variáveis úteis para a tarefa é denominado *seleção de atributos* (SA). Sob a perspectiva de dimensionalidade dos dados, SA está inserida como um dos processos de redução de dimensionalidade de dados.

SA promove a redução do custo computacional e de armazenamento, evitando perdas significativas de informação ou degradação do desempenho de aprendizagem (LI et al., 2017). Além disso, em alguns contextos, como é o caso do uso de algoritmos de AM para diagnóstico e prognóstico de pacientes, a utilização de um grande número de variáveis pode ser proibitivo ao se considerar a implementação de modelos preditivos

para o uso durante a rotina clínica.

De modo geral, os métodos de SA podem ser categorizados entre *filter-based*, *wrapper* e *embedded*. Os métodos do tipo *filter-based* fazem a seleção de um subconjunto de variáveis apenas com base em propriedades presentes nos dados. Já os métodos *wrapper* levam em consideração o desempenho preditivo de um algoritmo de AM predefinido para avaliar a qualidade das variáveis selecionadas. Por sua vez, os métodos *embedded* são intrinsecamente acoplados aos algoritmos de aprendizagem (LI et al., 2017). Exemplos de algoritmos de AM que possuem métodos de seleção de atributos intrínsecos são os algoritmos de regressão com algum fator de penalização, como penalização *LASSO* (L1) ou *Ridge* (L2) (TIBSHIRANI, 1996), que permitem desconsiderar variáveis durante o procedimento de modelagem.

É importante mencionar que, embora algoritmos com métodos *embedded* de SA estejam incluídos entre os algoritmos avaliados neste trabalho, a avaliação dos algoritmos se deu após a aplicação prévia de um método de SA do tipo *filter-based* comum a todos os algoritmos avaliados. A próxima seção apresenta o método de SA adotado em nossos experimentos.

2.6 *Correlation-based Feature Selection*

Correlation based Feature Selection (CFS) (HALL, 1999) é um método de SA que leva em consideração as correlações entre as variáveis independentes, bem como as correlações entre as variáveis independentes e a variável dependente. O método avalia o *mérito* de um subconjunto de atributos S:

$$merito(S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{2.2}$$

onde k refere-se ao número de variáveis do subconjunto S, $\overline{r_{cf}}$ é a correlação média entre as variáveis independentes e a variável dependente e $\overline{r_{ff}}$ é a correlação média entre as variáveis independentes do subconjunto S.

Conforme a equação 2.6, o numerador indica o poder preditivo das variáveis, enquanto que o denominador mensura o quão redundante é o conjunto de variáveis. Dessa forma, a equação retorna escores de *mérito* superiores para conjuntos de variáveis independentes que tenham baixa correlação entre si e alta correlação com as classes da variável dependente (LI et al., 2017).

Entre as vantagens desse método, pode-se citar o fato de que há uma avaliação conjunta das variáveis, ao contrário de outros métodos mais simples que avaliam isoladamente cada uma das variáveis independentes em relação à variável dependente, como é o caso do método *Mutual information*. Além do mais, devido a sua formulação, o método desestimula a seleção de variáveis altamente correlacionadas.

2.7 Métricas de avaliação

A avaliação dos classificadores consiste em uma etapa imprescindível em qualquer *pipeline* de AM. No que se refere aos algoritmos de classificação, é nessa etapa que se avalia a eficácia da classificação e a capacidade de generalização dos algoritmos em novos conjuntos de dados. Para isso, uma série de métricas, cada uma com suas forças e limitações, foram propostas para auxiliar no processo de avaliação.

Com o objetivo de avaliar os modelos de classificação, entre as métricas adotadas, pode-se citar métricas de Sensibilidade (Revocação), Precisão (ou Valor Preditivo Positivo - VPP), medida-F1 (F1), área sob a curva ROC (*Area Under Receiver Operating Characteristic Curve - AUROC*) e área sob a curva Precisão-Recall (Area Under Precision Recall - AUPRC). Essas métricas são tradicionalmente adotadas em tarefas de classificação. O restante desta seção apresenta com maior profundidade todas as métricas utilizadas neste trabalho.

- **Matriz de confusão** - A matriz de confusão é uma tabela na qual as linhas e colunas apresentam contagens das classificações corretas e incorretas realizadas pelo modelo. As contagens são categorizadas de forma a distinguir as classificações em quatro células. Duas células contabilizam os acertos do modelo: (i) Verdadeiros Positivos (VP) e (ii) Verdadeiros Negativos (VN). As demais células contabilizam as classificações incorretas: (iii) Falsos Positivos (FP) e (iv) Falsos Negativos (FN).

Tabela 2.2 – Exemplo de matriz de confusão

		Classe predita	
		1	0
Real	1	VP	FN
	0	FP	VN

Fonte: O Autor

A partir das informações presentes na matriz de confusão, pode-se calcular métricas como, revocação (sensibilidade), precisão, especificidade e medida-F1.

- **Sensibilidade (Sen) ou Revocação** - É a proporção de instâncias positivas classificadas corretamente entre todas as instâncias realmente pertencentes à classe positiva. A revocação é obtida pela equação 2.7.

$$rev = \frac{VP}{VP + FN} \tag{2.3}$$

A revocação também é recorrentemente chama de *sensibilidade* e pode ser interpretada como a métrica que avalia a capacidade do classificador de identificar as instâncias positivas. No contexto deste trabalho, a revocação informa a capacidade do classificador em identificar os pacientes que faleceram, foram admitidos na CTI ou necessitaram recursos de VMI.

- **Valor Preditivo Positivo (VPP) ou Precisão** - Informa a proporção de instâncias positivas classificadas corretamente entre todas as instâncias classificadas como pertencentes à classe positiva. A precisão é calculada a partir da equação 2.7.

$$pre = \frac{VP}{VP + FP} \tag{2.4}$$

A precisão pode ser interpretada como uma métrica para mensurar a performance do classificador entre as instâncias classificadas como positivas. No contexto deste trabalho, a precisão informa a proporção de pacientes corretamente classificados como pertencentes à classe positiva entre todos os pacientes classificados como pertencentes à classe positiva.

- **Medida-F1 (F1)** - É a média harmônica da precisão e a revocação. A F1 tem o objetivo de fornecer um único índice de desempenho que pondera igualmente a precisão e revocação. F1 é definido da seguinte forma:

$$F1 = 2 \times \frac{pre \times rec}{pre + rec} \tag{2.5}$$

Classificadores com índices altos de precisão e revocação resultam em índices altos de F1. Assim como para a revocação e precisão, quanto maior a F1, maior a eficácia do classificador.

- **Macro-F1 (ma-F1)** - Consiste na média aritmética entre as F1 de todas as classes. A macro-F1 é dada pela seguinte equação:

$$ma\_F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i, \tag{2.6}$$

onde C é o número de classes e $F1_i$ é a F1 da classe i. Por se tratar de uma média aritmética, a macro-F1 atribui pesos iguais para as F1 de ambas as classes.

- **Especificidade (Esp)** - Assim como se pode calcular a proporção de acerto para a classe positiva (revocação ou sensibilidade), também é pertinente calcular a proporção de previsões corretas na classe negativa. A especificidade mede essa proporção:

$$esp = \frac{VN}{VN + FP} \tag{2.7}$$

- **Valor preditivo negativo (VPN)** - Enquanto que o VPP mensura a proporção de instâncias positivas classificadas corretamente entre todas as instâncias classificadas como pertencentes à classe positiva, a métrica VPN calcula a proporção de instâncias da classe negativa classificadas corretamente entre todas as instâncias classificadas como pertencentes à classe negativa. A equação que calcula a VPN é dada por:

$$VPN = \frac{FP}{FP + TP} \tag{2.8}$$

- **Kappa** - Kappa (COHEN, 1960) é uma medida da concordância geral entre dois avaliadores que classificam itens em um determinado conjunto de categorias (KVÅLSETH, 1989).

$$Kappa = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)} \tag{2.9}$$

No contexto de avaliação de algoritmos de classificação, o classificador assume um dos papéis de avaliador e outro simboliza o observador que tem acesso ao *ground-truth*.

- **Curvas ROC e PR**

As métricas supracitadas são estimadas após a definição de um limiar de dicotomização sob as probabilidades retornadas pelo modelo. Visto que a escolha do limiar influencia os erros e acertos do modelo, a escolha do limiar é variável e dependente

da tarefa de classificação (SAITO; REHMSMEIER, 2015).

Os gráficos ROC e PR possibilitam mensurar a performance dos modelos sob diferentes limiares de probabilidade. A curva ROC apresenta a relação entre a taxa de verdadeiros positivos (sensibilidade ou revocação) e taxa de falsos positivos (especificidade -1) em todos os limiares possíveis para definir uma instância como pertencente à classe positiva.

Embora a curva ROC seja comumente utilizada como uma das principais métricas para avaliação dos classificadores (PAIVA et al., 2021), têm-se conhecimento de que quando utilizada em tarefas com desbalanceamento de classes, a curva ROC tende a superestimar a capacidade de discriminação dos classificadores (SOFAER; HOETING; JARNEVICH, 2019; SAITO; REHMSMEIER, 2015).

Utilizada na área de Recuperação da Informação (BAEZA-YATES; RIBEIRO-NETO et al., 1999), a curva PR possibilita visualizar a relação entre a precisão e revocação ao se adotar diferentes limiares de probabilidade. A curva PR também mostrou-se útil para a avaliação de tarefas de classificação, especialmente em casos de classes desbalanceadas (SAITO; REHMSMEIER, 2015). Ao contrário da curva ROC, a curva PR não incorpora as previsões corretas da classe negativa (especificidade) e portanto, é menos suscetível a superestimar a performance do modelo em tarefas com desbalanceamento de classes (SOFAER; HOETING; JARNEVICH, 2019).

Área sob a curva ROC - A partir da curva ROC, é possível mensurar estatísticas que resumem a capacidade inerente de um modelo de distinguir as classes a partir de todos os possíveis limiares de dicotomização. Entre as estatísticas que resumem a curva ROC, a área sob a curva ROC (AUROC) é a mais popular e amplamente utilizada (LIU et al., 2005).

Área sob a curva Precisão-Revocação - A área sob a curva PR sumariza em uma única estatística a performance do classificador, em termos de precisão e revocação, observada em diferentes limiares de probabilidade (semelhante à AUROC). Quanto maior a área sob a curva PR, maiores os índices de revocação e precisão atingidos pelo modelo.

2.8 Explicabilidade e SHAP

A eficácia não é a única propriedade importante durante o emprego de modelos preditivos. Em áreas sensíveis, a compreensão detalhada sobre o modelo e quais os fatores

levados em considerações durante as predições, também são propriedades de acentuada importância (BURKART; HUBER, 2021; LUNDBERG; LEE, 2017). A explicabilidade (*explainability*), por alguns autores também denominada de interpretabilidade (*interpretability*), consiste na propriedade que possibilita a compreensão humana acerca dos fatores levados em consideração pelos modelos durante o processo preditivo.

Embora a busca por modelos preditivos que permitam a compreensão por humanos remeta aos primórdios do campo de Inteligência Artificial (IA), esta continua sendo uma questão de pesquisa não exaurida, que inclusive, tem recebido crescente atenção devido ao advento dos chamados modelos de Aprendizado Profundo (*Deep Learning - DL*) e dos modelos de caixa preta (*Black-Box models*).

Neste contexto, abordagens que promovem explicabilidade têm sido propostas. LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e SHAP (LUNDBERG; LEE, 2017) são duas abordagens amplamente conhecidas que oferecem o que é conhecido como explicações locais. Tanto LIME quanto SHAP fazem uso de modelos substitutos (*surrogate models*) para prover a explicabilidade local. O racional é que esses modelos sejam capazes de incorporar o comportamento dos modelos de caixa preta durante uma predição individual. Esses modelos substitutos são interpretáveis (modelos lineares), treinados para apresentar predições aproximadas às predições realizadas pelos modelos de caixa preta. Após treinados, esses modelos interpretáveis são utilizados para a análise da contribuição exercida por cada atributo.

A abordagem SHAP – SHAP (LUNDBERG; LEE, 2017) é uma abordagem para prover explicabilidade de modelos preditivos. SHAP baseia-se nos *shapley values* — uma abordagem originada da teoria dos jogos por Shapley (1953) — para explicar a contribuição que cada variável exerce na saída predita por um modelo. Além de poder explicar predições individualmente (interpretação local), SHAP é extensivamente utilizado para análises de interpretabilidade global, onde busca-se compreender a relação entre valores das variáveis e as predições retornadas pelo modelo a partir de uma perspectiva geral.

Em nossos experimentos, utilizamos a abordagem KernelSHAP para a interpretação dos modelos, visto que essa é uma abordagem agnóstica, que não pressupõe um tipo de modelo específico. KernelSHAP faz uso de um conjunto de instâncias denominado *background*. As contribuições são calculadas com base nesse conjunto *background*, a partir do qual é calculada predição média. Durante a geração de coalizões entre as variáveis das quais os *shapley values* são calculados, as instâncias *background* são utilizadas para prover amostras de valores para simular a ausência de variáveis.

A abordagem SHAP foi implementada por seus autores e disponibilizada como uma biblioteca que recebe o mesmo nome. Além de disponibilizar um conjunto de técnicas para calcular os *shapley values*, a biblioteca também oferece visualizações que contribuem para a interpretação dos modelos. Uma das visualizações é a *summary-plot*, que possibilita a visualização global acerca das contribuições de cada variável na predição da classe alvo (mortalidade, CTI e VMI).

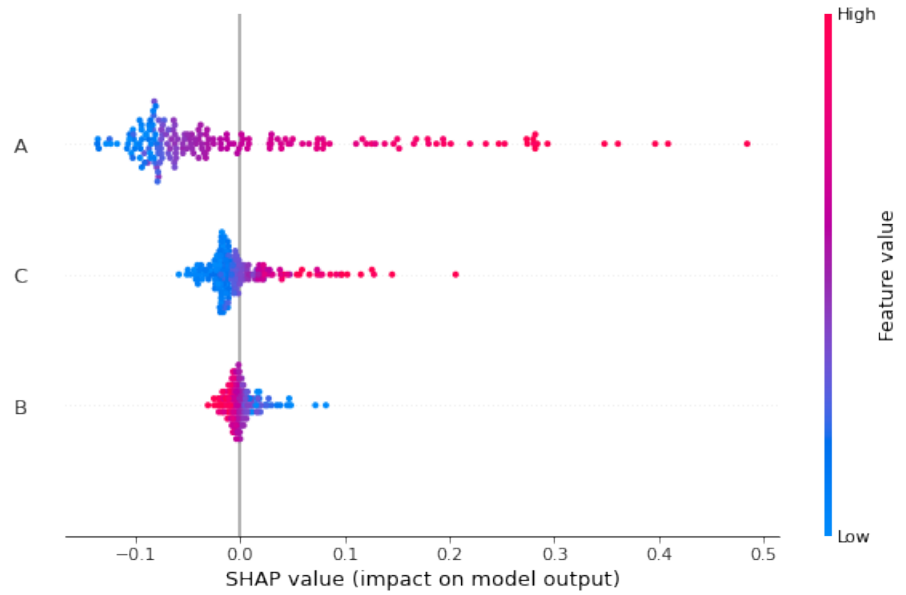


Figura 2.7 – Exemplo do *summary-plot* da biblioteca SHAP
Fonte: O Autor

A Figura 2.7 apresenta a visualização *summary-plot* da biblioteca SHAP. São apresentadas as contribuições de três atributos A, B e C em uma tarefa de classificação binária. Os atributos estão dispostos em relação a sua importância: quanto mais ao topo, maior a importância atribuída pelo modelo. Cada ponto projetado refere-se a uma instância e as colorações são atribuídas em relação a intensidade dos valores dos atributos. A linha vertical do gráfico separa as regiões de contribuição negativa e positiva. Quanto mais à esquerda estão projetados os pontos em relação à linha vertical, mais negativa é a contribuição daquele atributo para a predição da classe positiva. Por outro lado, quanto mais à direita estão projetados os pontos em relação à linha vertical, maior a contribuição daquele atributo para a predição da classe positiva.

Por exemplo, supondo que o *summary-plot* da Figura 2.7 seja referente a uma tarefa de predição de mortalidade (a classe positiva é a classe *mortalidade=sim*), ao considerar o atributo A, a grande concentração de pontos azuis na região esquerda à linha vertical indica que valores baixos para o atributo (representados em azul), apresentam tendência de contribuição negativa para a predição de *mortalidade=sim*. Pontos com valores

medianos para o atributo A (representados por tons roxos) tendem a estar concentrados próximos a linha vertical, local onde arranjam-se os pontos para os quais os valores dos atributos apresentam valores de contribuição de pequena magnitude. Ainda, é possível observar que pontos com valores altos para o atributo A (coloridos em tons vermelhos) situam-se à direita da linha vertical, sinalizando que valores altos para o atributo A tendem a contribuir positivamente para a predição de *mortalidade=sim*. Embora com importância de contribuição absoluta menor do que o atributo A, o atributo C apresenta uma tendência similar ao atributo A, com valores baixos apresentando tendência à contribuição negativa (embora de baixa magnitude) para a predição da classe *mortalidade=sim* e valores altos apresentando maiores contribuições para a predição da classe *mortalidade=sim*. Por fim, o atributo B é aquele que apresenta a menor contribuição absoluta entre os três atributos. O atributo B apresenta uma tendência inversa aos demais atributos: valores baixos para o atributo tendem a apresentar contribuições de pequena magnitude para a predição da classe *mortalidade=sim* e valores altos tendem a apresentar contribuição negativa de pequena magnitude para a predição da classe *mortalidade=sim*.

2.9 Resumo do Capítulo

Este capítulo apresentou os principais conceitos necessários para subsidiar a compreensão do presente trabalho. Os conteúdos apresentados abrangem os principais aspectos acerca dos algoritmos de classificação avaliados nesse trabalho, bem como os procedimentos e técnicas que permeiam a referida tarefa, como a seleção de atributos, a aplicação de matrizes de custo para o manejo de conjuntos de dados com desbalanceamento de classes e métricas de avaliação de desempenho dos classificadores. Além disso, sumarizamos duas técnicas, uma de visualização de dados (t-SNE) e outra de *explainability* (SHAP), as quais constituem as bases para a técnica de visualização proposta neste trabalho.

3 TRABALHOS RELACIONADOS

Este capítulo discute trabalhos existentes que abordam as tarefas de predição de mortalidade, necessidade de VMI e CTI para pacientes com COVID-19 com base nas informações disponíveis na admissão hospitalar. Consideramos tanto trabalhos avaliando escores de risco clássicos (incluindo aqueles projetados especialmente para pacientes com COVID-19 e aqueles que antecedem a pandemia), quanto classificadores baseados em AM.

No que se refere aos algoritmos adotados para o desenvolvimento dos modelos preditivos de forma geral, os modelos de regressão logística destacam-se entre os preditores propostos. Gupta et al. (2020) realizaram a avaliação externa de 22 modelos de prognóstico, dentre os quais, 12 eram modelos de regressão logística. Esta prevalência por modelos estatísticos tradicionais também foi apontada por Paiva et al. (2021), pela constatação de que a maioria dos 76 trabalhos revisados utilizaram métodos estatísticos tradicionais (60%), como regressão logística multivariada e modelo de Cox. A aplicação de abordagens de AM foi observada em aproximadamente 39% dos estudos.

No âmbito dos dados utilizados para a modelagem dos preditores, observa-se que a falta de adoção de um protocolo único pelos hospitais bem como a variabilidade na alocação de recursos resultam em conjuntos de dados heterogêneos. Isso significa que conjuntos de dados provenientes de diferentes hospitais nem sempre contêm os mesmos atributos. Assim, a avaliação de um modelo preditivo gerado a partir de um conjunto de dados proveniente de um hospital pode ser inviabilizada em um conjunto de dados de um outro hospital – em casos em que nem todos os atributos utilizados pelo modelo estão disponíveis (BURDICK et al., 2020; KNIGHT et al., 2020).

Os conjuntos de dados utilizados nos estudos geralmente variam de uma única unidade hospitalar com algumas centenas de pacientes, a conjuntos de dados que abrangem milhares de registros de hospitalização provenientes de centenas de unidades hospitalares em vários países. Apenas para citar alguns exemplos, Covino et al. (2020) e Zhao et al. (2020) realizaram seus estudos com base em conjuntos de dados contendo 334 e 641 pacientes, respectivamente, de uma única unidade hospitalar. Paiva et al. (2021) utilizaram um conjunto de dados com 5032 pacientes de 36 hospitais em 17 cidades de 5 estados brasileiros. Para desenvolver o escore 4C, Knight et al. (2020) utilizaram um grande conjunto de dados com 57824 pacientes coletados de 260 hospitais na Inglaterra, Escócia e País de Gales.

Embora os conjuntos de dados utilizados nos diferentes estudos não contenham as mesmas variáveis, há uma recorrência entre as variáveis consideradas mais relevantes pelos modelos preditivos. Para a maioria das tarefas, a idade é a variável mais citada. Outras variáveis incluem saturação de O2, níveis de neutrófilos, frequência respiratória, comorbidades, insuficiência cardíaca, lactato desidrogenase (LDH), proteína C-reativa (PCR), métricas de nível de consciência (por exemplo, Escala de Coma de Glasgow), entre outros.

Entre as métricas utilizadas para a comparação e avaliação dos modelos preditivos, destaca-se a AUROC, empregada para avaliar a capacidade de discriminação entre as classes. Menos frequentemente, outras métricas como sensibilidade, especificidade, VPP e VPN são adotadas para complementar o processo avaliativo dos modelos propostos.

Como enfatizado por Paiva et al. (2021), apesar de na pesquisa médica a métrica AUROC ser amplamente utilizada como única medida de capacidade discriminatória dos modelos, isoladamente, esta métrica é insuficiente para avaliar e comparar modelos. Além do mais, têm-se conhecimento que quando aplicada em cenários de desbalanceamento entre as classes, caso comum em tarefas relacionadas à área da saúde, a métrica AUROC pode superestimar o desempenho dos modelos (DAVIS; GOADRICH, 2006; SAITO; REHMSMEIER, 2015).

O restante desse capítulo está dividido em quatro seções. A Seção 3.1 trata dos escores de risco utilizados para tarefas prognósticas. A Seções 3.2 e 3.3 discutem trabalhos relacionados a esta dissertação que tratam de tarefas de predição de mortalidade, necessidade de uso de VMI e de internação em CTI. Por fim, a Seção 3.4 apresenta as revisões sistemáticas que foram publicadas sobre essas tarefas prognósticas no escopo da COVID-19.

3.1 Escores de risco para prognóstico

O desenvolvimento de ferramentas computacionais para auxiliar no diagnóstico e prognóstico de pacientes vem sendo estudado há décadas. No contexto das ferramentas de prognóstico, os escores de risco baseados em modelagem estatística clássica estão entre os mais estudados. Esses escores propõem-se a auxiliar nas mais variadas demandas da rotina hospitalar, desde facilitar a detecção e categorização de gravidade de doenças, bem como, alertar e possibilitar a intervenção imediata ou antecipada por parte da equipe médica.

Um escore extensamente utilizado na rotina clínica é o escore *Sequential Organ Failure Assessment* (SOFA) (VINCENT et al., 1996). SOFA foi desenvolvido na década de 90 com a proposta de descrever quantitativamente e, de forma objetiva, o grau de disfunção ou falha de órgãos ao longo do tempo em grupos de pacientes (embora também seja aplicado de maneira individual). Estudos mostraram que o escore também apresentava boa correlação entre a pontuação e a mortalidade e, com isso, passou a ser recorrentemente utilizado para avaliação de mortalidade em UTIs (LAMBDEN et al., 2019).

Neste contexto, foi proposta uma variedade de escores para alerta precoce. Entre esses escores, pode-se citar o *Early Warning Score* (EWS) proposto por (MORGAN; WILLIAMS; WRIGHT, 1997), projetado para auxiliar na assistência clínica à beira leito de pacientes que apresentam sinais fisiológicos de doença crítica estabelecida ou iminente (MORGAN; WRIGHT, 2007). Já o *Modifield Early Warning Score* (MEWS), foi concebido a partir de uma modificação do escore de (MORGAN; WILLIAMS; WRIGHT, 1997) com o intuito de identificar pacientes em risco e auxiliar na avaliação precoce e admissão em unidades de emergência (SUBBE et al., 2001). Proposto pelo *Royal College of Physicians*, o escore *National Early Warning Score* (NEWS) foi desenvolvido com o objetivo de auxiliar no rastreamento e categorização de pacientes com doença aguda (PHYSICIANS, 2021), enquanto o escore *Rapid Emergency Medicine Score* (REMS) foi proposto para prever mortalidade e tempo de permanência hospitalar de pacientes não cirúrgicos atendidos em departamento de emergência (OLSSON; TERÉNT; LIND, 2004). É importante ressaltar que tanto esses escores, quanto vários outros, têm sido avaliados para outras demandas em complemento àquelas para as quais foram originalmente concebidos.

Desde o surgimento da COVID-19, uma grande quantidade de trabalhos têm avaliado a adequação dos escores já existentes, bem como novos escores para as demandas ocasionadas pela doença (COVINO et al., 2020; KNIGHT et al., 2020; FRANCONE et al., 2020; JI et al., 2020; LIANG et al., 2020; LIU et al., 2020; HU; YAO; QIU, 2020). Concomitantemente, uma série de trabalhos têm avaliado e comparado esses escores com abordagens de AM mais recentes. Entre esses trabalhos, observa-se que as abordagens de AM recorrentemente atingem resultados superiores aos escores tradicionais no que se refere à capacidade de classificação (BURDICK et al., 2020; KNIGHT et al., 2020). Entretanto, a falta de transparência quanto aos critérios levados em consideração por esses modelos ao predizer um determinado prognóstico, têm sido apontada como um fator limitante para a sua adoção na rotina clínica (YU et al., 2021; KNIGHT et al., 2020; YADAW et al., 2020).

Tabela 3.1 – Análise comparativa dos trabalhos relacionados à mortalidade

Autor	Registros	Hospitais	Técnica	Principais Atributos	Principais resultados
Paiva et al. (2021)	5032	36	FNet ResNet *LightGBM* SVM KNN *Ensembles*	idade ureia proteína C-reativa LDH frequência respiratória frequência cardíaca neutrófilos sódio pCO2	Modelo *Ensemble*: AUROC: 0,871 F1-obito: 0,564 F1-alta: 0,913 Macro-F1: 0,739
Zhao et al. (2020)	641	1	Regr Log	insuficiência cardíaca procalcitonina lactato desidrogenase DPOC saturação de oxigênio frequência cardíaca idade	AUROC: 0,82 Sensibilidade: 0,071 Especificidade: 1,0
Covino et al. (2020)	334	1	MEWS NEWS NEWS2 qSOFA REMS	parâmetros do REMS: saturação de oxigênio frequência respiratória frequência cardíaca pressão arterial média escala de Glasgow idade	REMS (48 horas): AUROC: 0,882 Sensibilidade: 0,909 Especificidade: 0,718 REMS (7 dias): AUROC: 0,823 Sensibilidade: 0,961 Especificidade: 0,568
Knight et al. (2020)	57824	260	GAM+Regr Log *XGBoost*	idade sexo número de comorbidades frequência respiratória saturação de oxigênio escala de Glasgow ureia PCR	GAM - Regr Log: AUROC: 0,767 escore de Brier: 0,171 *XGBoost*: AUROC: 0,779 escore de Brier: 0,197

Em nosso trabalho, não foi possível comparar nossos modelos com os escores de risco, visto que nem todas as variáveis utilizadas por esses escores estavam disponíveis em nossos conjuntos de dados.

3.2 Predição de mortalidade

Nesta seção falaremos sobre trabalhos que abordam a tarefa de predição de mortalidade de pacientes com COVID-19. Várias técnicas foram empregadas para abordar a predição da mortalidade, como o uso de escores de risco baseados em abordagens estatísticas clássicas e algoritmos de classificação (COVINO et al., 2020; ZHAO et al., 2020; KNIGHT et al., 2020).

Covino et al. (2020) avaliaram seis escores de pontuação fisiológica recorrente-

mente usados como escores de alerta precoce para prever o risco de hospitalização em CTI e o risco de mortalidade - ambas as tarefas considerando até sete dias. O REMS atingiu os melhores resultados com uma AUROC de 0,823 (95% IC, [0,778 - 0,863]). Zhao et al. (2020) desenvolveram escores de risco específicos para prever mortalidade e internação em CTI. Os escores foram desenvolvidos com base no algoritmo de regressão logística que atingiu uma AUROC de 0,82 (IC 95%, [0,73 - 0,92]).

Outro escore projetado especificamente para prever o risco de mortalidade foi o *4C Mortality Score* (4C) (KNIGHT et al., 2020). O escore foi concebido a partir do uso de modelos aditivos generalizados, os quais foram utilizados para pré-selecionar variáveis. As variáveis foram categorizadas e utilizadas para o treinamento de um modelo de regressão logística. O escore 4C atingiu uma AUROC de 0,77 (95% IC, [0,76 - 0,77]), que foi superado pelo algoritmo *XGBoost* por uma pequena margem. Ainda assim, os autores afirmaram que o seu escore é preferível devido à dificuldade em compreender as previsões feitas pelo *XGBoost*.

Paiva et al. (2021) também avaliaram algoritmos de classificação para previsão de mortalidade, incluindo um modelo *transformers* (FNet), redes neurais convolucionais (CNN), algoritmos de *boosting* (LightGBM), *Random Forest* (RF), *Suport Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN). Além disso, um metamodelo *ensemble* foi treinado para combinar todos os classificadores mencionados. Os melhores resultados foram alcançados pelo modelo *ensemble*, com uma F1 média para a classe *mortalidade=sim* de 0,56 e uma AUROC de 0,87 (IC 95% [0,86 - 0,88]). Para verificar as variáveis e suas contribuições para as predições, o estudo utilizou a abordagem de explicabilidade SHAP.

A Tabela 3.1 resume as principais características da literatura relacionada à predição de mortalidade, incluindo os conjuntos de dados utilizados, as técnicas utilizadas, as principais características e seus resultados. Nosso estudo complementa os trabalhos mencionados ao avaliar seis algoritmos em novos conjuntos de dados de dois hospitais.

3.3 Predição de necessidade de admissão em CTI e uso de VMI

Nesta seção discorremos acerca de trabalhos que desenvolveram e/ou avaliaram modelos para a predição de necessidade de internação em CTI e necessidade de uso de recursos de VMI. A predição de admissão à CTI consiste em identificar quais pacientes apresentam maior probabilidade de necessitar leitos de CTI durante a hospitalização. Esta é uma tarefa crucial para o gerenciamento de recursos, visto que a disponibilidade de

leitos de CTI é limitada. Dada a sua importância, vários trabalhos têm proposto métodos para auxiliar na triagem de pacientes acometidos pela COVID-19.

A Tabela 3.2 sumariza os estudos discutidos em nosso trabalho que avaliaram a tarefa de CTI. O estudo anteriormente citado conduzido por Covino et al. (2020), também avaliou seis escores de risco na tarefa de predição de admissão à CTI. Embora não tenha sido concebido especificamente para a utilização em pacientes acometidos pela COVID-19, o escore NEWS apresentou o melhor desempenho tanto para a predição durante as primeiras 48 horas quanto para a admissão em até 7 dias, atingindo respectivamente, AUROC de 0,80 (IC 95%, [0,76 - 0,84]) e 0,78 (IC 95%, [0,73 - 0,83]), sensibilidade de 0,66 e 0,71 e, especificidade de 0,85 (IC 95%, [0,80 - 0,89]) e 0,77 (IC 95%, [0,72 - 0,82]), para os horizontes de 48 horas e 7 dias.

Tabela 3.2 – Análise comparativa dos trabalhos relacionados sobre predição de necessidade de internação em CTI

Autor	Registros	Hospitais	Técnica	Principais Atributos	Principais resultados
Zhao et al. (2020)	641	1	Regr Log	procalcitonina lactato desidrogenase saturação de oxigênio hitórico de fumante linfócitos	AUROC: 0,74 Sensibilidade: 0.105 Especificidade: 0,992
Covino et al. (2020)	334	1	MEWS NEWS NEWS2 qSOFA REMS	parâmetros do NEWS: saturação de oxigênio frequência respiratória frequência cardíaca pressão arterial média escala de Glasgow idade	NEWS (48 horas): AUROC: 0,802 Sensibilidade: 0,660 Especificidade: 0,849 NEWS (7 dias): AUROC: 0,783 Sensibilidade: 0,714 Especificidade: 0,773

O estudo de Zhao et al. (2020) também desenvolveu e avaliou escores de risco para admissão em CTI para pacientes com COVID-19. O modelo atingiu AUROC de 0,74 ([95% IC, 0,63 - 0,85]). As variáveis com maior poder preditivo foram Lactato desidrogenase, procalcitonina, saturação de O2, histórico de tabagismo e contagem de linfócitos.

A tarefa de predição de VMI, assim como a predição de CTI, é de acentuada importância no contexto da COVID-19, doença que compromete diretamente o sistema respiratório dos infectados. As principais características dos estudos discutidos neste trabalho, que avaliam a tarefa de VMI, estão apresentados na Tabela 3.3.

Burdick et al. (2020) desenvolveram um modelo para predição de necessidade de

suporte de VMI nas primeiras 24 horas a contar a admissão hospitalar. O algoritmo utilizado foi o *XGBoost*. O modelo desenvolvido atingiu sensibilidade de 0,90, especificidade de 0,58 e AUROC de 0,87. O modelo obteve resultados superiores ao escore MEWS, que atingiu sensibilidade de 0,78, especificidade de 0,40 e AUROC de 0,64. Embora os resultados obtidos com o modelo desenvolvido sejam promissores, evidencia-se o baixo número de pacientes do conjunto de dados que demandaram suporte de ventilação mecânica (10 pacientes).

Tabela 3.3 – Análise comparativa dos trabalhos relacionados sobre VMI

Autor	Registros	Hospitais	Técnica	Principais Atributos	Principais resultados
Yu et al. (2021)	1980	8	XGBoost	idade temperatura corporal frequência respiratória saturação de oxigênio histórico tabagismo diabetes doença cardíaca doença pulmonar	acurácia: 0.86 NPV 0,878% especificiade 0,976% AUROC 0,68%
Burdick et al. (2020)	197	1	XGBoost	pressão diastólica pressão sistólica frequência cardíaca temperatura corporal frequência respiratória saturação de oxigênio células brancas sanguíneas contagem de plaquetas lactato nível de ureia creatinina bilirrubina	AUROC: 0,87 sensibilidade 0,90 especificidade 0,58
Rodriguez et al. (2021)	3111	1	XGBoost Regr Log	doenças respiratórias complicações renais	XGBoost: AUROC: 0,743 AUPRC: 0,137

Outro modelo *XGBoost* para predição de VMI foi desenvolvido por Yu et al. (2021). Diferente do estudo de Burdick et al. (2020), o trabalho de Yu et al. (2021) não limitou-se em avaliar a predição de necessidade de VMI nas primeiras 24 horas. O modelo obteve uma acurácia de 0,86 (IC 95%, [0,834 - 0,886]), especificidade de 0,976 e AUROC de 0,68. O trabalho também aplicou a abordagem SHAP, com a qual constatou-se que o aumento da idade, aumento da temperatura corporal e aumento da frequência respiratória, bem como, valores baixos para saturação de O2, histórico de tabagismo e diabetes *mellitus* foram fatores que aumentaram a chance do paciente necessitar VMI.

Além de um modelo *XGBoost*, Rodriguez et al. (2021) também desenvolveram modelos de regressão logística com regularização *Elastic Net* (RL-EN) e modelos de regressão logística com regularização LASSO. Considerando a AUROC, *XGBoost* obteve

o maior resultado (0,738 (95% IC, [0,682 - 0,812])). Ao considerar a AUPRC, *XGBoost* atingiu o segundo maior resultado (0,137 (95% IC, [0,047 - 0,175])) e RL-NE a maior AUPRC (0,141 (95% IC, [0,046 - 0,183]).

Assim como os demais trabalhos que avaliaram VMI, Rodriguez et al. (2021) utilizaram a abordagem SHAP para prover a explicabilidade das variáveis utilizadas pelos modelos. Embora a idade não tenha sido considerada a variável com maior contribuição durante as predições, inesperadamente, valores altos para a idade dos pacientes foram associados negativamente à necessidade de VMI. Os autores argumentaram que este comportamento é possivelmente um reflexo de diretivas de antecipação e dos processos de decisão clínicos, em vez de uma menor incidência de insuficiência respiratória grave. No geral, as variáveis mais relevantes para a predição de VMI foram síndromes e doenças respiratórias, bem como, complicações renais.

3.4 Revisões sistemáticas

Nesta seção falaremos sobre trabalhos de revisão que realizaram avaliações de modelos preditivos propostos, salientando as principais limitações identificadas por essas revisões.

Dado o cenário crítico desencadeado pela COVID-19, já nos primeiros meses após o seu surgimento, observou-se a proliferação de vários estudos reportando o desenvolvimento e avaliação de modelos preditivos para o diagnóstico e prognóstico de pacientes acometidos pela COVID-19.

Entretanto, devido às dificuldades inerentes ao contexto da pandemia e às evidentes limitações impostas ao se conduzir estudos sobre uma doença recente e, ainda em curso, muitos trabalhos foram realizados em caráter de urgência (WYNANTS et al., 2020; BOOTH; ABELS; MCCAFFREY, 2021; ROBERTS et al., 2021). Desta forma, muitos modelos foram desenvolvidos com base em conjuntos de dados pequenos, que com frequência continham pacientes provenientes de uma única unidade hospitalar, o que inviabiliza uma adequada avaliação dos modelos. Com o objetivo de auxiliar no desenvolvimento e avaliação desses modelos, revisões sistemáticas aplicaram esforços para selecionar modelos propostos, avaliando seus pontos fortes e limitações.

Neste contexto, Wynants et al. (2020) realizaram uma revisão sistemática com o objetivo de avaliar a validade e utilidade de trabalhos que propuseram modelos preditivos para diagnóstico e prognóstico de pacientes, bem como para detecção de pessoas na po-

pulação em geral com risco aumentado de infeção por COVID-19 ou de hospitalização devido a doença. O estudo avaliou 196 trabalhos, os quais descreveram 232 modelos preditivos, sendo que 107 objetivavam a predição de um ou mais prognósticos relacionados à COVID-19. Como resultado, todos os modelos foram classificados com risco de viés alto ou inconclusivo.

Gupta et al. (2020) realizaram a avaliação e validação externa de 22 modelos preditivos de prognóstico para deterioração clínica e mortalidade dos pacientes. A definição de deterioração clínica adotada pelo estudo contempla aqueles pacientes que necessitaram de suporte ventilatório durante a hospitalização ou morreram.

A principal métrica utilizada na avaliação dos modelos foi a AUROC. A eficácia dos modelos foi comparada com o uso das variáveis *saturação de O2* e *idade* do paciente para predição de deterioração e mortalidade, respectivamente. Como resultado, constatou-se que dos modelos que retornavam probabilidades, tanto de deterioração clínica quanto de mortalidade, nenhum apresentou calibração adequada. O estudo concluiu que nenhum dos modelos prognósticos avaliados ofereceu aumento substancial de eficácia em relação à utilização dos atributos *saturação de O2* e *idade* do paciente. Achado semelhante foi constatado em (YADAW et al., 2020), onde observou-se a estagnação da métrica AUROC ao considerar outros atributos além da *idade*, *nível de saturação de O2* e *tipo de encontro com o paciente* (internação *versus* consultas ambulatoriais e telessaúde).

Recentemente, (MILLER et al., 2022) também realizaram uma revisão sistemática com o objetivo de avaliar os trabalhos publicados que utilizaram modelos preditivos para doenças graves causadas pela COVID-19. A avaliação foi realizada utilizando a ferramenta *Prediction Model Risk Of Bias Assessment Tool* (PROBAST) (WOLFF et al., 2019). Os aspectos avaliados foram: tipo do estudo, configuração, tamanho da amostra, tipo de validação e desfecho (intubação, ventilação, qualquer tipo de tratamento de suporte ou mortalidade). O estudo concluiu que dos 79 artigos que atenderam os critérios de inclusão, 70 artigos foram identificados como tendo alto risco ou risco inconclusivo de viés, ou preocupação alta ou pouco clara quanto à aplicabilidade.

Tanto Wynants et al. (2020) quanto Gupta et al. (2020) e (MILLER et al., 2022) advertiram quanto às limitações dos modelos preditivos para o uso na rotina clínica. Tais conclusões reforçam tanto a necessidade de aprimoramento, quanto de intensificação de avaliação acerca dos fatores levados em consideração por esses artefatos preditivos.

3.5 Resumo do Capítulo

Este capítulo apresentou e discutiu estudos que abordaram as tarefas de predição de mortalidade, internação em CTI e necessidade de VMI. Inicialmente, apresentamos as principais características referentes ao desenvolvimento de modelos para predição prognóstica. Na sequência, discorremos acerca dos escores de risco e sua relação com as tarefas prognósticas contempladas em nosso trabalho. Após, discorremos acerca de estudos que abordaram as tarefas de predição de mortalidade, necessidade de VMI e internação em CTI. Por fim, discutimos estudos que realizaram a avaliação de modelos preditivos descritos na literatura.

4 MATERIAIS E MÉTODOS

Este trabalho tem como objetivo prever a mortalidade, necessidade de VMI e CTI para pacientes com COVID-19 durante a internação do paciente com base nas informações disponíveis no momento da admissão. Abordamos estas tarefas através do uso de algoritmos de classificação.

Em nosso trabalho, realizamos a comparação de seis algoritmos de classificação. Além da métrica AUROC – recorrentemente utilizada como principal métrica de avaliação – utilizamos outras oito métricas que detalham a capacidade de classificação dos modelos. Além disso, desenvolvemos e avaliamos versões dos classificadores com apenas uma variável independente. Dentre esses, treinamos classificadores com a variável *idade* para as tarefas de predição de mortalidade e CTI, bem como apenas *saturação de O2* para a tarefa de predição de necessidade de VMI. Escolhemos estas variáveis pois elas são recorrentemente apontadas como relevantes para estas tarefas prognósticas. Por fim, desenvolvemos uma técnica de visualização baseada nos valores de contribuição obtidos com SHAP, que permite obter *insights* sobre a relação entre as variáveis levadas em consideração pelos modelos preditivos e suas previsões corretas e incorretas.

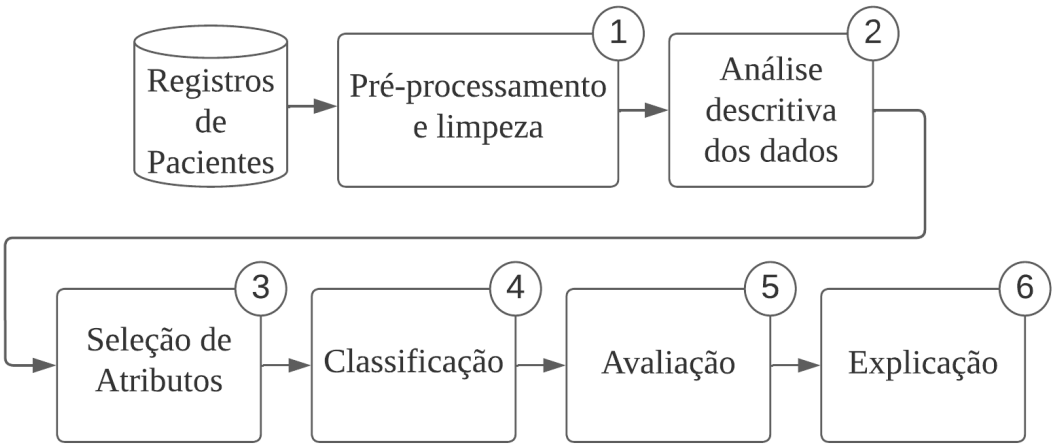


Figura 4.1 – Etapas da abordagem empregada neste trabalho
Fonte: O Autor

Nosso método compreende seis etapas, descritas na Figura 4.1. Inicialmente, nosso conjunto de dados de entrada passa pelo processo de pré-processamento e limpeza dos dados e então, realiza-se a análise descritiva dos dados. Após, realiza-se a seleção de atributos e na sequência, é submetido aos algoritmos de classificação. Os resultados dos classificadores são avaliados e submetidos à abordagem de explicabilidade.

As próximas seções descrevem os dados e cada etapa com mais detalhes.

4.1 Conjuntos de Dados e Pré-processamento

O estudo envolveu pacientes internados em duas instituições brasileiras: *Hospital Moinhos de Vento (HMV)* e *Hospital de Clínicas de Porto Alegre (HCPA)*. O HMV é um hospital privado considerado um dos melhores da América Latina [1]. O HCPA é um hospital público de ensino terciário vinculado academicamente à Universidade Federal do Rio Grande do Sul (UFRGS) e está entre os principais hospitais universitários públicos do Brasil [2].

Ambos os hospitais estão localizados em Porto Alegre, capital do estado do Rio Grande do Sul, no sul do Brasil, e são centros de assistência médica líderes regionais. O projeto foi aprovado pelo Comitê de Ética em Pesquisa institucional[3]. O consentimento informado foi dispensado devido à natureza retrospectiva do estudo. Os dados coletados não incluíam nenhuma informação identificável para garantir a privacidade do paciente.

Os pacientes incluídos no estudo são adultos (>= 18 anos) diagnosticados com COVID-19 através do *Reverse Transcription Polymerase Chain Reaction* (RT-PCR). Os atributos disponíveis nos conjuntos de dados referem-se a dados clínicos e demográficos: idade, sexo, etnia, comorbidades (hipertensão, obesidade, diabetes, etc.), exames da admissão (gasometria arterial, lactato, hemograma, plaquetas, proteína C reativa, d-dímeros, CK, creatinina, ureia, sódio, potássio, bilirrubina), necessidade de admissão em UTI, tipo de suporte ventilatório na admissão e no período em UTI (O2 cateter, ventilação não invasiva, cânula nasal de alto fluxo, ventilação mecânica invasiva), data de início e fim da hospitalização, bem como a mortalidade. A lista completa dos atributos disponíveis na admissão hospitalar utilizadas em nosso trabalho pode ser consultada nas tabelas de apêndice A.1, A.2 e A.3. O conjunto de dados do hospital HMV contempla registros tanto de hospitalizações regulares quanto de internações em CTI, enquanto que o conjunto de dados do hospital HCPA contém exclusivamente registros de internações em CTI.

Conjunto de dados HMV: O conjunto de dados HMV refere-se aos registros anonimizados de pacientes com confirmação de COVID-19 hospitalizados no Hospital Moinhos de Vento entre março de 2020 e junho de 2021. No total, o conjunto de dados possui registros de 1526 pacientes hospitalizados, dos quais 58,7% são homens. O conjunto de

[1] <https://www.hospitalmoinhos.org.br/institucional/o-hospital/premios-e-certificacoes>; Acesso: 29/12/2022.

[2] <https://www.hcpa.edu.br/institucional/institucional-apresentacao/institucional-premios-e-destaques>; Acesso: 29/12/2022.

[3] Aprovação do Comitê de Ética: HCPA CAAE: 32314720.8.0000.5327, HMV CAAE: 32314720.8.3001.5330

dados abrange tanto pacientes com hospitalização regular quanto pacientes internados em CTI. O processo de coleta dos dados foi conduzido por profissionais da saúde, que realizaram a transcrição manual das informações contidas no prontuário dos pacientes para um formulário contendo as variáveis previamente estabelecidas para o estudo (informações clínicas disponíveis na admissão hospitalar). Pacientes sem informação de mortalidade foram desconsiderados durante o processo de transcrição dos dados. Assim, registros de hospitalização nos quais os pacientes foram transferidos para outros hospitais e/ou casos de evasão não foram incluídos no estudo. Processos de verificação automática foram aplicados sobre os dados coletados com o objetivo de encontrar inconsistências decorrentes do processo de transcrição dos dados. Este procedimento incluiu a verificação de consistência de datas, bem como, a presença de *outliers*. Quando valores considerados suspeitos foram encontrados, os anotadores foram avisados para revisão dos dados.

Conjunto de dados HCPA: O conjunto de dados HCPA contém registros anônimos de 2269 pacientes com confirmação de COVID-19 internados na CTI do HCPA entre março de 2020 e agosto de 2021. Os dados foram coletados diretamente do banco de dados do hospital. Os desfechos de interesse foram a mortalidade e necessidade de VMI durante a internação. Assim, os registros foram estratificados nas classes *mortalidade=não* e *mortalidade=sim*, bem como *vmi=não* e *vmi=sim* de acordo com os dados fornecidos pelos hospitais. Os registros de exames, bem como, dados vitais foram coletados de uma base de dados diretamente alimentada pelos equipamentos que realizam exames e monitoram os pacientes. A exportação dos dados foi realizada pela equipe responsável pelo armazenamento e gerenciamento dos dados do hospital HCPA. Para isso, após a definição dos atributos de interesse e, com a prévia aprovação do comitê, gerou-se um conjunto de dados contendo as informações anonimizadas de internações em CTI dos pacientes.

4.1.1 Pré-processamento e limpeza dos dados

A primeira etapa após a coleta dos dados consiste no pré-processamento e limpeza dos dados. A Figura 4.2 apresenta um esquema com as etapas de pré-processamento aplicadas. As Subseções a seguir abordam os procedimentos realizados.

Exclusão de pacientes sem desfecho: O conjunto de dados HCPA desconsidera 20 registros de hospitalizações nas quais os pacientes morreram durante as primeiras 24 horas

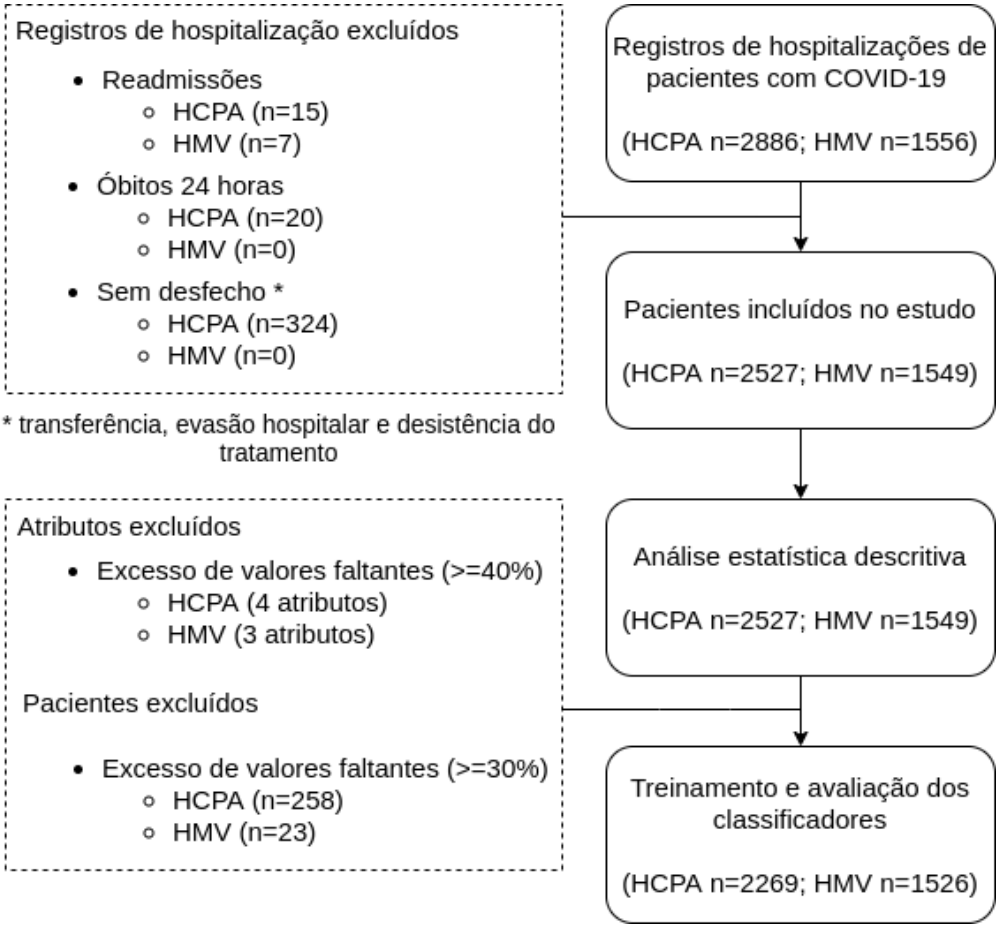


Figura 4.2 – Etapas de pré-processamento aplicadas nos conjuntos de dados
Fonte: O Autor

após a internação, 310 registros de hospitalização de pacientes que foram transferidos para outros hospitais e 6 registros de hospitalização de pacientes sem desfecho (3 pacientes por evasão e 3 por desistência do tratamento). Para o conjunto de dados HMV, os pacientes sem informação de desfecho foram desconsiderados durante o processo de transcrição dos dados. Assim, registros de hospitalização nos quais os pacientes foram transferidos para outros hospitais e/ou casos de evasão não foram incluídos no estudo.

Exclusão de variáveis e instâncias com valores faltantes: Variáveis com índice de valores faltantes superior à 40% foram removidas. No conjunto de dados HMV, a variável BNP, que refere-se à *brain natriuretic peptide* foi a única variável que atendeu o critério de exclusão. No conjunto de dados HCPA, as variáveis *peso*, *altura* e *tipo sanguíneo* foram excluídas. As variáveis *troponina* do conjunto de dados HMV e *fibrinogênio* do conjunto de dados HCPA foram as únicas exceções a esta regra. Embora as variáveis estivessem indisponíveis para, respectivamente, 44.67 e 41,87% dos registros, optamos por preservá-las, devido a evidências prévias quanto a sua importância como indicadores de severidade para a COVID-19 (BI et al., 2020; LIPPI; LAVIE; SANCHIS-GOMAR,

2020).

Registros de pacientes com índice de valores faltantes superior a 30% também foram removidos. Respectivamente, para conjuntos de dados HMV e HCPA, foram excluídos 23 (5 mortes e 18 altas) e 258 (93 mortes e 165 altas) registros de pacientes que ultrapassavam esse limiar.

Gestão de pacientes reinternados: Entre os registros de hospitalização abrangidos pelo período do estudo, havia pacientes com mais de uma passagem pelos hospitais. As bases de dados HMV e HCPA continham respectivamente 7 e 19 pacientes com duas passagens pelos hospitais. Desses pacientes, os 7 pacientes do HMV e 13 pacientes do HCPA foram re-hospitalizados dentro de um intervalo de 30 dias a contar a data da alta médica da hospitalização índice, enquanto que os outros 6 pacientes do HCPA foram re-hospitalizados após o intervalo de 30 dias.

Os pacientes re-hospitalizados dentro de um intervalo de 30 dias foram considerados como casos de readmissão. A adoção do período de 30 dias teve por base o critério adotado pela agência *Centers for Medicare & Medicaid Services* (CMS) para definir uma readmissão hospitalar[4].

Portanto, nos casos dos pacientes readmitidos (reinternação dentro de 30 dias), utilizamos apenas o desfecho hospitalar da readmissão para compor o registro do paciente. Ou seja, se na readmissão o paciente evoluiu para óbito, consideramos o paciente apenas como uma instância da classe *mortalidade=sim*, independentemente de ele ter recebido alta médica na hospitalização índice.

Tabela 4.1 – Exemplo meramente ilustrativo de um paciente com duas hospitalizações dentro de um intervalo de 30 dias

paciente	entrada	saída	D-dímeros	eritrócitos	pCO2	mortalidade
5f3823r	20/06/2021	01/07/2021	772	4,98	28	não
			...			
5f3823r	07/07/2021	23/07/2021	744	5,76	45	sim

Fonte: O Autor

A Tabela 4.1 apresenta um exemplo meramente ilustrativo de um paciente com duas hospitalizações, uma hospitalização índice e uma readmissão dentro de 30 dias à contar a data da alta médica da hospitalização índice. A entrada da hospitalização índice ocorreu na data 20/06/2021 e a saída na data 01/07/2021. Seis dias após sua saída, o paciente voltou a ser hospitalizado (07/07/2021). Nesta readmissão, o paciente teve óbito

[4] *Website* do programa de redução de readmissões hospitalares da CMS[5]

como desfecho.

Tabela 4.2 – Exemplo meramente ilustrativo de um paciente com duas hospitalizações dentro de um intervalo de 30 dias – instância resultante do pré-processamento

paciente	entrada	saída	D-dímeros	eritrócitos	PCO2	mortalidade
5f3823r	20/06/2021	23/07/2021	772	4,98	28	sim

Fonte: O Autor

A Tabela 4.2 apresenta o resultado do pré-processamento aplicado para pacientes readmitidos. A instância final para o paciente é constituído pelas informações da hospitalização índice (células em verde) e pelo desfecho da readmissão (células em vermelho).

Aplicamos este critério por interpretarmos que o quadro clínico desses pacientes pode não ser adequadamente representado pelo desfecho *mortalidade=sim*, visto que estes pacientes necessitaram ser readmitidos em um curto período de tempo e vieram à óbito. Embora o número de readmissões em relação ao número de registros dos conjuntos de dados seja pequeno, a utilização dessas hospitalizações como exemplos de pacientes pertencentes à classe *mortalidade=não*, poderia ser prejudicial para o processo de treinamento dos classificadores.

Para pacientes com reinternação ocorrida após 30 dias a contar a hospitalização índice, apenas os registros da hospitalização índice foram considerados e os registros das readmissões foram desconsiderados. Este critério foi adotado, pois tem-se conhecimento de pacientes que apresentam disseminação viral persistente, apresentando resultado positivo para RT-PCR, mesmo após a recuperação dos sintomas clínicos (ZHANG et al., 2020).

Após os ajustes e remoção dos registros que atenderam os critérios acima mencionados, 15 registros do conjunto de dados HCPA foram removidos ou agregados. Informações de 9 registros de readmissão foram agregados aos registros de suas respectivas hospitalizações índices e 6 registros de reinternações ocorridas após 30 dias a contar a hospitalização índice foram excluídos. Como no conjunto de dados HMV apenas 7 pacientes foram readmitidos, as informações dos 7 registros de readmissão foram agregados com os respectivos registros das hospitalizações índices.

4.1.2 Características dos Conjuntos de Dados

A Tabela 4.3 apresenta as características dos conjuntos de dados após o pré-processamento.

Tabela 4.3 – Estatísticas dos conjuntos de dados

	Registros	mortalidade=sim	mortalidade=não	cti=sim	cti=não	vmi=sim	vmi=não
HMV	1526	182	1344	458	1068	252	1274
HMV_{CTI}	458	143	315	458	0	252	206
HCPA	2269	849	1420	2269	0	1342	927

Fonte: O Autor

Dos 1526 registros que compõem o conjunto de dados HMV, em 458 (30%) os pacientes deram entrada na CTI, em 252 (16,51%) houve demanda por ventilação mecânica invasiva e, em 182 (11,93%) os pacientes morreram. Quando considerados os 458 pacientes que internaram na CTI, 142 morreram, os quais representam uma taxa de mortalidade de 9,30% em relação ao número total de pacientes do conjunto de dados e 31% em relação aos pacientes internados na CTI.

Do total de 2269 pacientes do conjunto de dados HCPA, os pacientes do sexo masculino representam 54,69% do conjunto de dados HCPA e a taxa de mortalidade é de 37,42%. A taxa de pacientes que necessitaram VMI é de 59,14%.

4.2 Análise descritiva dos dados

Para uma compreensão inicial dos dados disponíveis e dos principais fatores relacionados ao óbito, internação em CTI e necessidade de recursos de VMI, realizamos a análise estatística descritiva dos atributos disponíveis durante a admissão hospitalar (etapa 2 da Figura 4.1). Inicialmente, para as variáveis contínuas, aplicamos o teste de normalidade de *Shapiro-Wilk* (SHAPIRO; WILK, 1965). O nível de significância adotado para o teste foi de 95%.

Para todas as variáveis testadas, a aplicação do teste resultou em um *p-value* menor do que 0,05, dando subsídios para a rejeição da hipótese nula e, por consequência, a conclusão de que as variáveis não são normalmente distribuídas. Dessa forma, as variáveis contínuas foram descritas a partir de medianas e intervalos interquartis. As variáveis categóricas, foram resumidas a partir de contagens e valores percentuais.

O teste de *Kruskal-Wallis* foi utilizado para comparar as medianas entre as classes de cada variável, enquanto que as variáveis categóricas foram comparadas com os testes *Chi-squared* e *Fisher's exact test*. O nível de significância adotado para os testes foi de 95%, de forma que valores *p-values* menores do que 0,05 deram subsídios para rejeitar a hipótese nula e concluir que as distribuições são diferentes.

Os testes foram realizados estratificando os conjuntos de dados com base nas variáveis dependentes (*mortalidade*, *cti* e *vmi*). Os resultados da análise estatística descritiva são apresentados na Seção 5.3.

4.3 Seleção de atributos

A terceira etapa do nosso método consiste na seleção de atributos. Esta etapa tem como objetivo selecionar um subconjunto de atributos úteis para as tarefas de predição.

Para a seleção de atributos, utilizamos o método *Correlation based Feature Selection* (CFS). O procedimento de seleção de atributos foi realizado sobre os dados não imputados. Optamos por não imputar previamente os dados para preservar as características das distribuições dos dados das variáveis.

Após a seleção de atributos, os atributos com valores contínuos foram imputados com a mediana entre os valores observados. Atributos com valores categóricos foram imputados com valores de moda.

4.4 Classificação

A classificação é a quarta etapa do método proposto. As tarefas de predição de mortalidade, CTI e VMI, foram abordadas como problemas de classificação binária. Dessa forma, os eventos de interesse a serem identificados foram tratados como a classe positiva. Portanto, o valor sim das variáveis *mortalidade*, *cti* e *vmi* foram considerados como as classes positivas para as tarefas de predição.

Os algoritmos de classificação avaliados foram: (i) *Logistic Regression* (LR); (ii) J48; (iii) *Random Forest* (RF); (iv) *XGBoost* (XGB); (v) *AdaBoost* (AB) e; (vi) *Multi Layer Perceptron* (MLP). Selecionamos estes algoritmos por serem capazes de apresentar bom desempenho em tarefas de classificação para as quais há disponibilidade de um conjunto de dados de tamanho moderado, que julgamos refletir o caso desse estudo.

Os classificadores foram treinados com os conjuntos de treino e testados com os conjuntos de teste. Mais detalhes sobre o procedimento de reamostragem de dados utilizado para a composição dos conjuntos de treino e teste são apresentados na Seção 4.5.1.

Conforme revisão sistemática realizada em Gupta et al. (2020), dos 22 modelos para predição de prognóstico avaliados, nenhum apresentou resultados substancialmente

superiores quando comparado aos resultados obtidos a partir do uso isolado das variáveis *idade* e *saturação de O2* como variáveis preditoras para mortalidade e degradação do quadro clínico do paciente, respectivamente.

Com o objetivo de avaliar o desempenho dos classificadores nas tarefas propostas, treinamos versões desses classificadores utilizando apenas um atributo como variável independente. Para as tarefas de predição de mortalidade, treinamos classificadores que utilizaram apenas a idade como variável independente. Para a tarefa de predição de internação em CTI, treinamos classificadores que utilizaram apenas a idade, bem como apenas saturação de O2. Para a predição de necessidade de VMI, treinamos versões de classificadores de variável única que utilizaram as variáveis *saturação de O2* e *pCO2*.

A partir da estratificação dos conjuntos de dados pelas variáveis dependentes *mortalidade*, *cti* e *vmi*, é possível observar que as classes são desbalanceadas nos três casos. Para o conjunto de dados HMV a variável *mortalidade* apresenta o maior desbalanceamento entre as classes. Quando considerado todos os pacientes (internação regular + CTI) desbalanceamento é de aproximadamente 7,38 pacientes com *mortalidade=não* para cada paciente com desfecho *mortalidade=sim*.

As variáveis *vmi* e *cti* apresentam desbalanceamento de aproximadamente 5,05 e 2,33 vezes, respectivamente. Desta forma, proporcionalmente, a cada paciente que demandou recursos de VMI, existem 5 pacientes que não demandaram recursos de VMI e aproximadamente 2,33 pacientes que não foram internados em CTI.

Quando considerado o atributo alvo *mortalidade* apenas entre os pacientes que foram internados na CTI, o desbalanceamento do HMV é de 2,20 altas hospitalares para cada óbito. Para o conjunto de dados HCPA, o qual contém apenas pacientes internados na CTI, o desbalanceamento para a variável *mortalidade* é de 1,67.

Quando considerada a necessidade de suporte de VMI apenas entre os pacientes do HMV que foram internados na CTI, o desbalanceamento é de 1,22 vez, enquanto que para o HCPA é de 1,45 vez. Entretanto, nestes casos, a classe majoritária é a classe positiva (*vmi=sim*), visto que a maioria dos pacientes internados em CTI necessitaram suporte de VMI.

Para lidar com o desbalanceamento de classes, aplicamos a abordagem de matriz de custo (*Cost sensitive*) — que realiza a ponderação dos erros do classificador (falsos negativos ou falsos positivos) a depender da tarefa a ser abordada. Visto que a identificação dos eventos (*mortalidade=sim*, *cti=sim*, *vmi=sim*) foram considerados como classe positiva, aplicamos uma matriz de custo que penalizou os erros do tipo falso negativo. Dessa

forma, para as diferentes tarefas de predição (mortalidade, CTI e VMI) o erro cometido pelo classificador ao não identificar os eventos (a morte, a necessidade de internação em CTI e a necessidade de utilização de recursos de VMI) foi mais penalizado do que o erro cometido ao atribuir pacientes que não experienciaram a ocorrência dos eventos à classe positiva.

As matrizes de custo aplicadas levaram em consideração o desbalanceamento entre as classes, de forma que a penalização para os erros falsos negativos foi definida como:

$$pen = \frac{nc\_maj}{nc\_min}, \tag{4.1}$$

onde nc_maj é o número de instâncias da classe majoritária e nc_min é o número de instâncias da classe minoritária. Para os experimentos que avaliaram a necessidade de VMI, a classe majoritária é a classe positiva.

4.5 Avaliação

A quinta etapa do método consiste na avaliação dos classificadores. Esta etapa contempla a avaliação da eficácia dos classificadores e suas capacidades de generalização nos conjuntos de teste. As métricas adotadas durante a avaliação dos modelos foram previamente apresentados na Seção 2.7. A próxima seção apresenta a técnica de reamostragem utilizada para a avaliação dos classificadores.

4.5.1 Reamostragem Bootstrap

Fizemos uso da técnica de reamostragem *Bootstrap* (EFRON; TIBSHIRANI, 1994) para a avaliação dos modelos preditivos. Para cada conjunto de dados, aplicamos B iterações de reamostragem com repetição, onde $B = 100$. Para cada iteração, um número de instâncias — igual ao número de instâncias do conjunto de dados original — foi aleatoriamente selecionado para compor o conjunto de treino e o restante das instâncias não incluídas no conjunto de treino foram selecionadas para constituir o conjunto de teste. Esta estratégia é conhecida como *bootstrap out-of-bag*. Dessa forma, para cada uma das B iterações, treinou-se os classificadores com o conjunto de treino e, a partir do conjunto de teste, calculou-se as métricas de eficácia. Ao fim do processo, as métricas foram agregadas e expressas a partir de medias e intervalos de confiança.

4.6 Explicabilidade

Esta seção descreve a abordagem proposta para auxiliar na explicação de quais variáveis são levadas em consideração pelo modelo subjacente durante a inferência. O objetivo é extrair *insights* sobre a relação entre as variáveis levadas em consideração pelo modelos preditivos e suas previsões corretas e incorretas. A abordagem proposta corresponde ao passo seis, da Figura 4.1 e seus componentes são mostrados na Figura 4.3. A ideia é semelhante ao *clustering* supervisionado (LUNDBERG; ERION; LEE, 2018), no qual os valores de SHAP (não valores brutos dos atributos) são usados para agrupar instâncias. No entanto, em vez de agrupar, nós os usamos para gerar representações visuais para inspecionar previsões.

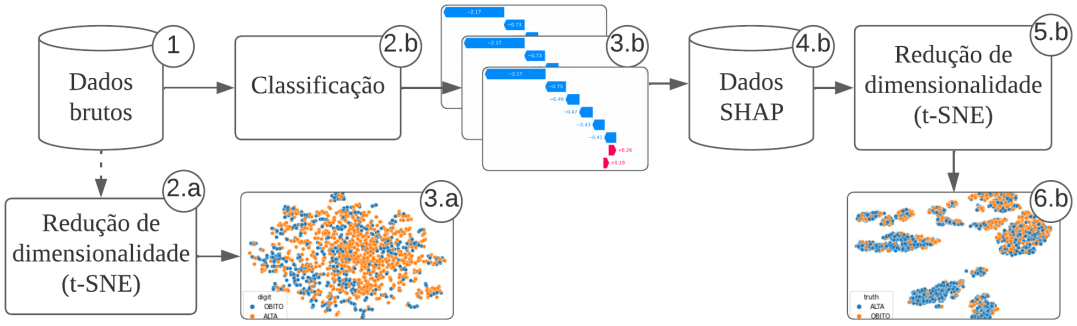


Figura 4.3 – Etapas aplicadas na explicação dos resultados retornados pelo modelo preditivo
Fonte: O Autor

Primeiro, partindo de um conjunto de dados genérico D que neste caso poderia ser qualquer um dos nossos conjuntos de dados, usamos a redução de dimensionalidade para entender aspectos de separabilidade linear dos dados. Espera-se que nos dados brutos não haja padrões claros de separabilidade. Os retângulos *2.a* e *3.a* representam a visualização dos dados brutos usando uma projeção t-SNE (MAATEN; HINTON, 2008). É possível observar que a disposição das instâncias não constitui agrupamentos, sugerindo pouca linearidade nos dados.

A partir do mesmo conjunto de dados D, a abordagem proposta consiste nas seguintes etapas: (*2.b*) treinar um algoritmo de classificação e (*3.b*) aplicar a abordagem de explicação SHAP sobre as previsões. O processo de treinamento em *2.b* segue as práticas padrão, como manipulação de valores ausentes, aplicação de técnicas para lidar com desequilíbrio de classe e avaliação de modelo. Em *3.b*, selecionamos instâncias do conjunto de teste e as submetemos à abordagem SHAP. Obtemos, para cada valor da variável v da instância i, seus valores de contribuição (valores SHAP). Em nossos experimentos,

Tabela 4.4 – Exemplo de um conjunto de dados *D* com quatro instâncias i e o respectivo conjunto de dados *C* com quatro instâncias i' representadas pelos *shapley values*

Conjunto de dados *D*			Conjunto de dados *C*		
a	*b*	*c*	*a'*	*b'*	*c'*
82	843	0,25	0,50	-0,80	2,25
75	640	0,30	0,20	2,00	1,00
26	450	0,10	-1,30	0,20	-0,03
65	560	0,14	0,30	0,35	-2,51

Fonte: O Autor

usamos KernelSHAP para estimar valores SHAP, mas outros métodos também poderiam ter sido usados. Com os valores SHAP, criamos uma instância i', que consiste nos valores SHAP calculados, em vez dos valores brutos do atributo. Na próxima etapa (*4.b*), criamos um conjunto de dados C constituído pelas instâncias i'.

A Tabela 4.4 apresenta um exemplo de um conjunto de dados D genérico e o respectivo conjunto de dados C formados pelos *shapley values*. O conjunto de dados D contém quatro instâncias, cada uma representada por três atributos, *a*, *b* e *c*. O conjunto de dados *C* contém as instâncias constituídas pelos *shapley values* estimados para cada atributo de cada uma das instâncias do conjunto de dados D.

Finalmente, na última etapa, a redução de dimensionalidade é aplicada às instâncias de valores SHAP em C para gerar um gráfico 2D que fornece informações sobre a separabilidade das instâncias. Este procedimento é representado pelo passo *5.b*. Aplicamos a técnica de redução de dimensionalidade t-SNE em nossos testes, mas outras técnicas de redução de dimensionalidade podem ser usadas. A representação em *6.b* é o resultado da projeção em 2D gerada pelo t-SNE.

É possível observar que as instâncias contidas na projeção podem formar agrupamentos bem definidos. Por exemplo, certas regiões podem ter prevalência de uma classe específica, enquanto outras podem apresentar maior incerteza. Além disso, os pontos projetados podem ser coloridos com base nos valores dos atributos, o que possibilita obter *insights* importantes, como a relação entre os atributos consideradas pelos modelos preditivos e suas previsões corretas e incorretas.

4.7 Resumo do Capítulo

Este capítulo apresentou a metodologia adotada em nosso trabalho. Apresentamos as etapas que à constituem, bem como as escolhas realizadas e suas respectivas justificativas. Além das etapas padrões do *pipeline* de AM, descrevemos a técnica de visualização proposta para auxiliar na interpretação dos modelos preditivos. O próximo capítulo apresenta os experimentos executados e resultados.

5 EXPERIMENTOS

Neste capítulo, apresentamos os experimentos realizados para responder as seguintes questões de pesquisa:

- *Q1* – É possível identificar, entre os pacientes com COVID-19, quais têm maior probabilidade de falecer devido à doença? Em outras palavras, a previsão de mortalidade pode ser realizada com acurácia com base nas informações disponíveis no momento da admissão?
- *Q2* – É possível identificar, entre os pacientes com COVID-19, aqueles que têm maior probabilidade de necessitar leito de CTI?
- *Q3* – É possível identificar, entre os pacientes com COVID-19, aqueles que têm maior probabilidade de necessitar recursos de VMI?
- *Q4* – Os modelos treinados em um conjunto de dados podem ser utilizados para fazer predições sobre pacientes de outro conjunto de dados?
- *Q5* – Quais atributos foram considerados mais importantes para as tarefas de predição avaliadas?
- *Q6* – O desempenho obtido com os classificadores de múltiplas variáveis é significativamente superior ao desempenho obtido por classificadores de variável única?

5.1 Cenários de Avaliação

A avaliação dos classificadores foi conduzida levando-se em consideração dois cenários.

- *C1* – os dados de todos os pacientes foram considerados (internação regular + CTI)
- *C2* – apenas pacientes internados em CTI foram considerados.

Em ambos os casos, nos concentramos nas informações disponíveis durante a admissão hospitalar. A definição dos cenários *C1* e *C2* objetivam a distinção para as avaliações sobre o conjunto de dados HMV, visto que o conjunto de dados HCPA apenas contém pacientes internados na CTI (por definição, só pertence ao cenário *C2*). Ou seja, o cenário *C2* para o conjunto de dados HMV é conduzido sobre o subconjunto HMV_{CTI}. Além do mais, visto que no cenário *C2* considera-se apenas pacientes internados em CTI, a tarefa de predição de CTI não é avaliada neste cenário.

5.2 Configuração dos Experimentos

Esta seção detalha os recursos utilizados nos experimentos e a configuração das execuções.

5.2.1 Ferramentas

O pré-processamento dos dados foi realizado com o uso da linguagem de programação *Python v3.5* e das bibliotecas *Pandas* e *Numpy*. A análise descritiva das variáveis foi realizada com a biblioteca *tableone v0.7.10* (POLLARD et al., 2018).

A modelagem dos classificadores foi realizada sobre as implementações disponibilizadas pela *API Weka v3.9.6* através do *wrapper python-weka-wrapper3 v0.2.9* — com exceção do algoritmo de *Gradient Boosting Tree*, que utilizou a implementação *XGBoost*. A imputação de valores para os atributos contínuos foi realizada com as medianas dos valores observados em cada atributo. Atributos com valores categóricos foram imputados com los valores de moda. *XGBoost* pode lidar com valores faltantes, então os experimentos rodados com *XGBoost* utilizaram as versões dos conjuntos de dados sem imputação de valores faltantes. Para a interpretação dos modelos gerados, utilizamos a biblioteca *Alibi* [1], a qual disponibiliza uma função denominada *KernelSHAP*, que encapsula a função *KernelExplainer* (LUNDBERG; LEE, 2017) da biblioteca *SHAP*, disponibilizando utilitários que facilitam o seu uso. Utilizamos a abordagem *KernelExplainer* por ser agnóstica no que se refere aos tipos de algoritmos compatíveis.

5.2.2 Algoritmos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina avaliados neste trabalho foram: *Logistic Regression* (LR), C4.5 (J48), *Random Forest* (RF), *AdaBoost* (AB), *Multi Layer Perceptron* (MLP) e *XGBoost* (XGB). Não comparamos os algoritmos de AM com escores de riscos tradicionais (como MEWS, NEWS, REMS, *etc*), pois nem todos os atributos utilizados por esses escores estavam disponíveis em nossos conjuntos de dados, o que não possibilitaria comparações justas.

[1] *Github da biblioteca Alibi* <https://docs.seldon.io/projects/alibi/en/stable/index.html>

5.2.3 Execuções

Para a avaliação dos modelos, realizamos 100 iterações de reamostragem de dados com repetição (*bootstrap*). Os algoritmos de classificação foram treinados a partir dos conjuntos de treino e, avaliados com base nas instâncias dos conjuntos de teste. Vale ressaltar que os conjuntos de treino e teste são disjuntos. Assim, diferentes instâncias constituem os conjuntos de treino e teste.

Treinamos um modelo extra para cada tarefa para realizar a análise das variáveis que apresentam maior contribuição durante as predições. Ao invés de adotar o processo de reamostragem *bootstrap*, adotamos a abordagem *hold-out*. Selecionamos aleatoriamente 60% e 40% das instâncias para compor os conjuntos de treino e teste, respectivamente. Após o treinamento dos modelos, executamos a abordagem SHAP para obter os valores de contribuição (*shapley values*) para as instâncias do conjunto de teste.

Para o experimento que avalia a questão *Q4* (na Seção 5.4.4), foram consideradas apenas as variáveis contidas em ambos conjuntos de dados (HMV_{CTI} e HCPA). Quando necessário, as variáveis foram convertidas para que os atributos dos dois conjuntos de dados estivessem expressos nas mesmas unidades de medida. Realizamos avaliações cruzadas (avaliação externa), de forma que cada conjunto de dados foi utilizado uma vez como conjunto de treino e outra como conjunto de teste.

5.3 Análise descritiva dos dados

Esta seção refere-se à etapa 2 do método descrito na Figura 4.1 e apresenta as principais características dos conjuntos de dados HMV e HCPA, bem com a análise descritiva dos dados.

A idade é o atributo com as diferenças mais evidentes entre as classes. Considerando todos os pacientes (internação regular + CTI) do conjunto de dados HMV a mediana é de 61 anos para os pacientes da classe *mortalidade=não* e 85 anos para a classe *mortalidade=sim* (Apêndice B.1). Os pacientes internados na CTI tendem a ser mais velhos, com uma idade mediana de 71 anos, 11 anos a mais que os pacientes que não foram admitidos na CTI (Apêndice B.3). Os pacientes internados na CTI que necessitaram VMI têm uma idade mediana de 73 anos, enquanto que os pacientes que não necessitaram VMI têm uma idade mediana de 61 anos (Apêndice B.5).

Além da idade, há outros atributos que recorrentemente apresentaram diferença

significativa entre as classes e que estão relacionado à prognósticos desfavoráveis. Dentre eles destacam-se a diminuição de valores de eritrócitos, hemoglobina, hematócritos e linfócito, bem como, o aumento dos leucócitos, neutrófilos, pCO2, troponina, tempo de protrombina, LDH, D-dímeros, aspartato e proteína C-reativa (Apêndices [B.1 – C.2]).

A Tabela 5.1 apresenta os principais atributos com diferença significativa ao estratificar os pacientes internados na CTI (HCPA e HMV_{CTI}) entre as classes *mortalidade=sim* e *mortalidade=não*. Os pacientes do conjunto de dados HCPA são mais jovens do que os pacientes do conjunto de dados HMV que internaram na CTI. Enquanto a idade mediana dos pacientes do conjunto de dados HMV que internaram na CTI é de 71 anos, a idade mediana dos pacientes do conjunto de dados HCPA é de 60 anos.

Tabela 5.1 – Comparação entre os conjuntos de dados HCPA e HMV_{CTI} - Principais atributos com diferença significativa.

		HCPA	HMV_{CTI}	P-Value
n		2269	458	
idade		60.0 [47.0,70.0]	71.0 [58.2,81.8]	0.001
plaquetas		214000.0 [163000.0,273000.0]	190500.0 [141000.0,257250.0]	0.001
leucócitos		8800.0 [6450.0,12180.0]	7845.0 [5582.5,10635.0]	0.001
neutrófilos		7145.0 [4960.0,10427.5]	6070.0 [3977.5,8720.0]	0.001
linfócitos		790.0 [540.0,1150.0]	900.0 [630.0,1260.0]	0.001
monócitos		420.0 [260.0,650.0]	510.0 [340.0,787.5]	0.001
pO2		86.2 [66.6,121.8]	74.0 [65.0,92.0]	0.001
troponina		10.3 [9.9,40.5]	11.0 [6.0,33.2]	0.001
DHL		469.0 [340.0,642.5]	514.5 [389.2,703.0]	0.001
D-dímeros		1180.0 [670.0,2570.0]	927.0 [537.0,1711.0]	0.001
prot. C-reativa		13.3 [7.5,20.8]	7.6 [3.4,14.1]	0.001
mortalidade	não	1420 (62.6)	315 (68.8)	0.014
	sim	849 (37.4)	143 (31.2)	
vmi	não	927 (40.9)	206 (45.0)	0.114
	sim	1342 (59.1)	252 (55.0)	

Fonte: O Autor

Além da diferença de idade dos pacientes, vários atributos apresentam diferenças consideráveis entre as populações dos conjuntos de dados HMV_{CTI} e HCPA.

Diferente do que ocorre no conjunto de dados HMV, não se observa diferença significativa na idade mediana do grupo de pacientes do conjunto de dados HCPA que necessitaram de VMI. Tanto a classe *vmi=sim* quanto a classe *vmi=não* apresentam idade mediana de 60 anos (Apêndice C.3). Em relação à predição de mortalidade para o conjunto de dados HCPA, observa-se uma diferença de 12 anos entre as medianas das classes

mortalidade=não (55 anos) e *mortalidade=sim* (67 anos) (Apêndice C.1).

Vários atributos apresentam diferenças entre os conjuntos de dados HMV$_{CTI}$ e HCPA. Esta heterogeneidade observada nos dados gerados pelos hospitais é destacada por Futoma et al. (2020) que contrapõem o recorrente hábito de, durante a avaliação dos modelos, enfatizar demasiadamente a capacidade de generalização dos modelos preditivos entre diferentes hospitais e protocolos clínicos. Além disso, os autores argumentam que a falta de generalização de um modelo desenvolvido em uma unidade hospitalar H_A, quando aplicado em uma unidade hospitalar H_B, não diminui a sua utilidade na unidade para a qual foi desenvolvido.

Na Seção 5.4.4 realizamos avaliações cruzadas entre os conjuntos de dados HMV$_{CTI}$ e HCPA – também denominadas de validação externa.

5.4 Resultados

Esta seção apresenta os resultados e respostas para as questões de pesquisa anteriormente definidas. Em nossas análises, enfatizamos os resultados de sensibilidade, pois o erro de não identificar pacientes como pertencentes às classes *mortalidade=sim*, *cti=sim* e *vmi=sim* é muito mais grave do que o de prever que o paciente tem alta probabilidade de óbito, necessitar de CTI e VMI.

5.4.1 É possível identificar, entre os pacientes com COVID-19, quais têm maior probabilidade de morrer devido à doença?

Para responder esta questão de pesquisa, avaliamos os algoritmos de classificação nos cenários *C1* (hospitalização regular + CTI) e *C2* (somente CTI). A Tabela 5.2 apresenta os resultados médios obtidos por cada algoritmo para os dois cenários avaliados. Os resultados médios com seus intervalos de confiança podem ser consultados na Tabela D.1 dos apêndices. O desempenho atingido pelos algoritmos varia em cada cenário a depender da métrica. Não há um algoritmo único que apresente desempenho superior para todas as métricas.

Considerando a sensibilidade, o algoritmo LR é o vencedor, atingindo os maiores valores em todos os cenários (para o conjunto de dados HCPA o algoritmo LR empatou com o XGB). AB apresentou o segundo maior valor de sensibilidade para *C1* e HMV$_{CTI}$

Tabela 5.2 – Resultados para a predição de mortalidade

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**0,84**	0,87	0,45	**0,98**	0,59	0,75	0,52	0,92	0,60
		J48	0,54	0,92	0,46	0,94	0,50	0,71	0,43	0,73	0,44
		RF	0,57	**0,95**	**0,59**	0,94	0,58	0,77	0,53	**0,93**	**0,61**
		AB	0,78	0,85	0,41	0,97	0,53	0,72	0,45	0,89	0,52
		MLP	0,61	0,92	0,52	0,95	0,56	0,75	0,50	0,87	0,57
		XGB	0,71	0,91	0,52	0,96	**0,60**	**0,77**	**0,54**	0,92	0,60
$C2$	$HCPA$	LR	**0,68**	0,73	0,60	0,79	0,64	0,70	0,40	0,77	0,65
		J48	0,56	0,70	0,53	0,72	0,54	0,63	0,25	0,63	0,53
		RF	0,62	**0,77**	**0,62**	0,78	0,62	0,70	0,40	0,78	0,66
		AB	0,67	0,71	0,59	0,78	0,62	0,69	0,37	0,76	0,64
		MLP	0,67	0,71	0,58	0,78	0,61	0,68	0,36	0,75	0,61
		XGB	**0,68**	0,73	0,61	**0,80**	**0,65**	**0,70**	**0,41**	**0,79**	**0,67**
$C2$	HMV_{CTI}	LR	**0,80**	0,82	0,66	**0,90**	**0,72**	**0,78**	**0,57**	**0,87**	**0,73**
		J48	0,67	0,80	0,60	0,84	0,63	0,72	0,44	0,73	0,59
		RF	0,76	**0,83**	**0,67**	0,88	0,71	0,78	0,56	0,87	0,69
		AB	0,77	0,80	0,63	0,89	0,70	0,76	0,53	0,86	0,69
		MLP	0,69	**0,83**	0,65	0,86	0,67	0,75	0,50	0,83	0,69
		XGB	0,73	**0,83**	0,67	0,88	0,69	0,77	0,54	0,87	0,71

Fonte: O Autor

e o terceiro maior valor para HCPA. Em todos os casos, J48 atingiu os valores de sensibilidade mais baixos.

Considerando o cenário *C1* do HMV, o modelo LR obteve sensibilidade de 0,84 e VPP de 0,45. Ou seja, o classificador foi capaz de identificar 84% do total de pacientes que evoluíram para óbito e, entre os pacientes que o classificador atribuiu a classe *mortalidade=sim*, em 45% das vezes os pacientes realmente faleceram. A especificidade de 0,87 informa que 87% dos pacientes que receberam alta foram corretamente classificados como pertencentes à classe *mortalidade=não*, ao instante em que o VPN de 0,98 informa que o classificador atribuiu corretamente os pacientes à classe *mortalidade=não* em 98% das vezes.

Também é possível observar que embora a sensibilidade e VPP tenham variado consideravelmente entre os diferentes classificadores, tanto a métrica F1+ quanto a métrica ma-F1 mantiveram-se relativamente estáveis. Esse comportamento deve-se ao fato de que essas métricas sumarizam, a partir de valores médios, diferentes capacidades dos algoritmos em classificar pacientes da classe *mortalidade=sim* e *mortalidade=não*. Isso justifica que tanto LR, que atingiu sensibilidade de 0,84 e VPP de 0,45, quanto RF e MLP, que atingiram respectivamente, sensibilidade de 0,57 e 0,61 e, VPP de 0,59 e 0,52, tenham apresentado métricas de F1+ e ma-F1 parecidas.

Em resposta à questão de pesquisa *Q1*, de modo geral e considerando a priori-

zação por maiores valores de sensibilidade, os algoritmos LR, AB e XGB apresentaram resultados promissores ao identificar pacientes da classe *mortalidade=sim*, enquanto que MLP, RF e J48 apresentaram eficácia inferior na detecção dos pacientes que morreram.

Os resultados obtidos são próximos aos reportados em trabalhos que avaliaram a tarefa de predição de mortalidade. Em termos de AUROC, F1+ e ma-F1, o classificador LR atingiu resultados semelhantes ao modelo *ensemble* de (PAIVA et al., 2021). O escore REMS avaliado por (COVINO et al., 2020) atingiu sensibilidade de 0,90 e 0,96 para os cenários que consideraram mortes ocorridas dentro de 48 horas e 7 dias, respectivamente. É importante ressaltar que não é possível realizar a comparação direta com os demais trabalhos da literatura visto que os conjuntos de dados utilizados diferem dos utilizados neste trabalho.

5.4.2 É possível identificar, entre os pacientes com COVID-19, aqueles que têm maior probabilidade de necessitar leito de CTI?

Com o objetivo de responder esta questão de pesquisa, utilizamos o conjunto de dados HMV. Não utilizamos o conjunto HCPA visto que o mesmo contempla apenas pacientes já internados na CTI. A Tabela 5.3 apresenta os resultados médios atingidos pelos classificadores para a predição de CTI. A Tabela de apêndice D.2 apresenta os intervalos de confiança para as métricas.

Nenhum algoritmo apresentou os melhores resultados para todas as métricas.

Tabela 5.3 – Resultados para a predição de CTI

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**0,66**	0,77	0,55	**0,84**	**0,60**	**0,70**	0,40	0,78	**0,63**
		J48	0,52	0,74	0,46	0,78	0,49	0,62	0,25	0,62	0,48
		RF	0,53	**0,86**	**0,62**	0,81	0,57	**0,70**	**0,41**	**0,79**	**0,63**
		AB	0,61	0,75	0,51	0,82	0,55	0,67	0,34	0,74	0,57
		MLP	0,58	0,77	0,52	0,81	0,55	0,67	0,34	0,73	0,59
		XGB	0,60	0,80	0,57	0,83	0,59	**0,70**	0,40	0,78	0,62

Fonte: O Autor

O algoritmo LR alcançou a maior sensibilidade (0,66), seguido pelos algoritmos AB e XGB, que atingiram sensibilidade de 0,61 e 0.60, respectivamente. O índice de sensibilidade atingido significa que o algoritmo LR foi capaz de identificar 66% entre todos os pacientes que necessitaram de CTI. Em contrapartida, o algoritmo apresentou o terceiro melhor índice de especificidade e VPP. A especificidade refere-se à fração de

pacientes que foram corretamente identificados como pacientes que não necessitaram de internação na CTI. O VPP é a precisão apresentada pelo algoritmo ao classificar os pacientes na classe *cti=sim*. Ou seja, VPP é a fração de pacientes que realmente necessitou de CTI entre aqueles pacientes que o classificador atribuiu como classe *cti=sim*. Em termos de especificidade, o algoritmo LR atingiu índices inferiores em 9 e 3 pontos percentuais (pp) quando comparado aos algoritmos RF e XGB, que atingiram os maiores índices de especificidade (0,86 e 0,80). Entretanto, os ganhos de especificidade atingidos por esses algoritmos são inferiores às perdas em termos de sensibilidade, visto que RF e XGB apresentaram índices de sensibilidade superiores em 13 pp e 6 pp, respectivamente.

Comportamento semelhante é observado em termos de VPP. Como consequência de classificar mais pacientes na classe positiva (*cti=sim*), o algoritmo LR apresentou menor VPP (precisão) entre os pacientes atribuídos a essa classe quando comparado aos algoritmos RF e XGB. Entretanto, embora apresente VPP inferior de 7 pp e 2 pp em relação aos algoritmos RF e XGB, o VPP de 0,55 atingido pelo algoritmo LR pode ser considerado promissor, dados os índices superiores de sensibilidade. Assim como para a tarefa de predição de mortalidade, o algoritmo J48 teve desempenho inferior aos demais algoritmos na predição de pacientes que necessitaram internação em CTI. Em termos de sensibilidade, os resultados atingidos em nosso trabalho são semelhantes aos outros estudos mencionados nos trabalhos relacionados. Em comparação ao escore de risco NEWS avaliado por (COVINO et al., 2020), os nossos resultados são inferiores em 14,2 pp e 5,4 pp, visto que no estudo os autores avaliaram o escore considerando as internações ocorridas dentro das primeiras 48 horas e nos 7 primeiros dias, respectivamente.

Conforme o relato de (ZHAO et al., 2020), o algoritmo *logistic regression* atingiu sensibilidade de aproximadamente 0,10, enquanto que a sensibilidade atingida pelo algoritmo LR em nosso trabalho foi de 0,66. Conforme argumentado por (ZHAO et al., 2020), a baixa sensibilidade atingida deveu-se ao desbalanceamento de classe observado no conjunto de dados utilizado no estudo. Entretanto, os autores não mencionam a adoção de técnicas para lidar com o desbalanceamento. Dessa forma, além das diferenças intrínsecas de cada conjunto de dados, o fato de termos aplicado a técnica *cost-sensitive* para atenuar os efeitos indesejados do desbalanceamento de classe, são fatores que podem justificar a diferença entre os níveis de sensibilidade atingidos. É importante ressaltar que embora tenha atingido sensibilidade de aproximadamente 0,10, o classificador de (ZHAO et al., 2020) atingiu uma AUROC de 0,74, contra a AUROC de 0,78 atingida pelo algoritmo LR em nosso trabalho. Mesmo com uma diferença de praticamente 50 pp em termos de

sensibilidade, a diferença em termos de AUROC é de aproximadamente 4 pp. Isso mostra o quão limitante pode ser a comparação de diferentes modelos preditivos ao adotar apenas a AUROC como métrica avaliativa (a Seção 5.5.3 estende a discussão acerca da métrica AUROC e suas limitações). Em resposta à questão de pesquisa *Q2*, pode-se dizer que foram atingidos resultados promissores no que se refere à capacidade de identificação de pacientes que internaram em CTI.

5.4.3 É possível identificar, entre os pacientes com COVID-19, aqueles que têm maior probabilidade de necessitar recursos de VMI?

Avaliamos a tarefa de predição de VMI nos cenários (*C1* e *C2*). A Tabela 5.4 apresenta os resultados médios alcançados pelos algoritmos. Os resultados com intervalos de confiança podem ser consultados na Tabela D.3 dos apêndices.

Tabela 5.4 – Resultados para a predição de necessidade de VMI

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**0,67**	0,77	0,36	**0,92**	**0,47**	0,65	**0,33**	**0,80**	**0,47**
		J48	0,40	0,83	0,32	0,88	0,35	0,60	0,21	0,61	0,32
		RF	0,35	**0,93**	**0,50**	0,88	0,41	**0,66**	0,32	0,79	0,45
		AB	0,61	0,74	0,32	0,91	0,42	0,62	0,26	0,74	0,39
		MLP	0,54	0,82	0,37	0,90	0,43	0,64	0,30	0,74	0,43
		XGB	0,53	0,85	0,40	0,90	0,46	**0,66**	**0,33**	0,78	0,44
$C2$	$HCPA$	LR	**0,89**	0,24	0,63	0,60	0,74	0,53	0,14	0,70	0,78
		J48	0,70	**0,49**	0,66	0,53	0,68	0,59	0,19	0,61	0,69
		RF	0,81	0,46	**0,69**	0,63	0,74	**0,64**	**0,28**	0,73	0,81
		AB	0,87	0,33	0,65	**0,64**	0,74	0,58	0,21	0,71	0,80
		MLP	0,80	0,38	0,66	0,58	0,72	0,57	0,19	0,69	0,77
		XGB	0,83	0,44	0,68	**0,64**	**0,75**	0,63	**0,28**	**0,74**	**0,82**
$C2$	HMV_{CTI}	LR	**0,73**	0,40	**0,61**	**0,55**	**0,65**	0,55	**0,13**	**0,62**	**0,68**
		J48	0,64	0,44	0,59	0,50	0,61	0,54	0,09	0,55	0,64
		RF	0,70	0,43	**0,61**	0,53	**0,65**	**0,56**	**0,13**	0,61	0,66
		AB	0,72	0,40	0,60	0,54	**0,65**	0,55	0,12	0,60	0,66
		MLP	0,63	**0,47**	**0,61**	0,51	0,61	0,54	0,11	0,59	0,66
		XGB	0,68	0,43	0,60	0,52	0,63	0,55	0,10	0,59	0,65

Fonte: O Autor

Assim como nas demais tarefas, nenhum algoritmo alcançou os melhores resultados para todas as métricas. Em termos de sensibilidade, o algoritmo LR recorrentemente apresentou os melhores resultados, seguido pelo algoritmo AB. De modo geral, o cenário *C2* obteve escores superiores de sensibilidade e VPP. Este resultado pode ser considerado intuitivo, visto que a proporção de pacientes que necessitaram de recursos de VMI é maior no cenário *C2*, já que este contempla apenas pacientes internados na CTI. Portanto,

ao considerar o *trade-off* entre sensibilidade e VPP, o cenário *C2* foi o que resultou em maior equilíbrio entre as métricas, com destaque para o conjunto de dados HCPA.

Para o conjunto de dados HCPA, exceto o algoritmo J48, todos os demais algoritmos atingiram sensibilidade igual ou superior a 0,80, ao instante em que atingiram níveis de VPP iguais ou superiores a 0,63. Quando considerado o cenário *C2*, tanto para os escores de sensibilidade quanto de VPP, os maiores valores foram observados no conjunto de dados HCPA.

Como resposta à questão de pesquisa *Q3*, pode-se dizer que os classificadores apresentaram resultados promissores para a tarefa de predição de necessidade de VMI. Assim como para as outras tarefas, a comparação direta entre os resultados não é possível. Ainda assim, os resultados obtidos em nossos experimentos são semelhantes aos valores reportados por outros trabalhos da literatura. Em termos de AUROC, os escores obtidos em nossos experimentos variaram de 0,55 à 0,80. Ao considerar o maior valor de AUROC (0,80), este valor é superior aos resultados reportados por (YU et al., 2021) e (RODRIGUEZ et al., 2021) e menor do que os resultados reportados por (BURDICK et al., 2020).

5.4.4 Os modelos treinados a partir do conjunto de dados de um hospital podem ser utilizados em pacientes de outro hospital?

Este experimento tem como objetivo avaliar a capacidade de generalização dos modelos em conjuntos de dados diferentes dos quais foram treinados. Para isso, avaliamos as tarefas de predição de mortalidade e necessidade de VMI no cenário *C2*, visto que esse é o cenário acessível tanto para o conjunto de dados HCPA quanto para o conjunto de dados HMV_{CTI}. Cada conjunto de dados foi utilizado para treino e teste de forma disjunta. A predição de CTI não foi avaliada nesse experimento, visto que o conjunto de dados HCPA apenas contém pacientes internados em CTI.

A Tabela 5.5 apresenta os resultados obtidos pelos classificadores para a tarefa de predição de mortalidade. Os resultados obtidos pelos classificadores variaram consideravelmente ao alterar os conjuntos de dados utilizados como treino e teste. Os algoritmos treinados com o conjunto de dados HCPA apresentaram resultados promissores quando testados no conjunto de dados HMV_{CTI}. O menor desempenho em termos de sensibilidade foi atingido pelo algoritmo RF (0,59), enquanto que a maior sensibilidade foi obtida com o algoritmo LR (0,91). Entretanto, quando treinados com o conjunto de

Tabela 5.5 – Resultados para predição de mortalidade - validação cruzada entre os conjuntos de dados HMV$_{CTI}$ e HCPA

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	F1$_+$	ma-F1	Kappa	AUROC	AUPRC
$C2$	$HCPA(treino)$ $HMV_{CTI}(teste)$	LR	**0,91**	0,66	0,55	**0,94**	**0,68**	**0,73**	**0,48**	**0,88**	**0,77**
		J48	0,59	0,66	0,44	0,78	0,50	0,61	0,22	0,63	0,49
		RF	0,73	**0,76**	**0,58**	0,86	0,64	0,72	0,45	0,82	0,62
		AB	0,83	0,63	0,51	0,90	0,62	0,68	0,39	0,80	0,60
		MLP	0,79	0,67	0,52	0,88	0,63	0,69	0,40	0,80	0,64
		XGB	0,86	0,62	0,51	0,91	0,64	0,69	0,40	0,81	0,63
$C2$	$HMV_{CTI}(treino)$ $HCPA(teste)$	LR	**0,48**	0,83	0,63	**0,73**	**0,54**	**0,66**	**0,32**	0,73	**0,60**
		J48	0,40	0,78	0,54	0,68	0,45	0,58	0,19	0,58	0,52
		RF	0,39	0,87	0,64	0,71	0,48	0,63	0,28	**0,74**	**0,60**
		AB	0,34	**0,89**	**0,65**	0,69	0,44	0,61	0,25	0,72	0,60
		MLP	0,39	0,85	0,60	0,70	0,47	0,62	0,25	0,69	0,58
		XGB	0,38	0,87	0,64	0,70	0,47	0,62	0,27	0,73	**0,60**

Fonte: O Autor

dados HMV$_{CTI}$ e testados com o conjunto HCPA, os algoritmos alcançaram índices de sensibilidade consideravelmente inferiores, sendo que o maior índice foi alcançado pelo algoritmo LR (0,48). Os demais algoritmos atingiram valores não superiores a 0,40, sendo que o menor valor de sensibilidade foi atingido pelo AB (0,34). Além disso, é possível observar que neste caso, os algoritmos apresentaram tendência de atribuir os pacientes à classe *mortalidade=não*, o que resultou em valores superiores de especificidade.

Tabela 5.6 – Resultados para predição de VMI - validação cruzada entre os conjuntos de dados HMV$_{CTI}$ e HCPA

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	F1$_+$	ma-F1	Kappa	AUROC	AUPRC
$C2$	$HCPA(treino)$ $HMV_{CTI}(teste)$	LR	0,75	0,32	0,57	**0,51**	**0,65**	**0,52**	**0,07**	**0,58**	**0,66**
		J48	0,60	**0,44**	0,57	0,47	0,58	**0,52**	0,04	0,53	0,62
		RF	0,72	0,31	0,56	0,47	0,63	0,50	0,03	0,56	0,64
		AB	**0,79**	0,23	0,56	0,47	**0,65**	0,48	0,03	0,56	0,63
		MLP	0,69	0,39	**0,58**	0,50	0,62	**0,52**	**0,07**	**0,58**	0,65
		XGB	0,67	0,35	0,56	0,46	0,61	0,50	0,01	0,54	0,63
$C2$	$HMV_{CTI}(treino)$ $HCPA(teste)$	LR	**0,76**	0,33	0,62	0,49	**0,68**	0,53	0,09	**0,60**	0,69
		J48	0,58	**0,49**	0,62	0,45	0,60	0,53	0,07	0,54	0,67
		RF	0,67	0,47	**0,64**	**0,50**	0,65	**0,57**	**0,14**	**0,60**	0,68
		AB	0,67	0,42	0,62	0,47	0,64	0,53	0,08	0,57	0,66
		MLP	0,65	0,46	0,63	0,48	0,64	0,55	0,11	0,58	**0,70**
		XGB	0,62	**0,49**	0,63	0,47	0,62	0,55	0,10	0,58	0,67

Fonte: O autor

Um aspecto interessante é que a execução, na qual o conjunto de dados HCPA foi utilizado como treino e o conjunto de dados HMV$_{CTI}$ como teste, resultou em níveis de sensibilidade superiores quando comparado à execução inversa. Como os conjuntos de dados HMV$_{CTI}$ e HCPA têm respectivamente desbalanceamento de classe de 2,20 e

1,67 quando considerado a mortalidade como variável dependente, o treinamento com o conjunto HMV_{CTI} penaliza mais os erros FN – o que tende a aumentar os níveis de sensibilidade – do que o treinamento com o conjunto HCPA. Entretanto, os níveis de sensibilidade superiores foram observados quando o conjunto HCPA foi utilizado para treino (com o qual aplicou-se menor penalização aos erros FN).

A Tabela 5.6 apresenta os resultados da validação cruzada na tarefa de predição de necessidade de suporte de VMI. Para a tarefa em questão, a avaliação cruzada entre os conjuntos de dados resultou em índices de desempenho semelhantes em ambas as direções (treino com HCPA e teste com HMV_{CTI} bem como, treino com HMV_{CTI} e teste com HCPA). Quando treinado com o conjunto HCPA e testado com o conjunto HMV_{CTI}, o melhor resultado em termos de sensibilidade foi obtido pelo algoritmo AB (0,79). Entretanto, o mesmo algoritmo atingiu um índice de especificidade de apenas 0,23, sugerindo que o algoritmo classificou um grande número de pacientes que não necessitou recursos de VMI como pertencentes à classe *vmi=sim*. Na sequência, o algoritmo LR atingiu o segundo melhor resultado em termos de sensibilidade (0,75). Enquanto LR apresentou uma redução de sensibilidade de 4 pp em relação ao AB, os ganho em VPP e especificidade foram de 1 pp e 9 pp, respectivamente.

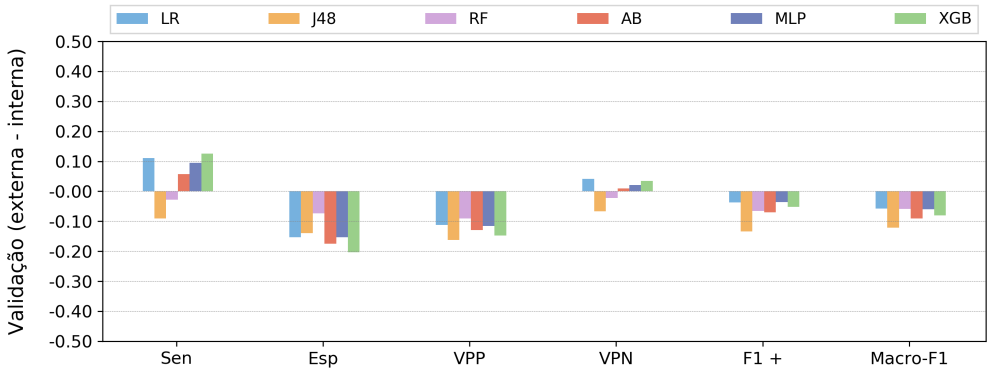
Quando considerado o treinamento com o conjunto de dados HMV_{CTI} e o teste com o conjunto de dados HCPA, o melhor resultado em termos de sensibilidade foi atingido pelo algoritmo LR (0,76), seguido pelos algoritmos RF e AB que atingiram um escore de sensibilidade de 0,67. Neste caso, o algoritmo LR atingiu níveis de sensibilidade superiores, enquanto que apresentou redução nos níveis de especificidade.

A Figura 5.1 apresenta as variações nos resultados entre realizar a validação interna, onde o treino e teste são conduzidos com porções de dados disjuntos provenientes de um único conjunto de dados (realizado nos experimentos 5.4.1, 5.4.2 e 5.4.3) e a validação externa (validação cruzada entre os conjuntos de dados HMV_{CTI} e HCPA). Dessa forma, a Figura 5.1 apresenta a variação obtida nos resultados ao utilizar um conjunto de dados de outro hospital como conjunto de treinamento em comparação à utilizar os conjuntos de dados de treino do próprio hospital. Deve-se observar que apenas um subconjunto de atributos estava disponível durante as execuções de validação cruzada, portanto as variações observadas não restringem-se exclusivamente aos diferentes conjuntos utilizados durante o treinamento.

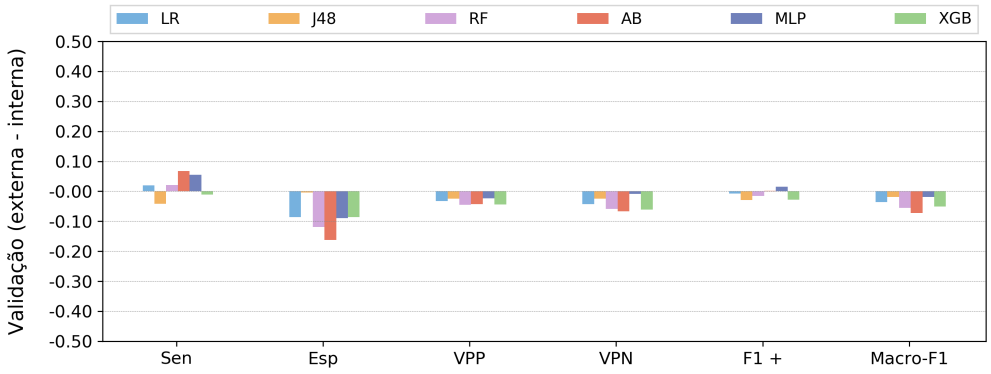
Tanto para a predição de mortalidade quanto para a predição de VMI, é possível observar padrões similares nos resultados. As Figuras 5.1a e 5.1b apresentam as variações

dos resultados obtidos no conjunto de dados HMV_{CTI} ao utilizar o HCPA para o treinamento nas tarefas de predição de mortalidade e necessidade de VMI, respectivamente. Em ambos os casos, o treinamento realizado com o conjunto de dados HCPA, resultou em aumento nos níveis de sensibilidade para a maioria dos classificadores. O aumento nos níveis de sensibilidade, é acompanhado pela queda dos níveis de especificidade e VPP.

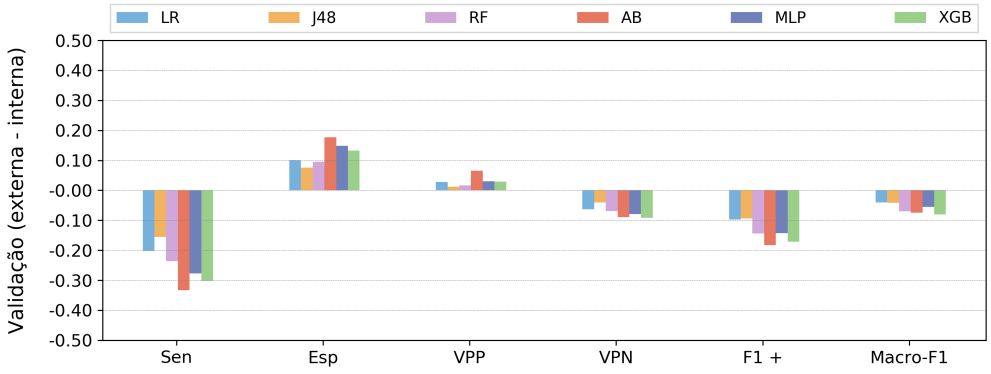
As Figuras 5.1c e 5.1d apresentam as variações nos resultados para o conjunto de dados HCPA ao utilizar o conjunto de dados HMV_{CTI} para o treinamento das tarefas de predição de mortalidade e necessidade de VMI. Nestes casos, ocorreu redução nos níveis de sensibilidade para todos os classificadores. Além disso, também observa-se redução nos níveis de VPN. A queda nos níveis de sensibilidade é acompanhada pelo aumento dos níveis de especificidade, especialmente para a tarefa de predição de mortalidade.



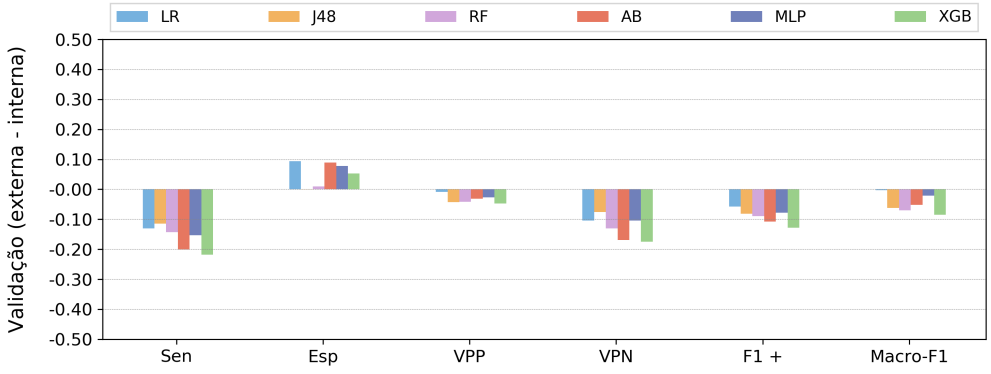
(a) Variação nos resultados do HMV_{CTI} ao treinar com $HCPA$ - predição de mortalidade



(b) Variação nos resultados do HMV_{CTI} ao treinar com $HCPA$ - predição de VMI



(c) Variação nos resultados do $HCPA$ ao treinar com HMV_{CTI} - predição de mortalidade



(d) Variação nos resultados do $HCPA$ ao treinar com HMV_{CTI} - predição de VMI

Figura 5.1 – Variação nos resultados ao treinar os classificadores com o mesmo conjunto de dados e ao treinar com outro conjunto de dados

Fonte: O Autor

Uma hipótese para justificar a redução dos índices de sensibilidade ao utilizar o conjunto de dados HMV_{CTI} para treino é que as margens de classificação, formadas no hiperplano pelas instâncias do conjunto de dados HCPA, podem ter sido melhor delimitadas - tanto devido ao maior número de instâncias, quanto pelas características das instâncias, ou até ambos os fatores. Logo, as margens formadas circundaram as instâncias do conjunto de dados HMV_{CTI}, ao instante em que o baixo número de instâncias do HMV_{CTI}, bem como suas características, podem não terem contribuído para a formação de margens de classificação adequadas para abranger as instâncias do conjunto HCPA.

Em resposta à questão de pesquisa *Q4*, a utilização de modelos preditivos treinados com um conjunto de dados proveniente de um hospital pode não generalizar adequadamente quando aplicado em outro hospital. Embora para certas tarefas e algoritmos, tenha-se atingido níveis de sensibilidade semelhantes, ou até mesmo superiores, também foram observadas quedas consideráveis nos níveis de VPP e especificidade. Assim, é importante que a utilização de conjuntos de dados externos seja cuidadosamente avaliada caso a caso.

5.4.5 Quais atributos foram considerados mais importantes para as tarefas de predição avaliadas?

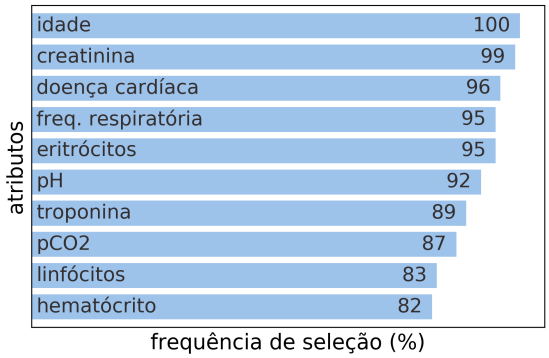
Esta análise tem como objetivo examinar os atributos considerados importantes durante a avaliação das tarefas de predição. Para isso, examinamos os atributos que foram frequentemente selecionados pelo algoritmo de seleção de atributos e também aplicamos a abordagem SHAP para avaliar os atributos considerados mais relevantes pelos algoritmos durante as predições.

Entre as etapas do *pipeline* aplicado em nossos experimentos, a seleção de atributos é a primeira etapa onde ocorre o julgamento e seleção de um subconjunto de atributos considerados úteis para a tarefa de predição em questão. A Figura 5.2 apresenta, para cada tarefa, a lista com os 10 atributos mais frequentemente selecionados ao longo da execução das B iterações de reamostragem $bootstrap$.

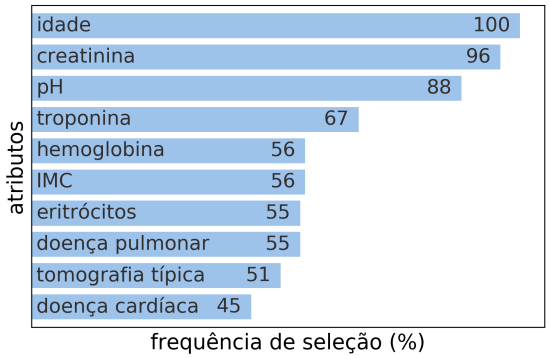
Na etapa de seleção de atributos, a idade foi o atributo selecionado com maior frequência. Na maioria das tarefas, a idade foi selecionada em todas as B iterações de reamostragem, exceto para as tarefas de predição de VMI no conjunto de dados HMV, nas quais a idade foi selecionada respectivamente 99% e 80% das vezes para predição de VMI para todos os pacientes e predição de VMI para pacientes da CTI. Ainda, considerando

as execuções de todas as tarefas, o atributo *eritrócitos* foi o segundo mais selecionado (74,29%), seguido do atributo *pH* (53,85%).

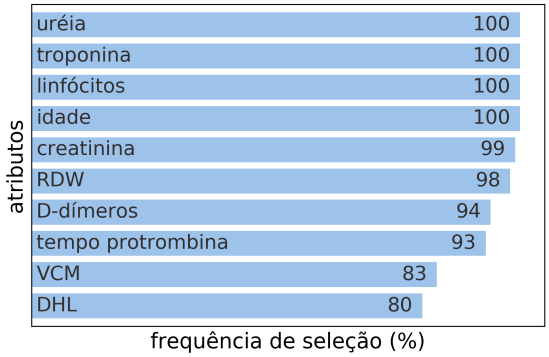
Quando consideradas as Figuras 5.2b e 5.2f, referentes às tarefas de predição de mortalidade e VMI apenas para os pacientes internados na CTI, é possível observar uma menor consistência entre as variáveis mais selecionadas. Na predição de mortalidade, apenas os atributos *idade*, *creatinina* e *pH* foram selecionados em ao menos 70% das reamostragens. Já para a predição de VMI, apenas os atributos *pH*, *idade* e *eritrócitos* foram selecionados acima desse limiar. Esta característica contrasta o comportamento observado durante a seleção de atributos nas demais tarefas, nas quais mesmo os atributos ranqueados na décima posição apresentaram uma frequência de seleção superior à 80%. Curiosamente, este comportamento não se repetiu nas tarefas de predição de mortalidade e VMI para o conjunto de dados HCPA, embora este também contenha apenas pacientes internados na CTI. Acreditamos que uma explicação para este fenômeno seja devido ao fato de que o conjunto HMV_{CTI} possui um número menor de pacientes, acarretando em uma maior variabilidade durante a seleção de atributos.



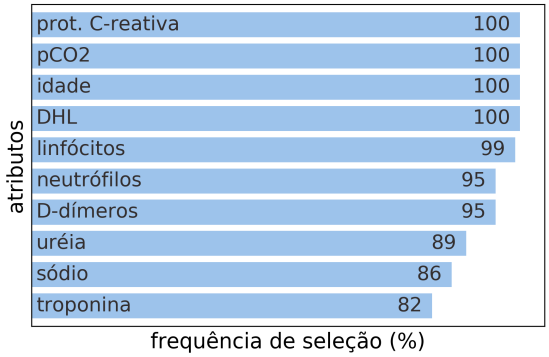
(a) HMV mortalidade



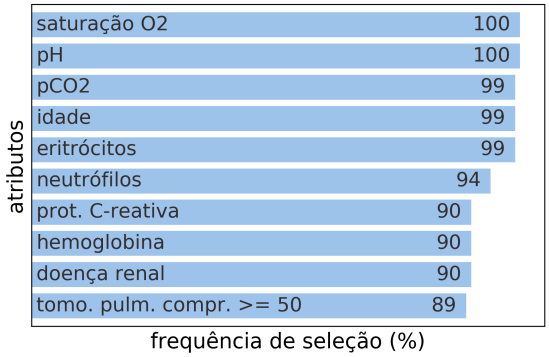
(b) HMV_{CTI} mortalidade



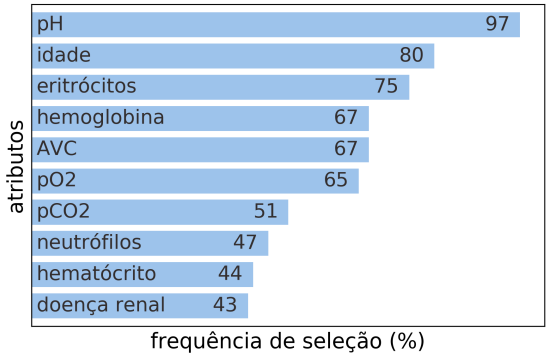
(c) $HCPA$ mortalidade



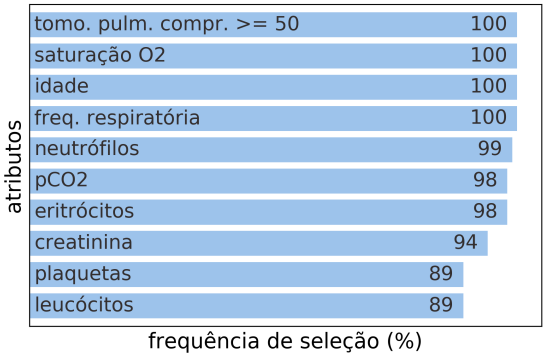
(d) $HCPA$ VMI



(e) HMV VMI



(f) HMV_{CTI} VMI



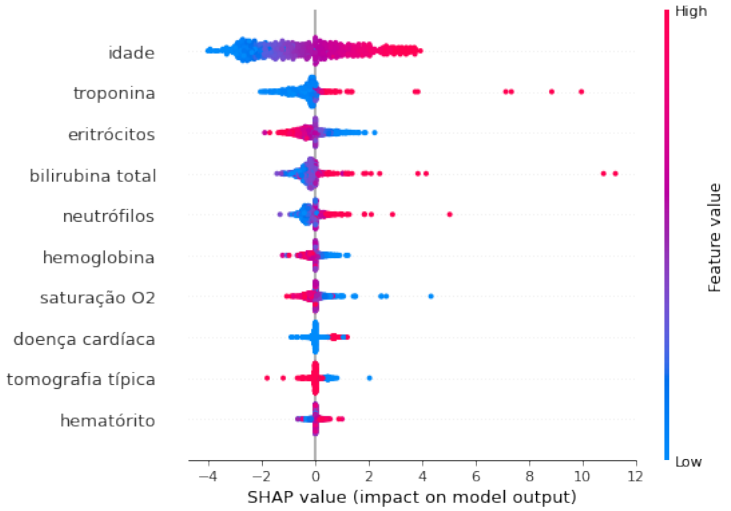
(g) HMV CTI

Figura 5.2 – Atributos mais selecionados durante o procedimento de seleção de atributos
Fonte: O Autor

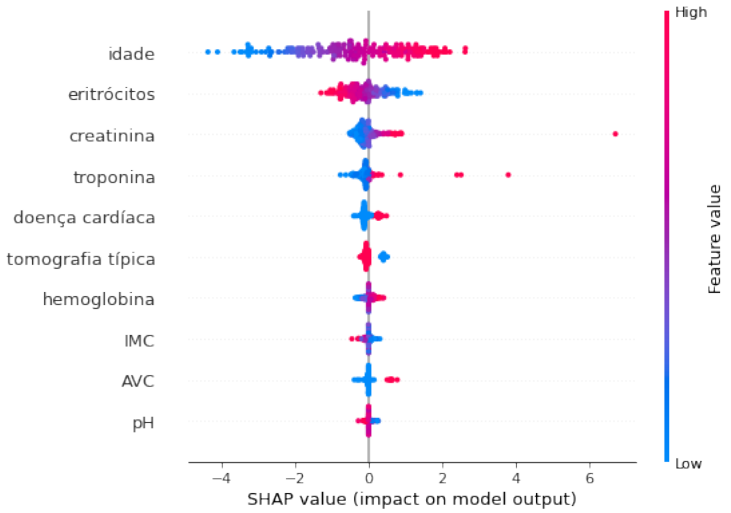
O conjunto de dados HCPA possui atributos que não estão disponíveis no conjunto HMV. Dentre esses, pode-se citar a *uréia*, *sódio*, *tempo de protrombina*, *RDW* e *VCM*, os quais foram frequentemente escolhidos durante a seleção de atributos. A uréia foi selecionada em todas as reamostragens para a tarefa de predição de mortalidade e em 89% das vezes na tarefa de predição de VMI. Os atributos *RDW*, *tempo de protrombina* e *VCM* foram selecionados respectivamente 98%, 93% e 83% das vezes na tarefa de predição de mortalidade, enquanto que o atributo sódio foi selecionado 86% das vezes para a tarefa de predição de VMI. Também é possível observar que tanto na tarefa de predição de CTI, quanto nas tarefas de predição de necessidade de VMI, destacaram-se atributos relacionados à função respiratória do paciente.

Como já mencionado, os atributos apresentados na Figura 5.2 referem-se às dez variáveis mais selecionadas nas $B = 100$ iterações de reamostragem conduzidas para estimar o desempenho dos modelos preditivos (resultados apresentados nas Subseções 5.4.1, 5.4.2 e 5.4.3). Os atributos selecionados mais frequentemente indicam maior consistência quanto a utilidade desses atributos, enquanto que atributos selecionados com menor frequência expõem fragilidade e suscetibilidade desses atributos à variações e sinais ruidosos presentes no conjunto de dados. Embora seja útil para reduzir o espaço dimensional e pré-selecionar atributos com maior chance de serem úteis, é importante frisar que esta análise indica quais variáveis tendem a ser mais úteis com base no critério de julgamento do método de seleção de atributos e não do algoritmo de classificação.

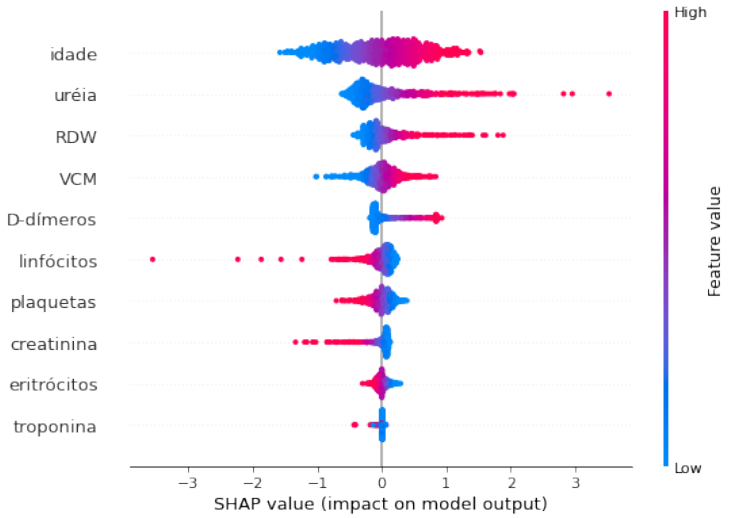
Com o objetivo de compreender quais foram os atributos levados em consideração pelos classificadores e como os valores dos atributos se relacionam com as saídas preditas, fizemos uso da abordagem SHAP, a qual viabiliza a análise das contribuições que os valores dos atributos exercem na predição. As Figuras 5.3, 5.4 e 5.5 apresentam os gráficos *summary-plot* com a relação dos atributos e suas contribuições para as tarefas de predição de mortalidade, de internação em CTI e predição de necessidade de VMI. Os gráficos apresentados referem-se aos algoritmos que atingiram os melhores resultados em termos de sensibilidade nos experimentos das Subseções 5.4.1, 5.4.2 e 5.4.3.



(a) LR para predição de desfecho no HMV



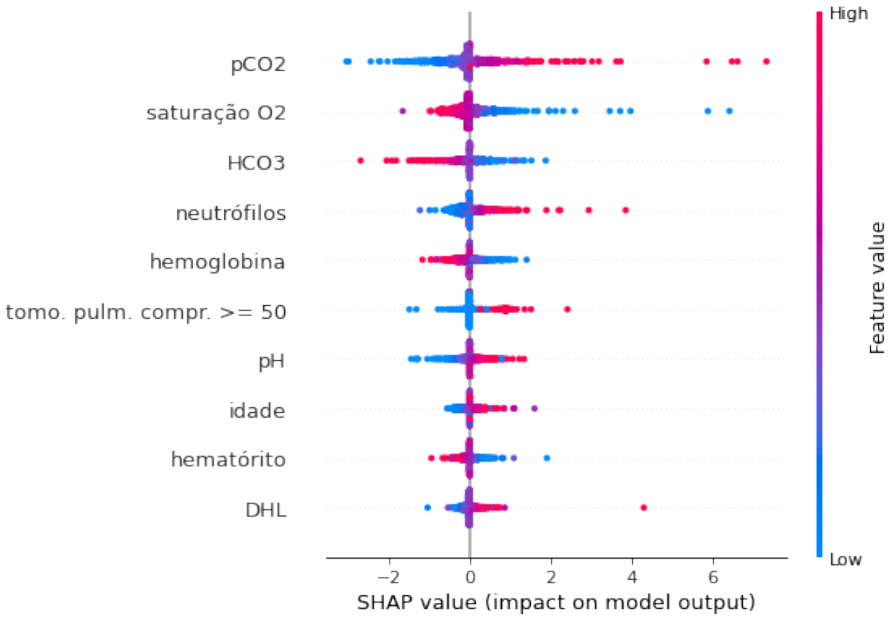
(b) LR para predição de desfecho no HMV_{CTI}



(c) LR para predição de desfecho no $HCPA$

Figura 5.3 – Visualizações dos principais atributos e suas contribuições nas tarefas de predição de mortalidade
Fonte: O Autor

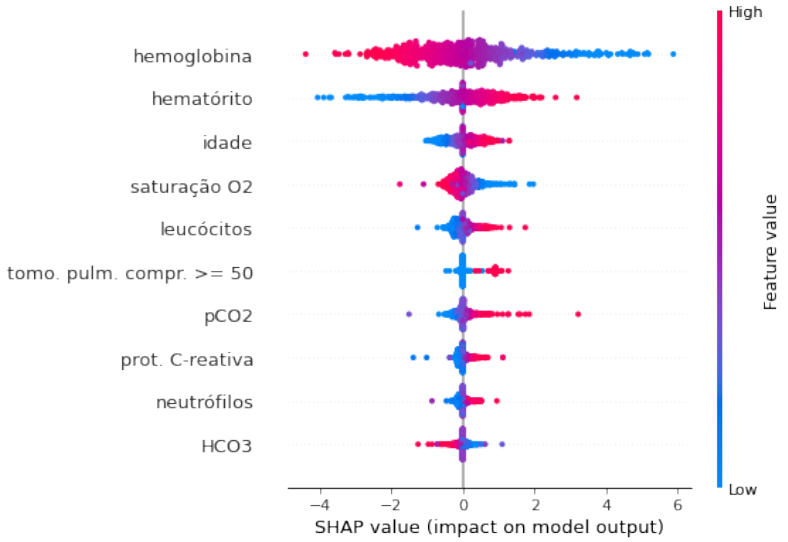
A depender da tarefa, diferentes atributos foram elencados como os mais relevantes durante as predições. De modo geral, ao considerar todas as tarefas avaliadas, a idade sobressaiu-se como o atributo de maior contribuição durante as predições. A relevância da idade do paciente foi observada para a tarefa de predição de mortalidade em ambos os conjuntos de dados (HMV e HCPA) e para os dois cenários *C1* e *C2* do HMV.



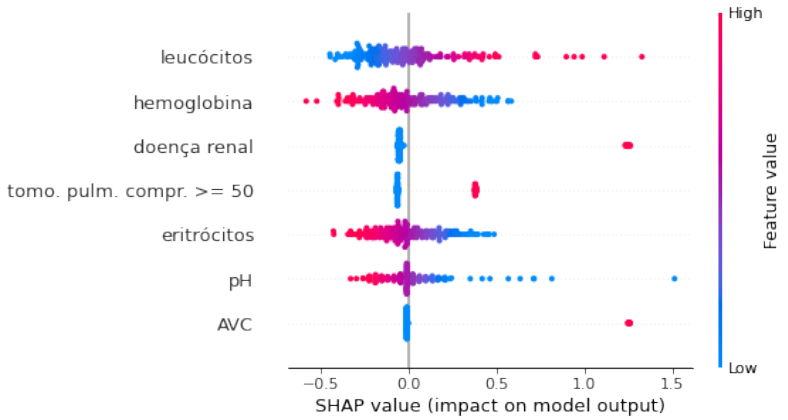
(a) LR para predição de CTI no HMV

Figura 5.4 – Visualizações dos principais atributos e suas contribuições nas tarefas de predição de CTI
Fonte: O Autor

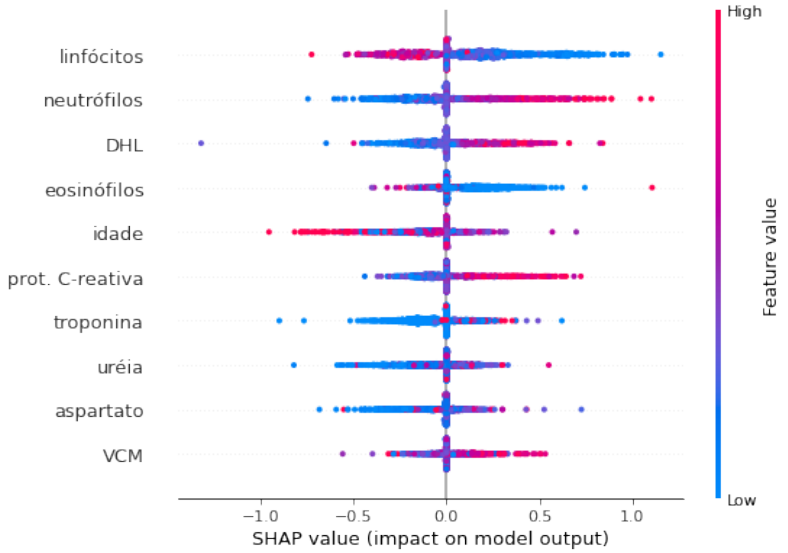
Quando o atributo *idade* foi elencado, a idade avançada dos pacientes foi sinalizada como fator agravante para o prognóstico do paciente, indicando maior contribuição para a predição da classe *mortalidade=sim*. A idade avançada é um fator de risco para COVID-19 bem conhecido e recorrentemente citado na literatura por estudos que aplicam algoritmos de AM para tarefas de diagnóstico e prognóstico de pacientes no contexto da COVID-19 (KANG; JUNG, 2020; LEUNG, 2020; ZHOU et al., 2020).



(a) LR para predição de VMI no HMV



(b) LR para predição de VMI no HMV_{CTI}



(c) RF para predição de VMI no $HCPA$

Figura 5.5 – Visualizações dos principais atributos e suas contribuições nas tarefas de predição de VMI
Fonte: O Autor

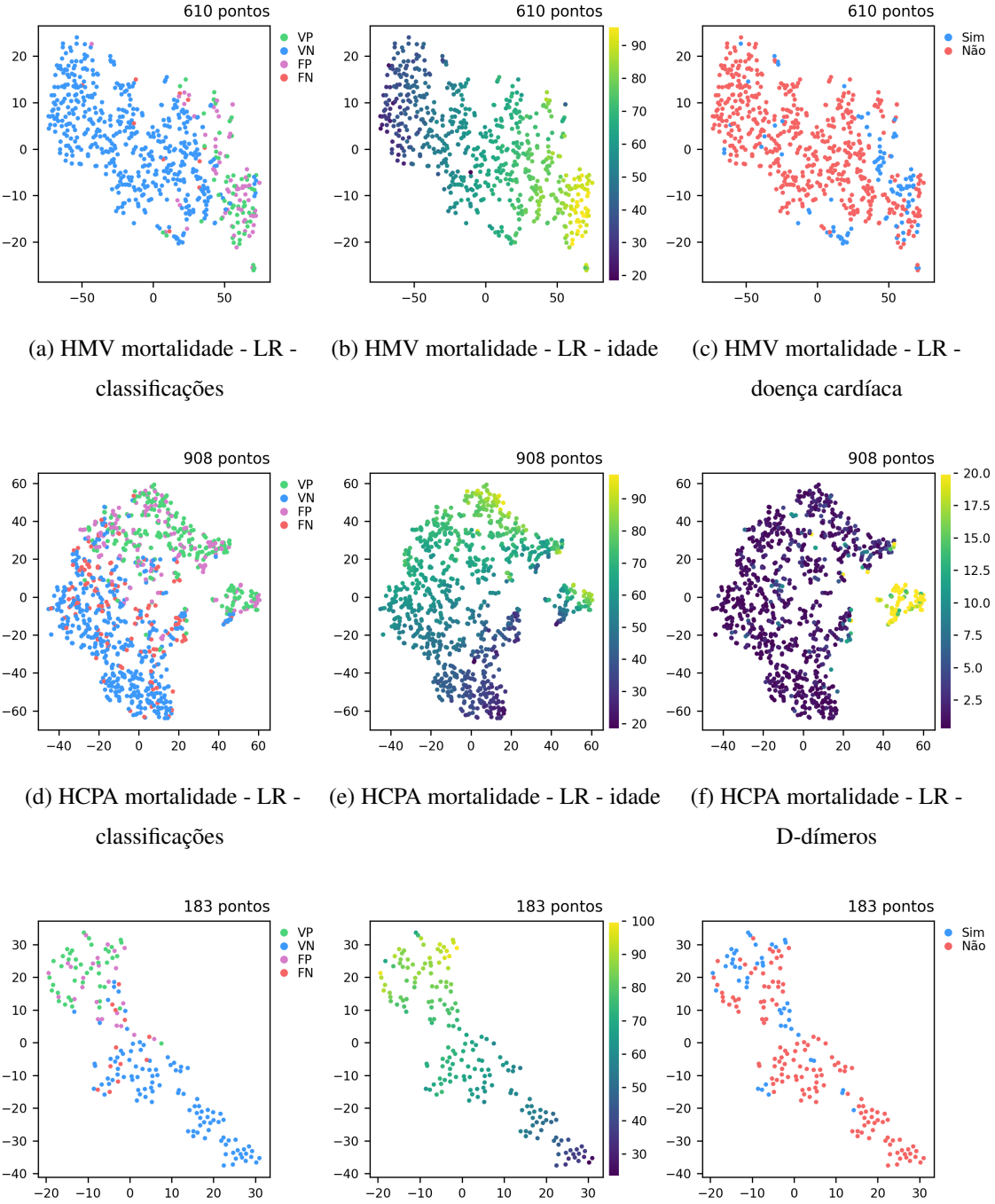
Embora apenas disponível no conjunto de dados HCPA, o nível de ureia também foi apontado como uma informação de alta contribuição para a predição de mortalidade, sugerindo que níveis baixos de ureia tendem a estar associados a pacientes pertencentes à classe *mortalidade=não* e que níveis medianos e altos são fatores observados em pacientes da classe *mortalidade=sim*. Outros contribuintes elencados para a predição de mortalidade incluem os atributos *neutrófilos*, *eritrócitos*, *linfócitos*, *hemoglobina*, *troponina*, *creatinina*, *VCM*, *D-dimeros*, *plaquetas*, *AVC* e *doença cardíaca*.

Tanto para a predição de CTI quanto de VMI, a relevância do atributo *idade* é ofuscada por outros atributos. Para a predição de CTI, fatores como aumento de *pCO2*, quedas de *saturação de O2*, baixos índices de *HCO3*, altas dosagens para *neutrófilos*, baixas dosagens para *hemoglobina* e *comprometimento pulmonar* estão associados à predição de necessidade de internação em CTI. Para a predição de necessidade de VMI, estão relacionadas baixas dosagens de *hemoglobina*, valores altos de *hematócrito*, valores altos para proteína C-reativa, baixa *saturação de O2*, valores altos para *leucócitos*, *comprometimento pulmonar*, bem como elevados índices de *pCO2*. Especificamente para a predição de VMI dos pacientes internados em CTI (HMV_{CTI}), valores baixos de *pH*, bem como as comorbidades de histórico de *doença renal*, *doença pulmonar*, *AVC* e *doenças do sistema nervoso*, foram elencadas como fatores associados à necessidade de suporte ventilatório.

5.4.6 Visualizações com *Shapley Values*

Com o objetivo de compreender como as predições realizadas pelos classificadores estão relacionadas com os valores dos atributos, utilizamos a técnica de visualização descrita na Seção 4.6.

A Figura 5.6 mostra as visualizações obtidas com t-SNE a partir dos *shapley values* calculados para a tarefa de predição de mortalidade. Cada ponto corresponde a uma instância (paciente). Os pontos são coloridos com base nos valores de diferentes atributos para auxiliar na interpretação das projeções geradas. O primeiro gráfico exibe as previsões corretas e incorretas (verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN)). O segundo e o terceiro gráfico demonstram pontos colorido por diferentes atributos.



(a) HMV mortalidade - LR - classificações

(b) HMV mortalidade - LR - idade

(c) HMV mortalidade - LR - doença cardíaca

(d) HCPA mortalidade - LR - classificações

(e) HCPA mortalidade - LR - idade

(f) HCPA mortalidade - LR - D-dímeros

(g) HMV$_{CTI}$ mortalidade - LR - classificações

(h) HMV$_{CTI}$ mortalidade - LR - idade

(i) HMV$_{CTI}$ mortalidade - LR - doença cardíaca

Figura 5.6 – Visualizações dos padrões obtidos com base nos *shapley values* para a predição de mortalidade – o eixo x corresponde à dimensão 1 do t-SNE e o eixo y à dimensão 2 do t-SNE
Fonte: O Autor

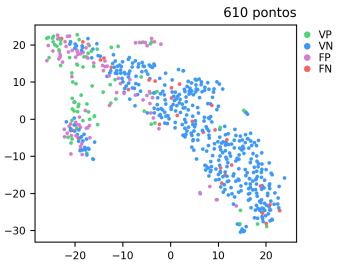
Os gráficos t-SNE reforçam a alta importância atribuída pelos classificadores ao atributo *idade* durante a predição de mortalidade. Ao analisar os gráficos com os resultados de classificação (primeiro gráfico de cada linha) e os gráficos coloridos com base na idade dos pacientes (Figuras 5.6b, 5.6e e 5.6h), é possível observar que erros recor-

rentes de FP (pontos roxos) e VP (pontos verdes) tendem a ocorrer nas regiões com alta concentração de pacientes mais velhos. Por outro lado, os erros FN (pontos vermelhos) e VN (pontos azuis) encontram-se em regiões onde predominam pacientes com idades medianas, bem como os jovens.

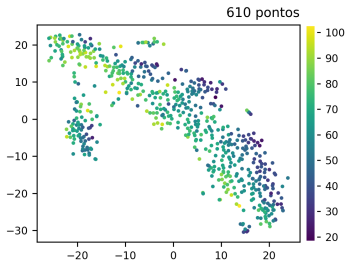
As Figuras 5.6b e 5.6h demonstram que as maiores concentrações de pacientes com doença cardíaca, cujos quais são frequentemente classificados como *mortalidade=sim*, também encontram-se nas regiões de idade avançada. Tanto a idade avançada, quanto doenças cardíacas, são conhecidas como fatores relacionados ao prognóstico desfavorável do paciente (MERCÊS; LIMA; NETO, 2020). Comportamento semelhante também é constatado no conjunto de dados HCPA e HMV$_{CTI}$. Analisando conjuntamente as Figuras 5.6d e 5.6e referentes ao conjunto de dados HCPA, bem como as Figuras 5.6g e 5.6h do conjunto de dados HMV$_{CTI}$, observa-se que a grande concentração de classificações VP e FP ocorreram para os pacientes com idade avançada.

A Figura 5.6f, referente ao conjunto de dados HCPA, mostra um agrupamento de pacientes com valores elevados de D-dímeros. Valores altos para D-dímeros podem indicar disfunções de coagulação (como a trombose e tromboembolismo), sendo este um fator de risco para os pacientes (ROSTAMI; MANSOURITORGHABEH, 2020). Quando analisado conjuntamente com a Figura 5.6e, observa-se que a região possui tanto pacientes jovens, quanto pacientes com idades medianas e avançadas. Ao consultar a Figura 5.6d, constata-se que nesta região, a maioria dos pacientes foram classificados como classe *mortalidade=sim* (VP e FP), exceto algumas instâncias na parte inferior da região, as quais foram classificadas como pertencentes à classe *mortalidade=não* (VN + FN). Os pacientes classificados como pertencentes à classe *mortalidade=não* eram pacientes mais jovens. Embora esta região seja predominantemente constituída de pacientes com valores altos para D-dímeros, para pacientes mais jovens, o classificador tendeu a atribuir os pacientes à classe *mortalidade=não*, enquanto que para os pacientes de idade mediana e avançada, o classificador tendeu a atribuí-los à classe *mortalidade=sim*. Para ambos os casos, o classificador acertou a maioria das classificações. Dessa forma, uma hipótese é que o classificador pode ter incorporado que valores mais altos de D-dímeros, embora seja um fator de risco para os pacientes com COVID-19, não está tão associado à morte em pacientes mais jovens.

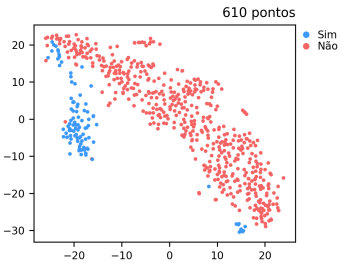
A Figura 5.7 apresenta as visualizações obtidas com o t-SNE a partir dos *shapley values* calculados para a tarefa de predição de VMI nos conjuntos de dados HMV e HMV$_{CTI}$.



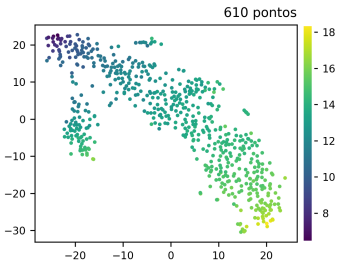
(a) HMV VMI - LR - classificações



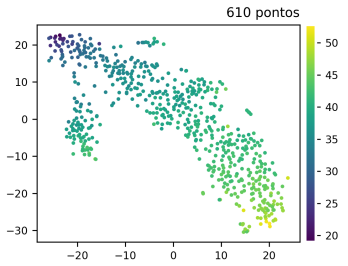
(b) HMV VMI - LR - idade



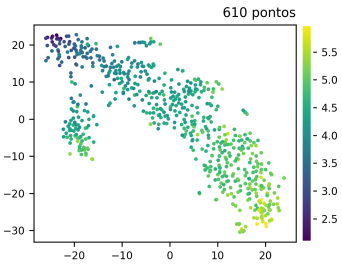
(c) HMV VMI - LR - tomo. pulm. compr. >= 50%



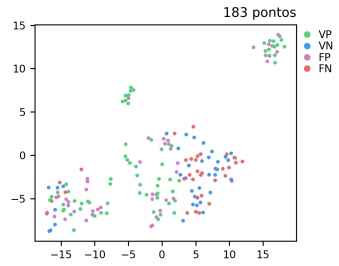
(d) HMV VMI - LR - hemoglobina



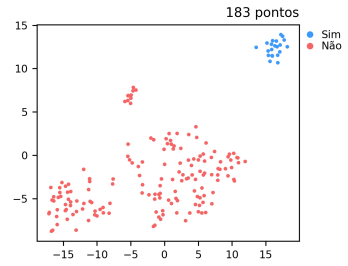
(e) HMV VMI - LR - hematócrito



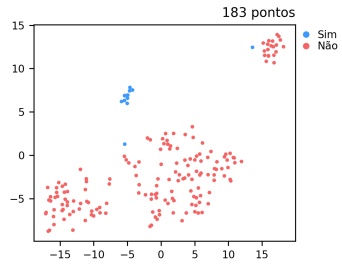
(f) HMV VMI - LR - eritrócitos



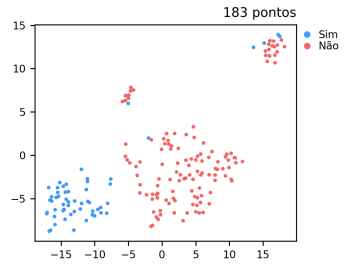
(g) HMV_{CTI} VMI - LR - classificações



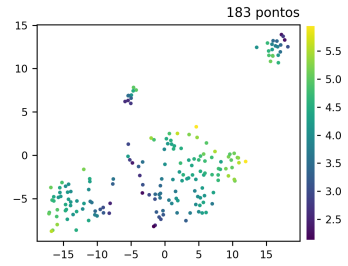
(h) HMV_{CTI} VMI - LR - doença renal



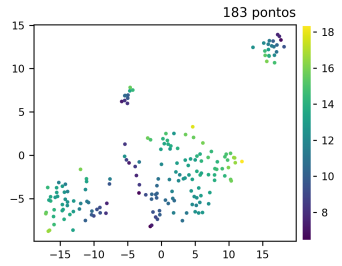
(i) HMV_{CTI} VMI - LR - doença cardíaca



(j) HMV_{CTI} VMI - LR - tomo. pulm. compr >= 50%



(k) HMV_{CTI} VMI - LR - eritrócitos



(l) HMV_{CTI} VMI - LR - hemoglobina

Figura 5.7 – Visualizações dos padrões obtidos com base nos *shapley values* para a predição de VMI – o eixo x corresponde à dimensão 1 do t-SNE e o eixo y à dimensão 2 do t-SNE
Fonte: O Autor

A Figura 5.7a apresenta as classificações realizadas pelo algoritmo LR para a ta-

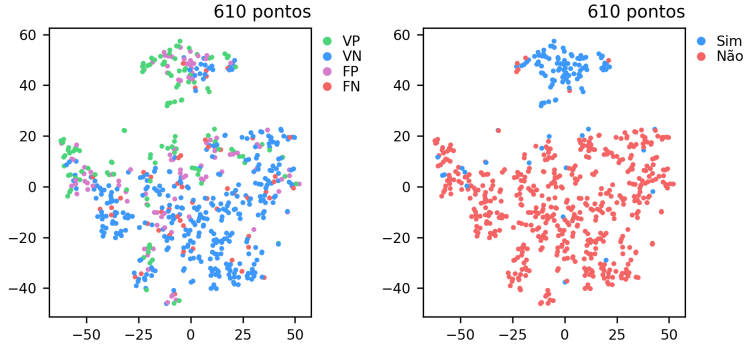
refa de predição de VMI nos conjuntos de dados HMV, enquanto que as Figuras 5.7b, 5.7c, 5.7d, 5.7e e 5.7f apresentam respectivamente as visualizações nas quais as instâncias foram coloridas pelos valores dos atributos *idade*, ocorrência de *tomografia com comprometimento pulmonar igual ou superior à 50%*, *hemoglobina*, *hematócrito* e *eritrócitos*.

Ao contrário do que é observado na tarefa de predição de mortalidade, a idade avançada tende a não ser um fator determinante para a atribuição dos pacientes à classe positiva (*vmi=sim*). A Figura 5.7b apresenta a visualização onde os pontos (pacientes) são coloridos com base no atributo *idade*. Ao analisar conjuntamente com a Figura 5.7a, é possível observar que as regiões onde encontram-se os pacientes mais velhos, não correspondem necessariamente onde predominam os pacientes classificados como *vmi=sim*. Por sua vez, a Figura 5.7c mostra que a região formada por pacientes que tinham comprometimento dos pulmões igual ou superior à 50%, converge com a região da Figura 5.7a na qual houve predomínio da atribuição à classe *vmi=sim*.

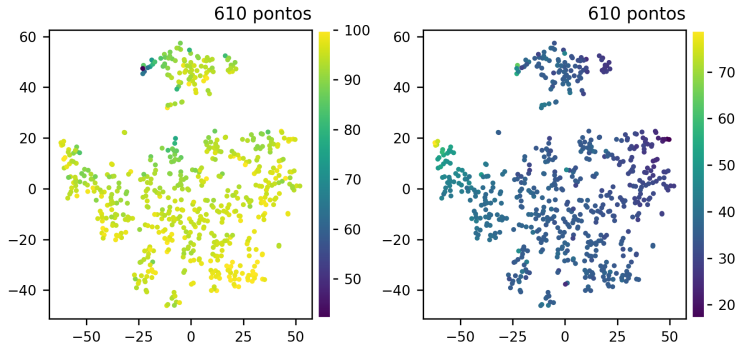
As Figuras 5.7d, 5.7e e 5.7f apresentam as visualizações nas quais as instâncias foram coloridas respectivamente pelos atributos *hemoglobina*, *hematócrito* e *eritrócitos*. Para os três atributos, observa-se um padrão semelhante, nos quais parte-se de valores baixos (cores escuras) presentes na região superior esquerda dos gráficos, até os valores mais altos (tons esverdeados e amarelos) na região inferior direita do gráfico. A partir dessas três visualizações, de forma complementar aos gráficos *summary-plot* da biblioteca SHAP, é possível constatar que os pacientes que apresentaram os valores mais altos para o atributo *hemoglobina*, também foram aqueles que apresentaram os valores mais altos de *hematócrito* e *eritrócitos*.

As Figuras 5.7g, 5.7h, 5.7i, 5.7j, 5.7k e 5.7l apresentam as visualizações obtidas com base nos *shapley values* estimados com o algoritmo LR na tarefa de predição de necessidade de VMI para o conjunto de dados HMV_{CTI}. Nas Figuras 5.7h e 5.7i, pode-se observar duas regiões onde arranjaram-se respectivamente, instâncias de pacientes com doença renal e doença cardíaca. Para ambas as regiões, o classificador atribuiu as instâncias à classe *vmi=sim*, sugerindo que o aumento de risco dessas comorbidades pode ter sido incorporado pelo classificador. Assim como observado para a tarefa de predição de necessidade VMI do conjunto de dados HMV, ao colorir as instâncias da visualização com base nos valores do atributo que indica pacientes com comprometimento pulmonar igual ou superior à 50%, também é possível observar um agrupamento onde predominam instâncias atribuídas à classe *vmi=sim*. Outra semelhança com os padrões observados na tarefa de predição de necessidade de VMI no conjunto de dados HMV, compete aos

atributos *eritrócitos* e *hemoglobina*, que apresentam padrões de coloração muito semelhantes, indicando uma correlação entre os atributos. Assim como observado no conjunto de dados HMV, as instâncias com valores inferiores para os atributos *eritrócitos* e *hemoglobina* tenderam a ser atribuídas à classe *vmi=sim*.



(a) HMV CTI - LR - classificações (b) HMV CTI - LR - tomo. pulm. compr. >= 50%



(c) HMV CTI - LR - saturação de O2 (d) HMV CTI - LR - pCO2

Figura 5.8 – Visualizações dos padrões obtidos com base nos *shapley values* para a predição de CTI – o eixo x corresponde à dimensão 1 do t-SNE e o eixo y à dimensão 2 do t-SNE
Fonte: O Autor

A Figura 5.8 apresenta as visualizações obtidas com o t-SNE a partir dos *shapley values* calculados para a tarefa de predição de CTI nos conjuntos de dados HMV. A Figura 5.8a apresenta as classificações realizadas pelo classificador LR. As Figuras 5.8b, 5.8c e 5.8d apresentam as instâncias coloridas pelos atributos *comprometimento pulmonar igual ou superior à 50%*, *SO2* e *pCO2*. Assim como observado na tarefa de predição de VMI, pacientes com comprometimento pulmonar tenderam a ser atribuídos à classe positiva (*cti=sim*). Isso pode ser observado na região superior da visualização. Nesta região, as instâncias com níveis de saturação de O2 superiores tenderam a ser classificadas como pertencentes à classe negativa (*cti=não*). Outro padrão relacionado à necessidade de

VMI (*vmi=sim*), diz respeito às instâncias com valores altos de *pCO2*, as quais localizam-se principalmente no lado esquerdo da visualização.

Deve-se mencionar que os achados evidenciados pelas visualizações não implicam necessariamente relações de causa e efeito, mas ainda assim, são de suma importância para a explicabilidade dos modelos preditivos. Além do mais, os achados estão alinhados com com os relatos da literatura, apresentados no Capítulo 3.

5.4.7 O desempenho obtido com os classificadores de múltiplas variáveis é significativamente superior ao desempenho obtido por classificadores de variável única?

Como constatado na análise da Seção 5.4.5, a idade foi frequentemente sinalizada como o atributo de maior contribuição durante as tarefas de predição de mortalidade, ao instante em que atributos relacionados com a função respiratória (*saturação de O2* e *pCO2*) foram elencados como atributos úteis para a tarefa de predição de internação em CTI e necessidade de VMI.

A presente questão de pesquisa tem como objetivo comparar o desempenho entre os classificadores com múltiplas variáveis independentes e suas respectivas versões com apenas uma variável independente. Para isso, comparamos o desempenho dos classificadores apresentados nas Seções 5.4.1, 5.4.2 e 5.4.3. As Figuras 5.9, 5.10 e 5.11 apresentam as diferenças entre os desempenhos obtidos pelos classificadores treinados com apenas uma variável independente, em relação aos classificadores de múltiplas variáveis para as tarefas de predição de mortalidade, CTI e VMI. Para as tarefas de predição de mortalidade, foram comparados classificadores que utilizaram apenas a idade como variável independente. Para a tarefa de predição de internação em CTI, comparamos classificadores que utilizaram apenas a idade e apenas saturação de O2 como variáveis independentes. Por fim, nas tarefas de predição de necessidade de VMI, comparamos classificadores que utilizaram apenas *saturação de O2* e *pCO2* variáveis independentes.

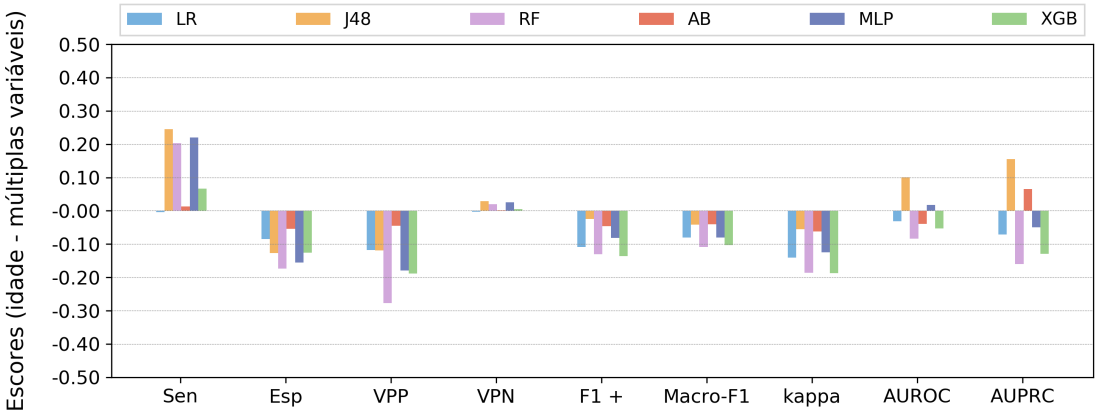
Na maioria dos experimentos, os classificadores que utilizaram apenas uma variável independente não apresentaram degradação de sensibilidade, sendo que em alguns casos constatou-se índices superiores de sensibilidade. Embora este efeito positivo nos níveis de sensibilidade tenha sido observado, também foi recorrente a redução nos níveis de VPP e especificidade. Como consequência, na maioria dos casos, não observou-se ganhos para as métricas F1+, ma-F1, Kappa, AUROC e AUPRC. Estes comportamentos foram observados para todas as tarefas de predição nos dois conjuntos de dados (HMV e

HCPA).

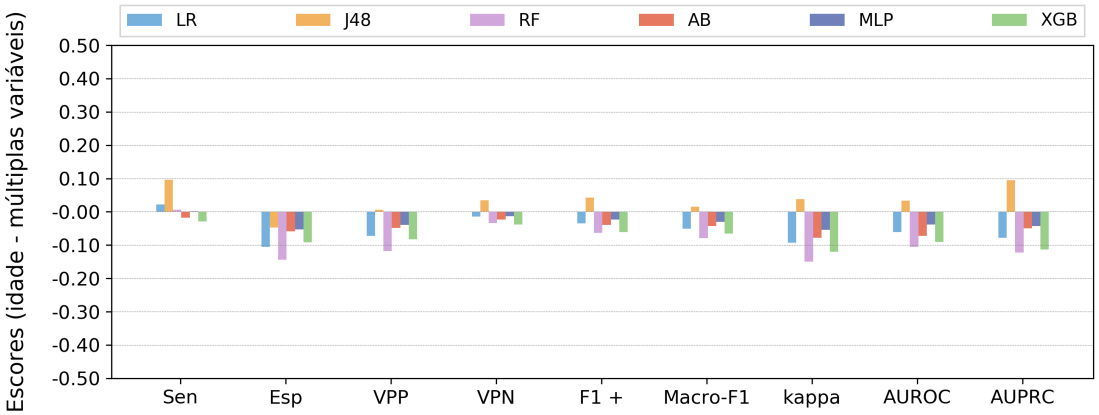
Nas tarefas de predição de mortalidade (Figura 5.9), o algoritmo LR, que apresentou os melhores resultados nos experimentos da Seção 5.4.1, tendeu a não apresentar quedas substanciais nos níveis de sensibilidade. Para o conjunto de dados HMV, ocorram aumentos consideráveis nos níveis de sensibilidade para alguns algoritmos. Em contrapartida, ocorreram reduções consideráveis nos níveis de VPP e especificidade, principalmente no cenário *C1* do conjunto de dados HMV e no conjunto de dados HCPA.

O algoritmo J48, que apresentou os menores índices de sensibilidade no experimento da Seção 5.4.1, foi recorrentemente o que obteve os maiores ganhos de sensibilidade entre os classificadores que utilizaram apenas a idade como variável independente.

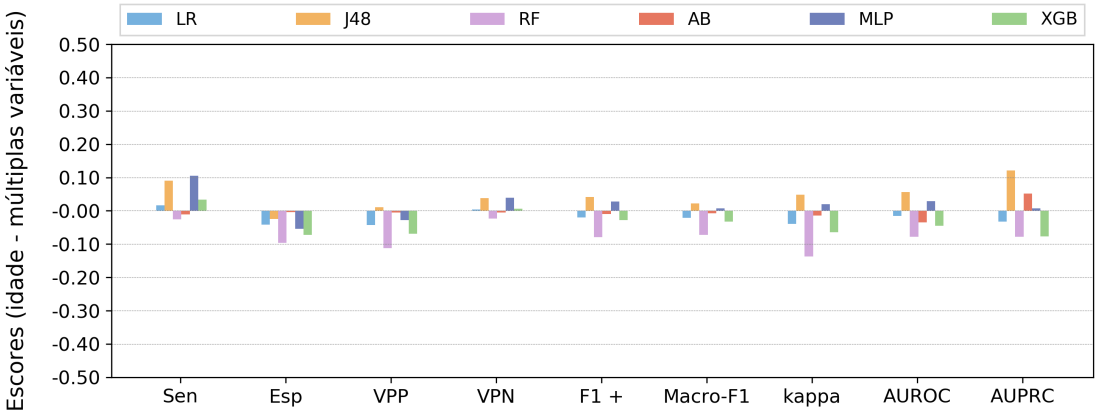
Quando considerado o conjunto de dados HMV no cenário *C1* (Figura 5.9a), além do algoritmo J48, os algoritmos RF e MLP destacaram-se pelo aumento de sensibilidade. Entretanto, as perdas nos níveis de VPP e especificidade não proporcionaram ganhos para as métricas F1+, ma-F1 e Kappa. Por fim, os algoritmos J48 e MLP foram os únicos que obtiveram ganhos de AUROC e o J48 e AB para AUPRC. No conjunto de dados HCPA (Figura 5.9b), apenas o algoritmo J48 obteve aumento para as demais métricas, além da sensibilidade. Para o cenário *C2* do conjunto de dados HMV (Figura 5.9c) os algoritmos J48 e MLP obtiveram os maiores ganhos de sensibilidade. Neste cenário, ambos os algoritmos apresentaram uma menor redução em termos de VPP e especificidade, resultando em aumento nos níveis das demais métricas.



(a) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de mortalidade no HMV



(b) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de mortalidade no $HCPA$



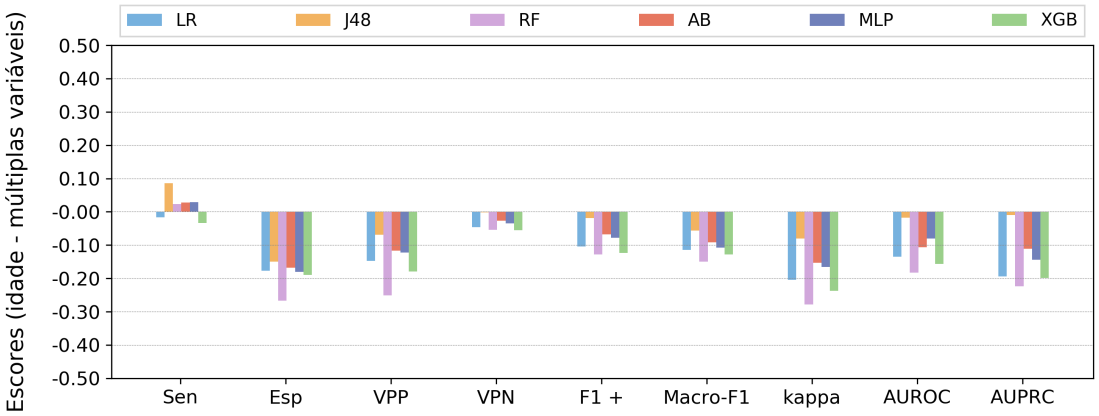
(c) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de mortalidade no HMV_{CTI}

Figura 5.9 – Comparação entre os classificadores de múltiplas variáveis e os de variável única (idade) para a predição de mortalidade

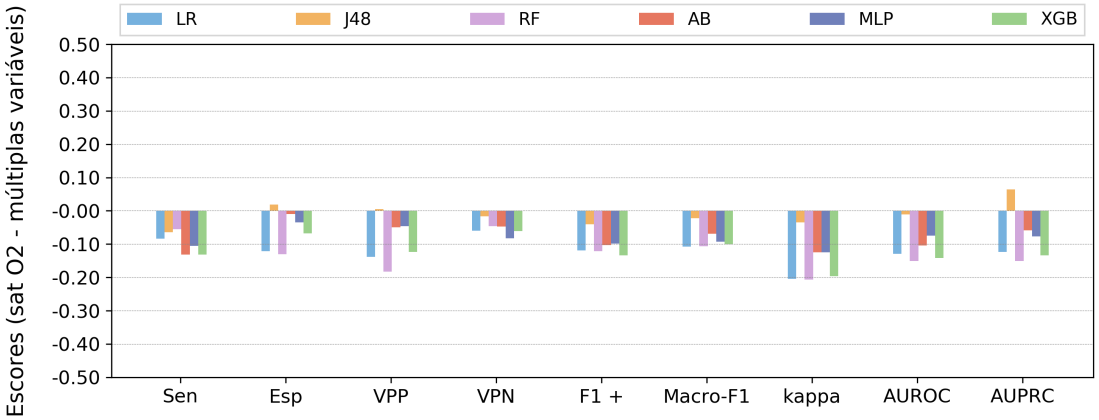
Fonte: O Autor

A Figura 5.10 apresenta as diferenças entre os desempenhos obtidos pelos classi-

ficadores que utilizam apenas a variável *idade* e apenas a variável *saturação de O2* como variável independente, em relação aos classificadores de múltiplas variáveis para as tarefas de predição de CTI. A Figura 5.10a apresenta as diferenças obtidas utilizando apenas a variável *idade*. Em termos de sensibilidade, apenas os algoritmos LR e XGB obtiveram resultados inferiores aos classificadores de múltiplas variáveis. Embora os algoritmos tenham atingido níveis de sensibilidade semelhantes aos atingidos pelos classificadores de múltiplas variáveis, ao utilizar apenas a idade como variável independente, ocorreram perdas nos índices de VPP e especificidade que excedem 10%. Exceto para a sensibilidade, nenhum algoritmo obteve ganhos de desempenho quando consideradas as demais métricas.



(a) Diferença entre os classificadores de múltiplas variáveis e os de variável única (idade) para a predição de CTI no HMV



(b) Diferença entre os classificadores de múltiplas variáveis e os de variável única (saturação de O2) para a predição de CTI no HMV

Figura 5.10 – Comparação entre os classificadores de múltiplas variáveis e os de variável única (idade) para a predição de CTI

Fonte: O Autor

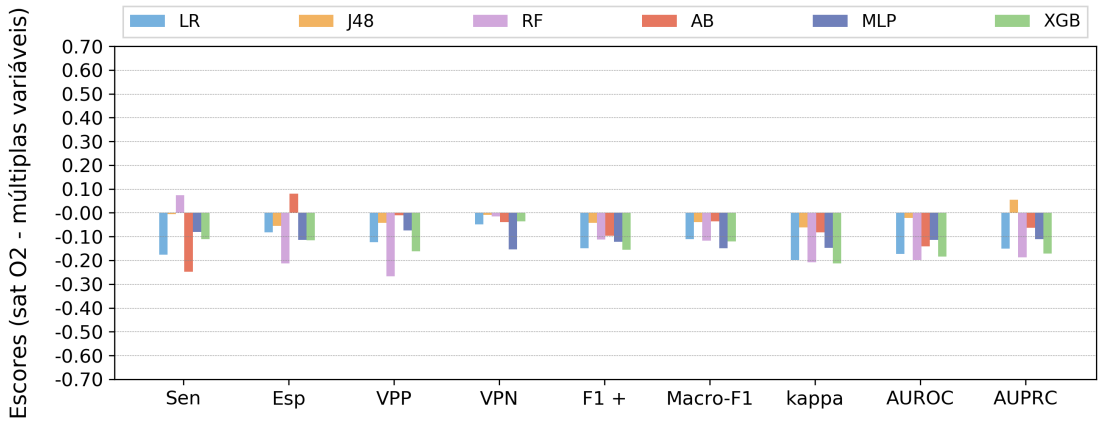
A Figura 5.10b apresenta as diferenças obtidas utilizando apenas a variável *satu-*

ração de O2 como variável independente. Neste caso, para quase todas as métricas, os algoritmos apresentaram redução nos escores. O algoritmo J48 foi o que apresentou a segunda menor redução nos níveis de sensibilidade, atrás do algoritmo RF. Entretanto, embora tenham sido os algoritmos que apresentaram as menores reduções de sensibilidade, é importante observar que os algoritmos J48 e RF foram os algoritmos que apresentaram os menores índices de sensibilidade pelos classificadores de múltiplas variáveis na predição de CTI.

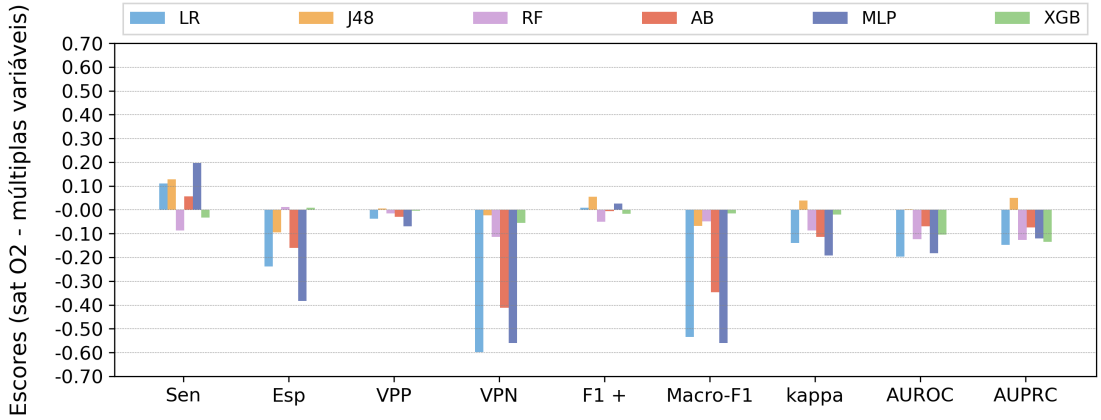
A Figura 5.11 apresenta as variações dos escores obtidas pelos classificadores treinados apenas com a *saturação de O2* dos pacientes, em relação aos classificadores multivariados para a predição de necessidade de VMI (experimento da Seção 5.4.3). Quando considerado o cenário *C2*, tanto para o conjunto de dados HMV_{CTI} (Figura 5.11c) quanto para o conjunto de dados HCPA (Figura 5.11b), obteve-se aumento nos níveis de sensibilidade. Entretanto, o ganho nos níveis de sensibilidade apenas foram obtidos devido a excessiva tendência dos classificadores em atribuir os pacientes à classe *vmi=sim*. Dessa forma, em ambos os casos observa-se a degradação dos níveis de especificidade e VPN, que por consequência, degradam os escores de Kappa e AUROC. É interessante observar que embora nesses casos os classificadores tenham excessivamente atribuído os pacientes à classe *vmi=sim*, a maioria dos classificadores obteve aumento em termos de F1 para a classe *vmi=sim* (F1+). Isto ocorre pois a métrica F1+ realiza a média harmônica entre sensibilidade e VPP, sendo que ambas as métricas assumem o mesmo peso durante a ponderação. Dessa forma, visto que para a maioria dos classificadores, o aumento nos níveis de sensibilidade excederam as perdas nos níveis de VPP, registrou-se aumento de F1+. Além disso, é possível observar que para a maioria dos classificadores, ocorreu a diminuição para a métrica ma-F1. A diminuição nos níveis de ma-F1 – que realiza a média entre as F1 de cada classe (*vmi=sim* e *vmi=não*) – é resultado da queda nos níveis de especificidade e VPN, as quais constituem as variáveis para o cálculo da F1- (F1 para a classe *vmi=não*). Como houve pequenas variações nos níveis de F1+ e degradações consideráveis nos níveis de especificidade e VPN, ocorreram reduções nos níveis de ma-F1. Tais comportamentos demonstram a importância de utilizar múltiplas métricas durante a etapa de avaliação, visto que cada métrica possui suas limitações.

Além das comparações entre os classificadores com múltiplas variáveis e os de variável única que utilizaram apenas o atributo *saturação de O2*, também treinamos classificadores de variável única utilizando *pCO2* como a variável independente para a predição de necessidade de VMI (Apêndice F). Os resultados foram semelhantes aos obtidos

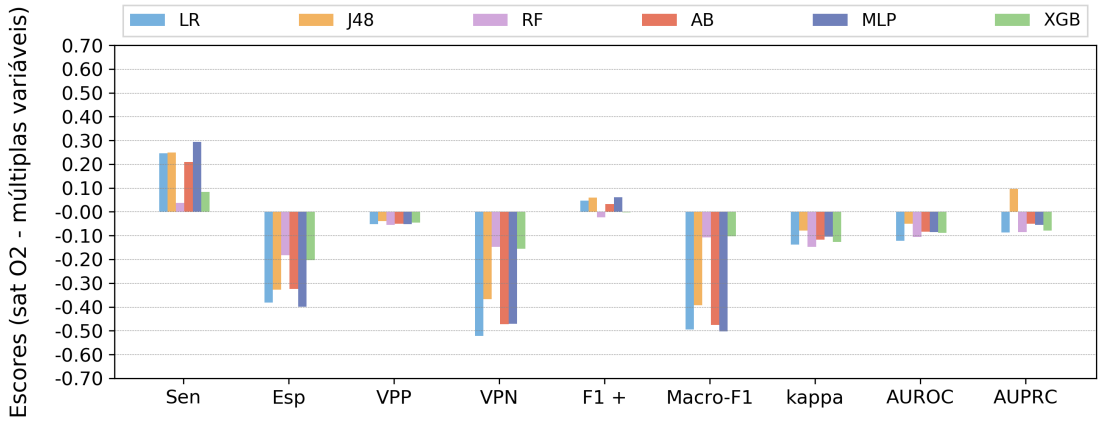
com os classificadores de variável única que utilizaram a variável *saturação de O2*.



(a) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no HMV



(b) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no $HCPA$



(c) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no HMV_{CTI}

Figura 5.11 – Comparação entre os classificadores de múltiplas variáveis e os de variável única (saturação de O2) para a predição de VMI

Fonte: O Autor

De modo geral, os classificadores que utilizaram apenas a idade do paciente como variável independente tenderam a apresentar índices de sensibilidade semelhantes ou até superiores aos classificadores treinados com múltiplas variáveis para a predição de mortalidade e internação em CTI, embora estes níveis tenham sido recorrentemente atingidos em detrimento dos níveis de precisão e especificidade. Tanto os níveis de sensibilidade quanto os níveis de precisão e revocação tenderam a apresentar diferenças estatisticamente significativas entre as versões dos classificadores de múltiplas variáveis e de apenas uma variável (abordagem alternativa). Os resultados dos testes de significância estatística podem ser consultados nas tabelas de apêndice E.1, E.2 e E.3 ($\alpha = 0.05$). Resultados semelhantes foram alcançados pelos classificadores treinados apenas com saturação de oxigênio e apenas pressão parcial de CO2 para a predição de necessidade de VMI. Os resultados obtidos indicam que mesmo que níveis de sensibilidade semelhantes ou superiores possam ser atingidos pelos classificadores que utilizaram apenas a variável idade e apenas saturação de oxigênio, os demais atributos tendem a ser importantes para alcançar níveis superiores de precisão e especificidade.

5.5 Sumário dos Principais Achados

Esta seção apresenta as considerações gerais acerca dos experimentos realizados.

5.5.1 Predição

Conforme os resultados obtidos, para todas as tarefas de predição avaliadas, nenhum algoritmo atingiu os melhores índices de desempenho para todas as métricas. Entretanto, quando considerada a métrica de sensibilidade – a qual mensura a capacidade dos classificadores em identificar os pacientes pertencentes às classes positivas (*mortalidade=sim* e *cti=sim*, *vmi=sim*) – o algoritmo *logistic regression* atingiu os maiores escores para todas as tarefas. Na sequência, *Adaboost* e *XGBoost* também tenderam a apresentar bons resultados nas tarefas avaliadas.

Embora a comparação direta dos resultados não tenha sido possível, visto que os conjuntos de dados utilizados nas avaliações não foram os mesmos dos conjuntos de dados utilizados por outros trabalhos, no geral, os resultados se assemelham aos reportados na literatura.

A avaliação cruzada entre os conjuntos de dados HMV e HCPA indicou que os classificadores podem ser sensíveis às características dos pacientes de cada hospital. Enquanto que a generalização parece ter sido alcançada ao treinar os classificadores com o conjunto de dados HCPA e testar no conjunto de dados HMV, o mesmo não ocorreu ao treinar os classificadores com o conjunto de dados HMVe testar com o conjunto de dados HCPA, resultando em considerável redução nos níveis de sensibilidade.

5.5.2 Atributos mais relevantes

Dentre as variáveis selecionadas pelos algoritmos, a idade foi recorrentemente elencada entre os atributos de maior contribuição, especialmente na tarefa de predição de mortalidade. Quando elencada, a idade avançada dos pacientes foi relacionada ao prognóstico desfavorável dos pacientes (*mortalidade=sim*). Para as tarefas de predição de internação em CTI e predição de necessidade de VMI, variáveis associadas à função respiratória foram elencadas como fatores agravantes para o prognóstico. Baixos índices de saturação de O2, bem como altos valores para pressão parcial de PO2 foram recorrentemente elencados como fatores relacionados à internação em CTI e necessidade de VMI. O histórico de doença cardíaca também foi um fator relacionado ao prognóstico desfavorável.

5.5.3 Métricas

Durante a análise dos resultados, foram observados aspectos interessantes acerca das métricas de avaliação que reforçam a necessidade de fazer uso de um conjunto de métricas complementares umas às outras, a fim de viabilizar a melhor análise comparativa entre os modelos preditivos propostos.

AUROC - Com base nos resultados obtidos, pôde-se observar que a métrica AUROC isoladamente não foi adequada para a comparação dos classificadores. Um exemplo disso pode ser observado na tarefa de predição de mortalidade (Tabela 5.2), a qual apresenta o maior desbalanceamento de classe entre as tarefas avaliadas. Em nossos experimentos, utilizamos o limiar de 0,5 (50%) sobre a probabilidade retornada pelos modelos preditivos para realizar a dicotomização das classes preditas. Na tarefa em questão, mesmo modelos com uma diferença de 1 pp para a métrica AUROC, chegaram a apresen-

tar uma diferença de 27 pp para a métrica sensibilidade (diferenças entre os algoritmos LR e RF para o conjunto de dados HMV). Isso ocorre pois a AUROC resume o desempenho dos modelos para todos os possíveis limiares de dicotomização. É importante observar que mesmo modelos que apresentam comportamentos distintos no que se refere a relação entre sensibilidade e taxa de falsos positivos ao considerar os diferentes limiares – o que resulta em diferentes curvas ROC – podem resultar em uma área sob a curva semelhante. Portanto, mesmo que a AUROC obtida pelos classificadores tenha sido semelhante, com o limiar de 0,5 os classificadores apresentaram desempenhos distintos no que se refere à capacidade de identificar os pacientes pertencentes às classes *mortalidade=sim*, *cti=sim* e *vmi=sim* (sensibilidade). Assim, é importante observar que a AUROC sumariza o comportamento do modelo como um todo, mas não é capaz de indicar que dois classificadores apresentam desempenho semelhante para uma determinada métrica, dado um limiar de dicotomização específico.

6 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho avaliamos seis algoritmos de AM para as tarefas de predição de mortalidade, predição de necessidade de internação em CTI e predição de necessidade de recursos de VMI.

Nenhum algoritmo atingiu os maiores escores para todas as métricas ou cenários avaliados. No geral, considerando a métrica de sensibilidade, os escores obtidos apontaram capacidade promissora dos classificadores em prever mortalidade, a internação em CTI e necessidade de VMI com base nas informações disponíveis no momento da admissão hospitalar. O algoritmo *logistic regression* obteve os maiores escores de sensibilidade, seguido pelos algoritmos, *AdaBoost* e *XGBoost*. A idade dos pacientes foi elencada entre os atributos de maior contribuição para a predição de mortalidade. Pacientes idosos foram mais suscetíveis ao prognóstico desfavorável (óbito). Variáveis associadas à função respiratória dos pacientes, como saturação de oxigênio e pressão parcial de CO2 foram elencadas como fatores relacionados à internação em CTI e necessidade de recursos de ventilação mecânica invasiva. Além desses, o histórico de doença cardíaca também foi elencado como um fator relacionado ao prognóstico desfavorável.

Os classificadores treinados apenas com a idade do paciente tenderam a apresentar índices de sensibilidade semelhantes ou até superiores aos classificadores treinados com múltiplas variáveis para a predição de mortalidade e internação em CTI, embora estes níveis tenham sido recorrentemente atingidos em detrimento dos níveis de precisão e especificidade. Resultados semelhantes foram alcançados pelos classificadores treinados apenas com saturação de oxigênio e apenas pressão parcial de CO2 para a predição de necessidade de VMI.

Além disso, desenvolvemos uma técnica de visualização de dados com base nos *shapley values* que permitiu estender a compreensão acerca dos fatores levados em consideração pelos classificadores durante as predições e, relaciona-los com os erros e acertos. É importante ressaltar que embora esta técnica tenha sido aplicada em tarefas de classificação com finalidade prognóstica no contexto da COVID-19, ela não restringe-se à tarefas de classificação, tampouco ao domínio ao qual foi aplicada.

6.1 Limitações

Os experimentos foram executados com a aplicação da técnica *cost-sensitive* para a penalização dos erros do tipo falso negativo. A penalização aplicada foi definida com base no desbalanceamento entre as classes da variável dependente de cada tarefa avaliada. Como cada tarefa apresenta diferentes graus de desbalanceamento entre as classes, cada tarefa foi avaliada com base em uma matriz de custo própria.

Conforme já mencionado, as ponderações foram definidas com base no desbalanceamento de classe observado em cada uma das B iterações da reamostragem *bootstrap* gerada para cada conjunto de dados. Para a tarefa de predição de mortalidade do conjunto de dados HMV, a ponderação aplicada nas 100 iterações de *bootstrap* resultou em valores de penalização para os erros do tipo falso negativo de em média 7,40 vezes $\pm$ 0,62, já para a tarefa de predição de mortalidade considerando apenas os pacientes internados na CTI (HMV$_{CTI}$), as penalizações aplicadas foram de em média 2,20 $\pm$ 0,26.

A diferença entre as penalizações aplicadas pode influenciar nos resultados dos classificadores. Portanto, para as tarefas de predição de mortalidade em CTI, internação em CTI e necessidade de VMI, as quais possuem um desbalanceamento de classe inferior à predição de mortalidade para todos os pacientes, aplicar matrizes de custo com maior penalização para os erros do tipo falso negativo poderia contribuir para resultados superiores em termos de sensibilidade, em comparação aos observados em nossos experimentos.

Nos experimentos realizados, os atributos com valores faltantes foram imputados a partir dos valores médios entre os valores observados em cada um dos atributos. Embora seja de fácil implementação, a imputação por valores médios pode alterar as características das distribuições dos dados. Com o objetivo de atenuar essas consequências indesejadas, técnicas mais sofisticadas de imputação de dados podem ser empregadas, como por exemplo *Multivariate Imputation By Chained Equations* (MICE) (BUUREN; GROOTHUIS-OUDSHOORN, 2011). Neste trabalho, não realizamos a avaliação do impacto da imputação multivariada em relação à imputação por valores médios.

Uma limitação da nossa técnica de visualização é que o SHAP é relativamente demorado quando aplicado a conjuntos de dados com um grande número de atributos(YANG, 2021). Essa limitação pode ser mitigada usando algoritmos SHAP otimizados (como TreeSHAP ou Fast treeSHAP (YANG, 2021)). Além do mais, visto que as visualizações são obtidas a partir da técnica t-SNE, a seleção dos parâmetros utilizados na execução do t-SNE, como perplexidade e taxa de aprendizado podem demandar ajustes.

6.2 Trabalhos futuros

Em nossos experimentos, utilizamos dois conjuntos de dados de dois hospitais situados no município de Porto Alegre, no estado do Rio Grande do Sul. Como trabalhos futuros, é necessário contemplar novos conjuntos de dados para reforçar os achados deste trabalho, bem como para estender a avaliação externa em conjuntos de dados provenientes de outras regiões do Brasil e até mesmo de outros países. Trabalhos futuros também podem contemplar a avaliação do efeito ao utilizar técnicas sofisticadas de imputação de dados, como MICE em comparação à imputação por valores médios. Também temos o intuito de expandir os testes com a técnica de visualização de dados empregada em nosso trabalho para outras tarefas Aprendizado de Máquina.

REFERÊNCIAS

ALBALLA, N.; AL-TURAIKI, I. Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: a review. **Informatics in Medicine Unlocked**, Elsevier, v. 24, p. 100564, 2021.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. [S.l.]: ACM press New York, 1999.

BI, X. et al. Prediction of severe illness due to covid-19 based on an analysis of initial fibrinogen to albumin ratio and platelet count. **Platelets**, Taylor & Francis, v. 31, n. 5, p. 674–679, 2020.

BISHOP, C. M. et al. **Neural networks for pattern recognition**. [S.l.]: Oxford university press, 1995.

BOOTH, A. L.; ABELS, E.; MCCAFFREY, P. Development of a prognostic model for mortality in covid-19 infection using machine learning. **Modern Pathology**, Nature Publishing Group, v. 34, n. 3, p. 522–531, 2021.

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

BURDICK, H. et al. Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial. **Computers in biology and medicine**, Elsevier, v. 124, p. 103949, 2020.

BURKART, N.; HUBER, M. F. A survey on the explainability of supervised machine learning. **Journal of Artificial Intelligence Research**, v. 70, p. 245–317, 2021.

BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate imputation by chained equations in r. **Journal of statistical software**, v. 45, p. 1–67, 2011.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

COVINO, M. et al. Predicting intensive care unit admission and death for covid-19 patients in the emergency department using early warning scores. **Resuscitation**, v. 156, 2020.

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: **Proceedings of the 23rd international conference on Machine learning**. [S.l.: s.n.], 2006. p. 233–240.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: CRC press, 1994.

FRANCONE, M. et al. Chest ct score in covid-19 patients: correlation with disease severity and short-term prognosis. **European radiology**, Springer, v. 30, n. 12, p. 6808–6817, 2020.

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

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of computer and system sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997.

FUTOMA, J. et al. The myth of generalisability in clinical research and machine learning in health care. **The Lancet Digital Health**, Elsevier, v. 2, n. 9, p. e489–e492, 2020.

GUPTA, R. K. et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with covid-19: an observational cohort study. **European Respiratory Journal**, Eur Respiratory Soc, v. 56, n. 6, 2020.

HALL, M. A. **Correlation-based feature selection for machine learning**. Thesis (PhD) — The University of Waikato, 1999.

HASTIE, T. et al. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, Ieee, v. 21, n. 9, p. 1263–1284, 2009.

HU, H.; YAO, N.; QIU, Y. Comparing rapid scoring systems in mortality prediction of critically ill patients with novel coronavirus disease. **Academic Emergency Medicine**, Wiley Online Library, v. 27, n. 6, p. 461–468, 2020.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intelligent data analysis**, IOS Press, v. 6, n. 5, p. 429–449, 2002.

JI, D. et al. Prediction for progression risk in patients with covid-19 pneumonia: the call score. **Clinical Infectious Diseases**, Oxford University Press US, v. 71, n. 6, p. 1393–1399, 2020.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–54, 2019.

KANG, S.-J.; JUNG, S. I. Age-related morbidity and mortality among patients with covid-19. **Infection & chemotherapy**, Korean Society of Infectious Diseases, v. 52, n. 2, p. 154, 2020.

KNIGHT, S. R. et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO clinical characterisation protocol: development and validation of the 4c mortality score. **BMJ**, v. 370, 2020.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, Springer, v. 5, n. 4, p. 221–232, 2016.

KUHN, D. M.; MOREIRA, V. P. Brcars: a dataset for fine-grained classification of car images. In: IEEE. **2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.], 2021. p. 231–238.

KVÅLSETH, T. O. Note on cohen's kappa. **Psychological reports**, SAGE Publications Sage CA: Los Angeles, CA, v. 65, n. 1, p. 223–226, 1989.

LAMBDEN, S. et al. The sofa score—development, utility and challenges of accurate assessment in clinical trials. **Critical Care**, BioMed Central, v. 23, n. 1, p. 1–9, 2019.

LEUNG, C. Risk factors for predicting mortality in elderly patients with covid-19: A review of clinical data in china. **Mechanisms of ageing and development**, Elsevier, v. 188, p. 111255, 2020.

LI, J. et al. Feature selection: A data perspective. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 50, n. 6, p. 1–45, 2017.

LIANG, W. et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19. **JAMA internal medicine**, American Medical Association, v. 180, n. 8, p. 1081–1089, 2020.

LIPPI, G.; LAVIE, C. J.; SANCHIS-GOMAR, F. Cardiac troponin i in patients with coronavirus disease 2019 (covid-19): evidence from a meta-analysis. **Progress in cardiovascular diseases**, Elsevier, v. 63, n. 3, p. 390, 2020.

LIU, H. et al. Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. **Journal of Data Science**, v. 3, n. 3, p. 257–278, 2005.

LIU, S. et al. Predictive performance of sofa and qsofa for in-hospital mortality in severe novel coronavirus disease. **The American Journal of Emergency Medicine**, Elsevier, v. 38, n. 10, p. 2074–2080, 2020.

LONGADGE, R.; DONGRE, S. Class imbalance problem in data mining review. **arXiv preprint arXiv:1305.1707**, 2013.

LUNDBERG, S. M.; ERION, G. G.; LEE, S.-I. Consistent individualized feature attribution for tree ensembles. **arXiv preprint arXiv:1802.03888**, 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. Curran Associates, Inc., p. 4765–4774, 2017.

MAATEN, L. V. D. et al. Dimensionality reduction: a comparative. **J Mach Learn Res**, v. 10, n. 66-71, p. 13, 2009.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. 11, 2008.

MERCÊS, S. O. das; LIMA, F. L. O.; NETO, J. R. T. de V. Associação da covid-19 com: idade e comorbidades médicas. **Research, Society and Development**, v. 9, n. 10, p. e1299108285–e1299108285, 2020.

MILLER, J. L. et al. Prediction models for severe manifestations and mortality due to covid-19: A systematic review. **Academic Emergency Medicine**, Wiley Online Library, v. 29, n. 2, p. 206–216, 2022.

MORGAN, R.; WILLIAMS, F.; WRIGHT, M. An early warning scoring system for detecting developing critical illness. **Clin Intensive Care**, v. 8, n. 2, p. 100, 1997.

MORGAN, R.; WRIGHT, M. In defence of early warning scores. **British Journal of Anaesthesia**, Oxford University Press, v. 99, n. 5, p. 747–748, 2007.

OLSSON, T.; TERÉNT, A.; LIND, L. Rapid emergency medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. **Journal of internal medicine**, Wiley Online Library, v. 255, n. 5, p. 579–587, 2004.

PAIVA, B. B. Miranda de et al. Effectiveness, explainability and reliability of machine meta-learning methods for predicting mortality in patients with covid-19: Results of the brazilian covid-19 registry. **medRxiv**, Cold Spring Harbor Laboratory Press, 2021.

PHYSICIANS, R. C. of. National early warning score (news) 2: standardising the assessment of acute-illness severity in the nhs. updated report of a working party 2017. 2021.

POLIKAR, R. Ensemble based systems in decision making. **IEEE Circuits and systems magazine**, IEEE, v. 6, n. 3, p. 21–45, 2006.

POLLARD, T. J. et al. tableone: An open source python package for producing summary statistics for research papers. **JAMIA open**, Oxford University Press, v. 1, n. 1, p. 26–31, 2018.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1135–1144.

ROBERTS, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. **Nature Machine Intelligence**, Nature Publishing Group, v. 3, n. 3, p. 199–217, 2021.

RODRIGUEZ, V. A. et al. Development and validation of prediction models for mechanical ventilation, renal replacement therapy, and readmission in covid-19 patients. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 28, n. 7, p. 1480–1488, 2021.

ROSTAMI, M.; MANSOURITORGHABEH, H. D-dimer level in covid-19 infection: a systematic review. **Expert review of hematology**, Taylor & Francis, v. 13, n. 11, p. 1265–1275, 2020.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 3, p. e0118432, 2015.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

SHAPLEY, L. S. A value for n-person games. In: ____. **Contributions to the Theory of Games (AM-28)**. [S.l.: s.n.], 1953. p. 307–318.

SOFAER, H. R.; HOETING, J. A.; JARNEVICH, C. S. The area under the precision-recall curve as a performance metric for rare binary events. **Methods in Ecology and Evolution**, Wiley Online Library, v. 10, n. 4, p. 565–577, 2019.

SUBBE, C. P. et al. Validation of a modified early warning score in medical admissions. **Qjm**, Oxford University Press, v. 94, n. 10, p. 521–526, 2001.

SUBUDHI, S.; VERMA, A.; PATEL, A. B. Prognostic machine learning models for covid-19 to facilitate decision making. **International Journal of clinical practice**, Wiley Online Library, v. 74, n. 12, p. e13685, 2020.

THAI-NGHE, N.; GANTNER, Z.; SCHMIDT-THIEME, L. Cost-sensitive learning methods for imbalanced data. In: IEEE. **The 2010 International joint conference on neural networks (IJCNN)**. [S.l.], 2010. p. 1–8.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

VINCENT, J.-L. et al. **The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure**. [S.l.]: Springer-Verlag, 1996.

WOLFF, R. F. et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. **Annals of internal medicine**, American College of Physicians, v. 170, n. 1, p. 51–58, 2019.

WRIGHT, R. E. Logistic regression. American Psychological Association, 1995.

WYNANTS, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. **bmj**, British Medical Journal Publishing Group, v. 369, 2020.

YADAW, A. S. et al. Clinical features of covid-19 mortality: development and validation of a clinical prediction model. **The Lancet Digital Health**, Elsevier, v. 2, n. 10, p. e516–e525, 2020.

YANG, J. Fast treeshap: Accelerating shap value computation for trees. **arXiv preprint arXiv:2109.09847**, 2021.

YU, L. et al. Machine learning methods to predict mechanical ventilation and mortality in patients with covid-19. **PLoS One**, Public Library of Science San Francisco, CA USA, v. 16, n. 4, p. e0249285, 2021.

ZHANG, L. et al. Persistent viral shedding lasting over 60 days in a mild covid-19 patient with ongoing positive sars-cov-2. **Quantitative Imaging in Medicine and Surgery**, AME Publications, v. 10, n. 5, p. 1141, 2020.

ZHAO, Z. et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. **PloS one**, v. 15, n. 7, 2020.

ZHOU, F. et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. **The lancet**, Elsevier, v. 395, n. 10229, p. 1054–1062, 2020.

APÊNDICE A — LISTA DOS ATRIBUTOS ADMISSIONAIS DISPONÍVEIS NOS CONJUNTOS DE DADOS

Tabela A.1 – Lista dos atributos disponíveis na admissão hospitalar

Atributo	Conjunto de dados	Descrição
sexo	HMV; HCPA	Sexo biológico do paciente
idade	HMV; HCPA	Idade do paciente (em anos)
IMC	HMV	Índice de Massa Corporal
tomo. típica	HMV	Indicação de tomografia computadorizada como contendo achados radiológicos típicos para COVID-19. Classificação realizada por radiologistas
tomo. pulm. compr.	HMV	Indicação de tomografia computadorizada com comprometimento pulmonar igual ou superior à 50%
sistólica	HMV	Pressão sistólica, ou máxima, é aquela que marca a contração do músculo cardíaco
diastólica	HMV	Pressão diastólica marca o momento de repouso, ou seja, relaxamento do músculo cardíaco, no qual os vasos permanecem abertos para o sangue passar
tmp. axilar	HMV	Temperatura axilar é a mensuração da temperatura corporal verificada usando um termômetro abaixo do braço/axila
freq. cardíaca	HMV	Frequência cardíaca é a velocidade do ciclo cardíaco medida pelo número de contrações do coração por minuto (bpm)
freq. respiratória	HMV	Frequência respiratória é o número de ciclos respiratórios (inspiração e expiração) contabilizados por minuto
saturação	HMV	Referente a saturação de oxigênio na circulação
eritrócitos	HMV; HCPA	Eritrócitos, também chamados de hemácias, são as células vermelhas do sangue, responsáveis pelo transporte de oxigênio aos tecidos
hemoglobina	HMV; HCPA	Hemoglobina é uma proteína presente no interior dos eritrócitos, responsável pela coloração do sangue e oxigenação dos tecidos
eritroblastos	HCPA	Precursores dos eritrócitos
neut. bastonetes	HCPA	Neutrófilos jovens
mielócitos	HCPA	Glóbulos brancos imaturos; precursores dos neutrófilos.
metamielócitos	HCPA	Glóbulos brancos imaturos; precursores dos neutrófilos.
plasmócitos	HCPA	Célula do sistema imune responsável pela produção de anticorpos (resposta imune adaptativa)
aspartato	HCPA	Enzima geralmente utilizada como marcador de lesão hepática

Fonte: O Autor

Tabela A.2 – Lista dos atributos disponíveis na admissão hospitalar

Atributo	Conjunto de dados	Descrição
alanina	HCPA	Enzima geralmente utilizada como marcador de lesão hepática
hematócrito	HMV; HCPA	Hematócrito é a percentagem de eritrócitos no sangue total (proporção entre a parte líquida e sólida do sangue)
leucócitos	HMV; HCPA	Leucócitos, também chamados de glóbulos brancos, são células que atuam na defesa do organismo, divididos em vários tipos celulares (ex.: neutrófilos, eosinófilos, basófilos, linfócitos, monócitos
neutrófilos	HMV; HCPA	Neutrófilos são os leucócitos que constituem a primeira linha de reconhecimento e defesa contra agentes infecciosos no tecido
eosinófilos	HMV; HCPA	Eosinófilos são os leucócitos mais envolvidos em processos alérgicos e parasitários
basófilos	HMV; HCPA	Basófilos são os leucócitos que produzem histamina e heparina, relacionados, principalmente, a processos alérgicos. Constituem menos de 2% das leucócitos no sangue
linfócitos	HMV; HCPA	Linfócitos são os leucócitos mais envolvidos em infecções virais e neoplasias; compõem a resposta imune adaptativa juntamente com os linfócitos B (plasmócitos)
monócitos	HMV; HCPA	Monócitos são os leucócitos que fazem parte do sistema mononuclear fagocítico
plaquetas	HMV; HCPA	Plaquetas são fragmentos citoplasmáticos originados da medula óssea, cuja principal função está associada com o processo de coagulação do sangue
D-dímeros	HMV; HCPA	Produtos da degradação da fibrina, frequentemente relacionado com anormalidades hemostáticas e trombose
troponina	HMV; HCPA	Enzima geralmente utilizada como marcador de lesão cardíaca
DHL	HMV; HCPA	Desidrogenase lactática é uma enzima presente em células de diferentes órgãos e tecidos. Seu aumento indica injúria celular
Fibrinogênio	HCPA	Proteína envolvida na coagulação sanguínea
prot. C-creativa	HMV; HCPA	Proteína de fase aguda cujo aumento está geralmente associado a processos inflamatórios
bilirrubina total	HMV; HCPA	Produzida pela fígado, os níveis da substâncias indicam comprometimento hepático
creatinina	HMV; HCPA	Produto do metabolismo muscular, é excretada pela urina em condições normais. Sendo assim, seu aumento é indicativo de insuficiência renal.
pH	HMV	Potencial hidrogeniônico; grau de acidez ou alcalinidade do sangue
pO2	HMV; HCPA	Pressão parcial de oxigênio (gasometria arterial)
pCO2	HMV; HCPA	Pressão parcial de gás carbônico (gasometria arterial)
HCO3	HMV; HCPA	Concentração de ânion bicarbonato no sangue (gasometria arterial)
BE	HMV; HCPA	*Base Excess* representa o total de bases no sangue (gasometria arterial)

Fonte: O Autor

Tabela A.3 – Lista dos atributos disponíveis na admissão hospitalar

Atributo	Conjunto de dados	Descrição
TCO2	HCPA	Teor de gás carbônico no plasma (gasometria arterial)
SO2	HMV; HCPA	Saturação de hemoglobina indica a fração de hemoglobina transportando oxigênio
potássio	HCPA	Indica o nível de potássio no sangue
magnésio	HCPA	Indica o nível de magnésio no sangue
sódio	HCPA	Indica o nível de sódio no sangue
uréia	HCPA	É produzida pelo fígado. Geralmente utilizada para analisar função/dano hepático e renal
TTPA	HCPA	Tempo de tromboplastina parcial ativada (TTPA) avalia a coagulação sanguínea
HCM	HCPA	Hemoglobina corpuscular média (HCM)
CHCM	HCPA	Concentração de hemoglobina corpuscular média (CHCM)
RDW	HCPA	*Red Cell Distribution Width (RDW)*, indica a variação de tamanho entre as hemácias
tempo protromb.	HCPA	Tempo de protrombina avalia a coagulação sanguínea
doença cardíaca	HMV	Informação sobre o acometimento de algum tipo de doença cardíaca
hipertensão	HMV	Informação sobre o acometimento de hipertensão
doença pulmonar	HMV	Informação sobre o acometimento de algum tipo de doença pulmonar
diabetes	HMV	Informação sobre o acometimento de algum tipo de *diabetes*
doença renal	HMV	Informação sobre o acometimento de algum tipo de doença renal
doença hepática	HMV	Informação sobre o acometimento de algum tipo de doença hepática
AVC	HMV	Informação sobre o acometimento/histórico de acidente vascular cerebral (AVC)
doença do sistema nervoso	HMV	Informação sobre o acometimento de algum tipo de doença do sistema nervoso
câncer	HMV	Informação sobre o acometimento de algum tipo de câncer
depressão	HMV	Informação sobre o acometimento de algum tipo de depressão
desfecho	HMV; HCPA	Informação do desfecho da hospitalização
cti	HMV	Identificação de passagem por Centro de Terapia Intensiva
vmi	HMV; HCPA	Identificação de necessidade de recursos de Ventilação Mecânica Invasiva

Fonte: O Autor

APÊNDICE B — CARACTERÍSTICAS DO CONJUNTO DE DADOS HMV

Tabela B.1 – Características dos pacientes do conjunto de dados HMV acometidos pela COVID-19, estratificado pelo desfecho

		mortalidade=não	mortalidade=sim	P-Value
n		1344	182	
sexo	feminino	549 (40.8)	80 (44.0)	0.472
	masculino	795 (59.2)	102 (56.0)	
idade		61.0 [47.0,72.0]	85.0 [78.0,91.0]	0.001
IMC		27.8 [25.1,31.5]	25.7 [23.2,28.8]	0.001
tomo. típica	não	188 (14.3)	64 (37.4)	0.001
	sim	1131 (85.7)	107 (62.6)	
tomo. pulm. compr.	não	998 (83.4)	72 (62.1)	0.001
	sim	198 (16.6)	44 (37.9)	
sistólica		126.0 [115.0,138.0]	123.0 [108.0,142.0]	0.346
diastólica		73.0 [64.0,81.0]	68.0 [58.0,78.0]	0.001
tmp. axilar		36.8 [36.2,37.5]	36.7 [36.0,37.5]	0.034
freq. cardíaca		91.0 [80.0,103.0]	90.0 [75.2,101.0]	0.240
freq. respiratória		20.0 [19.0,21.0]	21.0 [20.0,24.0]	0.001
saturação		95.0 [93.0,97.0]	94.0 [91.0,97.0]	0.001
eritrócitos		4.6 [4.2,4.9]	3.8 [3.2,4.3]	0.001
hemoglobina		13.7 [12.5,14.8]	11.6 [9.8,12.9]	0.001
hematócrito		39.9 [36.5,42.7]	35.0 [30.2,38.1]	0.001
leucócitos		6390.0 [4690.0,8682.5]	8385.0 [5370.0,12007.5]	0.001
neutrófilos		4645.0 [3087.5,6762.5]	6470.0 [4020.0,10310.0]	0.001
eosinófilos		10.0 [0.0,30.0]	10.0 [0.0,50.0]	0.074
basófilos		10.0 [10.0,20.0]	10.0 [10.0,20.0]	0.001
linfócitos		1010.0 [730.0,1330.0]	815.0 [562.5,1160.0]	0.001
monócitos		530.0 [360.0,740.0]	500.0 [340.0,830.0]	0.578
plaquetas		187000.0 [148000.0,242000.0]	182000.0 [135250.0,236750.0]	0.272
D-dímeros		690.0 [450.8,1075.0]	1330.0 [770.0,2354.5]	0.001
troponina		7.0 [5.0,12.0]	37.0 [20.5,81.2]	0.001
DHL		476.0 [342.0,604.0]	520.0 [402.0,740.0]	0.002
prot. C-reativa		5.0 [2.0,10.3]	8.5 [3.5,16.8]	0.001
bilirrubina total		0.3 [0.2,0.5]	0.4 [0.3,0.6]	0.007
creatinina		0.9 [0.8,1.1]	1.2 [0.9,1.7]	0.001

Fonte: O Autor

Tabela B.2 – **(Continuação)** Características dos pacientes do conjunto HMV acometidos pela COVID-19, estratificado pelo desfecho

		mortalidade=não	mortalidade=sim	P-Value
pH		7.5 [7.4,7.5]	7.4 [7.4,7.5]	0.001
pO2		75.0 [66.0,86.0]	74.0 [65.0,94.5]	0.295
pCO2		35.0 [32.0,38.0]	36.0 [32.0,43.0]	0.001
HCO3		24.0 [22.0,26.0]	24.0 [22.0,27.5]	0.097
BE		0.7 [-0.7,2.2]	1.0 [-1.2,3.3]	0.844
SO2		95.0 [93.0,97.0]	95.0 [92.0,97.0]	0.350
doença cardíaca	não	1146 (85.3)	101 (55.5)	0.001
	sim	198 (14.7)	81 (44.5)	
hipertensão	não	743 (55.3)	67 (36.8)	0.001
	sim	601 (44.7)	115 (63.2)	
doença pulmonar	não	1149 (85.5)	140 (76.9)	0.004
	sim	195 (14.5)	42 (23.1)	
diabetes	não	1092 (81.2)	130 (71.4)	0.003
	sim	252 (18.8)	52 (28.6)	
doença renal	não	1278 (95.1)	150 (82.4)	0.001
	sim	66 (4.9)	32 (17.6)	
doença hepática	não	1326 (98.7)	177 (97.3)	0.182
	sim	18 (1.3)	5 (2.7)	
AVC	não	1298 (96.6)	152 (83.5)	0.001
	sim	46 (3.4)	30 (16.5)	
doença s. nervoso	não	1285 (95.6)	148 (81.3)	0.001
	sim	59 (4.4)	34 (18.7)	
câncer	não	1237 (92.0)	154 (84.6)	0.002
	sim	107 (8.0)	28 (15.4)	
depressão	não	1229 (91.4)	169 (92.9)	0.615
	sim	115 (8.6)	13 (7.1)	
cti	não	1029 (76.6)	39 (21.4)	0.001
	sim	315 (23.4)	143 (78.6)	
vmi	não	1210 (90.0)	64 (35.2)	0.001
	sim	134 (10.0)	118 (64.8)	

Fonte: O Autor

Tabela B.3 – Características dos pacientes do conjunto HMV acometidos pela COVID-19, estratificado por necessidade de internação na CTI

		cti=não	cti=sim	P-Value
n		1068	458	
sexo	feminino	455 (42.6)	174 (38.0)	0.105
	masculino	613 (57.4)	284 (62.0)	
idade		60.0 [46.0,73.0]	71.0 [58.2,81.8]	0.001
IMC		27.6 [24.8,31.1]	27.7 [24.7,31.6]	0.959
tomo. típica	não	158 (15.1)	94 (21.1)	0.006
	sim	887 (84.9)	351 (78.9)	
tomo. pulm. compr.	não	836 (89.0)	234 (62.7)	0.001
	sim	103 (11.0)	139 (37.3)	
sistólica		126.0 [116.0,139.0]	125.0 [112.0,138.0]	0.123
diastólica		73.0 [64.0,82.0]	70.0 [61.0,79.0]	0.001
tmp. axilar		36.8 [36.2,37.5]	36.8 [36.2,37.6]	0.307
freq. cardíaca		91.0 [80.0,103.0]	90.0 [77.8,102.0]	0.092
freq. respiratória		20.0 [18.0,21.0]	21.0 [19.0,23.0]	0.001
saturação		95.0 [94.0,97.0]	94.0 [91.0,96.0]	0.001
eritrócitos		4.6 [4.2,4.9]	4.2 [3.6,4.7]	0.001
hemoglobina		13.8 [12.5,14.9]	12.6 [10.8,14.0]	0.001
hematócrito		40.1 [36.9,42.9]	37.3 [32.6,41.0]	0.001
leucócitos		6190.0 [4570.0,8145.0]	7845.0 [5582.5,10635.0]	0.001
neutrófilos		4350.0 [2985.0,6325.0]	6070.0 [3977.5,8720.0]	0.001
eosinófilos		10.0 [0.0,30.0]	10.0 [0.0,47.5]	0.404
basófilos		10.0 [10.0,20.0]	10.0 [10.0,20.0]	0.001
linfócitos		1010.0 [750.0,1340.0]	900.0 [630.0,1260.0]	0.001
monócitos		530.0 [360.0,740.0]	510.0 [340.0,787.5]	0.789
plaquetas		185000.0 [148750.0,236250.0]	190500.0 [141000.0,257250.0]	0.105
D-dímeros		658.0 [439.0,1041.2]	927.0 [537.0,1711.0]	0.001
troponina		7.0 [5.0,12.0]	11.0 [6.0,33.2]	0.001
DHL		466.0 [337.2,585.0]	514.5 [389.2,703.0]	0.001
prot. C-reativa		4.6 [1.8,9.3]	7.6 [3.4,14.1]	0.001
bilirrubina total		0.3 [0.2,0.5]	0.4 [0.3,0.5]	0.005
creatinina		0.9 [0.8,1.1]	1.0 [0.8,1.3]	0.011
pH		7.5 [7.4,7.5]	7.4 [7.4,7.5]	0.001

Fonte: O Autor

Tabela B.4 – **(Continuação)** Características dos pacientes do conjunto HMV acometidos pela COVID-19, estratificado por necessidade de internação na CTI

		cti=não	cti=sim	P-Value
pO2		75.0 [67.0,85.5]	74.0 [65.0,92.0]	0.681
pCO2		35.0 [32.0,37.0]	36.0 [33.0,40.0]	0.001
HCO3		24.0 [22.0,25.0]	24.0 [22.0,27.0]	0.002
BE		0.7 [-0.7,2.1]	0.9 [-0.9,3.1]	0.049
SO2		95.0 [94.0,97.0]	95.0 [92.0,97.0]	0.018
doença cardíaca	não	921 (86.2)	326 (71.2)	0.001
	sim	147 (13.8)	132 (28.8)	
hipertensão	não	618 (57.9)	192 (41.9)	0.001
	sim	450 (42.1)	266 (58.1)	
doença pulmonar	não	916 (85.8)	373 (81.4)	0.039
	sim	152 (14.2)	85 (18.6)	
diabetes	não	891 (83.4)	331 (72.3)	0.001
	sim	177 (16.6)	127 (27.7)	
doença renal	não	1025 (96.0)	403 (88.0)	0.001
	sim	43 (4.0)	55 (12.0)	
doença hepática	não	1050 (98.3)	453 (98.9)	0.520
	sim	18 (1.7)	5 (1.1)	
AVC	não	1021 (95.6)	429 (93.7)	0.144
	sim	47 (4.4)	29 (6.3)	
doença s. nervoso	não	1006 (94.2)	427 (93.2)	0.546
	sim	62 (5.8)	31 (6.8)	
câncer	não	983 (92.0)	408 (89.1)	0.077
	sim	85 (8.0)	50 (10.9)	
depressão	não	985 (92.2)	413 (90.2)	0.220
	sim	83 (7.8)	45 (9.8)	
vmi	não	1068 (100.0)	206 (45.0)	0.001
	sim		252 (55.0)	
mortalidade	não	1029 (96.3)	315 (68.8)	0.001
	sim	39 (3.7)	143 (31.2)	

Fonte: O Autor

Tabela B.5 – Características dos pacientes do conjunto HMV acometidos pela COVID-19, estratificado pela necessidade de VMI

		vmi=não	vmi=sim	P-Value
n		1274	252	
sexo	feminino	535 (42.0)	94 (37.3)	0.189
	masculino	739 (58.0)	158 (62.7)	
idade		61.0 [47.0,74.0]	73.0 [63.0,84.0]	0.001
IMC		27.6 [24.8,31.2]	27.8 [24.8,31.3]	0.675
tomo. típica	não	199 (16.0)	53 (21.8)	0.033
	sim	1048 (84.0)	190 (78.2)	
tomo. pulm. compr.	não	949 (85.4)	121 (60.2)	0.001
	sim	162 (14.6)	80 (39.8)	
sistólica		125.5 [114.8,138.0]	126.0 [115.0,140.0]	0.562
diastólica		73.0 [63.8,81.0]	69.5 [60.8,79.0]	0.001
tmp. axilar		36.8 [36.2,37.5]	36.8 [36.2,37.6]	0.378
freq. cardíaca		91.0 [80.0,103.0]	89.5 [77.0,101.2]	0.130
freq. respiratória		20.0 [19.0,21.0]	21.0 [19.8,24.0]	0.001
saturação		95.0 [93.0,97.0]	94.0 [90.0,96.0]	0.001
eritrócitos		4.6 [4.2,4.9]	4.1 [3.4,4.5]	0.001
hemoglobina		13.7 [12.5,14.9]	12.1 [10.3,13.5]	0.001
hematócrito		39.9 [36.6,42.8]	36.2 [31.8,39.7]	0.001
leucócitos		6305.0 [4650.0,8507.5]	8460.0 [5800.0,12170.0]	0.001
neutrófilos		4550.0 [3070.0,6530.0]	6600.0 [4082.5,10172.5]	0.001
eosinófilos		10.0 [0.0,30.0]	10.0 [0.0,50.0]	0.408
basófilos		10.0 [10.0,20.0]	10.0 [10.0,20.0]	0.001
linfócitos		1000.0 [730.0,1330.0]	900.0 [630.0,1262.5]	0.004
monócitos		530.0 [350.0,730.0]	515.0 [360.0,840.0]	0.416
plaquetas		185500.0 [148000.0,238000.0]	190000.0 [140000.0,265000.0]	0.185
D-dímeros		682.0 [443.2,1075.0]	1020.0 [613.0,1816.0]	0.001
troponina		7.0 [5.0,13.0]	16.0 [7.0,40.5]	0.001
DHL		474.0 [340.0,594.2]	520.0 [402.8,736.2]	0.001
prot. C-reativa		4.9 [1.9,9.8]	8.6 [3.5,15.9]	0.001
bilirrubina total		0.3 [0.2,0.5]	0.4 [0.2,0.5]	0.754
creatinina		0.9 [0.8,1.1]	1.0 [0.8,1.4]	0.005

Fonte: O Autor

Tabela B.6 – **(Continuação)** Características dos pacientes do conjunto HMV acometidos pela COVID-19, estratificado pela necessidade de VMI

		vmi=não	vmi=sim	P-Value
pH		7.5 [7.4,7.5]	7.4 [7.4,7.5]	0.001
pO2		74.0 [67.0,85.0]	76.0 [65.0,98.0]	0.165
pCO2		35.0 [32.0,38.0]	36.0 [32.0,42.0]	0.001
HCO3		24.0 [22.0,25.0]	24.0 [22.0,28.0]	0.027
BE		0.7 [-0.7,2.2]	0.9 [-1.2,3.4]	0.429
SO2		95.0 [93.0,97.0]	95.0 [92.0,97.0]	0.768
doença cardíaca	não	1071 (84.1)	176 (69.8)	0.001
	sim	203 (15.9)	76 (30.2)	
hipertensão	não	713 (56.0)	97 (38.5)	0.001
	sim	561 (44.0)	155 (61.5)	
doença pulmonar	não	1092 (85.7)	197 (78.2)	0.003
	sim	182 (14.3)	55 (21.8)	
diabetes	não	1047 (82.2)	175 (69.4)	0.001
	sim	227 (17.8)	77 (30.6)	
doença renal	não	1217 (95.5)	211 (83.7)	0.001
	sim	57 (4.5)	41 (16.3)	
doença hepática	não	1256 (98.6)	247 (98.0)	0.568
	sim	18 (1.4)	5 (2.0)	
AVC	não	1222 (95.9)	228 (90.5)	0.001
	sim	52 (4.1)	24 (9.5)	
doença s. nervoso	não	1200 (94.2)	233 (92.5)	0.365
	sim	74 (5.8)	19 (7.5)	
câncer	não	1167 (91.6)	224 (88.9)	0.206
	sim	107 (8.4)	28 (11.1)	
depressão	não	1172 (92.0)	226 (89.7)	0.278
	sim	102 (8.0)	26 (10.3)	
cti	não	1068 (83.8)		0.001
	sim	206 (16.2)	252 (100.0)	
mortalidade	não	1210 (95.0)	134 (53.2)	0.001
	sim	64 (5.0)	118 (46.8)	

Fonte: O Autor

APÊNDICE C — CARACTERÍSTICAS DO CONJUNTO DE DADOS HCPA

Tabela C.1 – Características dos pacientes do conjunto HCPA acometidos pela COVID-19, estratificado pelo desfecho

		mortalidade=não	mortalidade=sim	P-Value
n		1420	849	
idade		55.0 [42.0,65.0]	67.0 [57.0,75.0]	0.001
sexo	feminino	635 (44.7)	393 (46.3)	0.494
	masculino	785 (55.3)	456 (53.7)	
plaquetas		221.0 [170.8,275.2]	199.5 [149.8,270.0]	0.001
leucócitos		8.5 [6.3,11.7]	9.6 [6.7,13.3]	0.001
neutrófilos		6.8 [4.8,9.6]	8.0 [5.3,11.3]	0.001
linfócito		0.9 [0.6,1.2]	0.7 [0.4,1.0]	0.001
monócito		0.4 [0.3,0.6]	0.4 [0.2,0.7]	0.043
eosinófilos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.014
basófilos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.476
eritrócitos		4.5 [4.0,4.8]	4.2 [3.6,4.7]	0.001
VCM		87.1 [84.0,90.0]	88.5 [84.9,92.7]	0.001
HCM		29.2 [27.9,30.3]	29.6 [28.2,30.8]	0.001
CHCM		33.5 [32.6,34.2]	33.2 [32.2,34.2]	0.001
RDW		13.2 [12.5,13.9]	13.7 [13.0,14.8]	0.001
hematócrito		38.7 [35.1,41.7]	37.2 [32.2,40.8]	0.001
hemoglobina		12.9 [11.7,14.1]	12.3 [10.5,13.8]	0.001
eritroblastos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.001
neutrófilos bast		0.0 [0.0,0.1]	0.0 [0.0,0.2]	0.243
mielócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.111
metamielócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.148
plasmócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.942
TCO2		25.0 [22.4,27.4]	23.4 [20.4,26.3]	0.001
BE		0.1 [-2.5,2.2]	-2.3 [-5.9,0.5]	0.001
pO2		84.0 [66.0,117.0]	87.7 [67.7,131.0]	0.031
pCO2		37.2 [33.1,42.3]	37.9 [32.4,48.7]	0.005
SO2		96.4 [93.2,98.3]	96.2 [92.5,98.4]	0.427
HCO3		23.8 [21.3,26.1]	22.1 [19.3,24.8]	0.001
uréia		37.0 [27.0,57.0]	61.0 [41.0,101.0]	0.001

Fonte: O Autor

Tabela C.2 – (**Continuação**) Características dos pacientes do conjunto HCPA acometidos pela COVID-19, estratificado pelo desfecho

		mortalidade=não	mortalidade=sim	P-Value
TTPA		34.8 [31.6,38.9]	35.9 [32.0,41.7]	0.001
troponina		9.9 [9.9,20.2]	24.3 [9.9,92.5]	0.001
tempo protrombina		13.4 [12.8,14.3]	14.1 [13.2,15.5]	0.001
potássio		4.2 [3.8,4.6]	4.4 [3.9,4.8]	0.001
magnésio		2.0 [1.8,2.2]	2.1 [1.9,2.4]	0.008
DHL		437.0 [328.0,590.5]	535.0 [384.2,743.5]	0.001
fibrinogênio		631.0 [534.0,722.0]	615.5 [502.0,730.2]	0.094
D-dímeros		1.0 [0.6,1.9]	1.7 [0.9,5.2]	0.001
bilirrubina total		0.5 [0.3,0.7]	0.5 [0.3,0.8]	0.180
aspartato		44.0 [30.0,69.0]	48.0 [33.0,74.0]	0.001
alanina		41.0 [25.0,66.5]	34.0 [22.0,54.8]	0.001
prot. C-reativa		122.4 [69.9,191.9]	150.2 [87.0,228.2]	0.001
sódio		138.0 [136.0,141.0]	138.0 [135.0,142.0]	0.624
creatinina		0.9 [0.8,1.3]	1.3 [0.9,2.1]	0.001
vmi	não	786 (55.4)	141 (16.6)	0.001
	sim	634 (44.6)	708 (83.4)	

Fonte: O Autor

Tabela C.3 – Características dos pacientes do conjunto HCPA acometidos pela COVID-19, estratificado por VMI

		vmi=não	vmi=sim	P-Value
n		927	1342	
idade		60.0 [44.0,71.0]	60.0 [48.0,69.0]	0.978
sexo	feminino	445 (48.0)	583 (43.4)	0.035
	masculino	482 (52.0)	759 (56.6)	
plaquetas		219.0 [169.0,275.5]	211.0 [157.0,272.0]	0.024
leucócitos		8.3 [6.1,10.9]	9.4 [6.7,13.1]	0.001
neutrófilos		6.5 [4.6,9.0]	7.8 [5.3,11.3]	0.001
linfócito		0.9 [0.6,1.3]	0.7 [0.5,1.1]	0.001
monócito		0.5 [0.3,0.7]	0.4 [0.2,0.6]	0.001
eosinófilos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.001
basófilos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.004
eritrócitos		4.4 [3.9,4.8]	4.3 [3.8,4.8]	0.114
VCM		87.1 [84.0,90.1]	87.9 [84.6,91.6]	0.001
HCM		29.3 [27.9,30.5]	29.4 [28.1,30.5]	0.226
CHCM		33.5 [32.6,34.3]	33.3 [32.4,34.2]	0.005
RDW		13.2 [12.5,14.1]	13.3 [12.7,14.4]	0.001
hematócrito		38.1 [34.4,41.6]	38.2 [34.0,41.4]	0.837
hemoglobina		12.8 [11.4,14.0]	12.7 [11.1,14.0]	0.297
eritroblastos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.001
neutrófilos bast		0.0 [0.0,0.0]	0.0 [0.0,0.2]	0.001
mielócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.023
metamielócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.181
plasmócitos		0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.248
TCO2		24.7 [21.6,26.7]	24.2 [21.4,27.1]	0.928
BE		0.3 [-2.5,2.0]	-1.2 [-4.6,1.3]	0.001
pO2		82.6 [66.0,107.5]	87.6 [67.1,126.5]	0.005
pCO2		35.7 [31.9,39.1]	38.9 [33.5,47.8]	0.001
SO2		96.2 [93.3,98.2]	96.3 [92.6,98.4]	0.918
HCO3		23.6 [20.6,25.5]	23.0 [20.3,25.7]	0.423
uréia		39.0 [28.0,65.0]	48.0 [33.0,80.0]	0.001

Fonte: O Autor

Tabela C.4 – (**Continuação**) Características dos pacientes do conjunto HCPA acometidos pela COVID-19, estratificado por VMI

		vmi=não	vmi=sim	P-Value
TTPA		34.9 [31.6,39.2]	35.5 [31.9,40.5]	0.017
troponina		9.9 [9.9,22.1]	13.7 [9.9,55.0]	0.001
tempo protrombina		13.5 [12.8,14.4]	13.7 [13.0,14.8]	0.001
potássio		4.2 [3.8,4.6]	4.3 [3.9,4.8]	0.008
magnésio		2.0 [1.8,2.2]	2.1 [1.9,2.3]	0.001
DHL		398.0 [304.0,527.0]	525.5 [382.0,714.8]	0.001
fibrinogênio		611.5 [509.0,705.0]	636.0 [534.0,736.5]	0.001
D-dímeros		1.0 [0.6,1.8]	1.4 [0.7,3.8]	0.001
bilirrubina total		0.5 [0.3,0.7]	0.5 [0.4,0.7]	0.002
aspartato		40.0 [28.0,64.0]	49.0 [33.0,74.0]	0.001
alanina		37.0 [21.2,62.0]	39.0 [25.0,62.0]	0.023
prot. C-reativa		112.5 [60.5,172.2]	150.1 [86.5,227.1]	0.001
sódio		138.0 [135.0,140.0]	139.0 [136.0,142.0]	0.001
creatinina		1.0 [0.8,1.4]	1.1 [0.8,1.7]	0.001
mortalidade	não	786 (84.8)	634 (47.2)	0.001
	sim	141 (15.2)	708 (52.8)	

Fonte: O Autor

APÊNDICE D — RESULTADOS DOS CLASSIFICADORES

Tabela D.1 – Resultado de classificação para a tarefa de predição mortalidade

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**,84 [,72-,94]**	,87 [,84-,89]	,46 [,40-,53]	**,98 [,96-,99]**	,59 [,52-,65]	,76 [,72-,79]	,52 [,44-,59]	,92 [,89-,95]	,60 [,48-,69]
		J48	,54 [,41-,68]	,91 [,88-,94]	,46 [,38-,55]	,94 [,91-,96]	,50 [,41-,58]	,71 [,66-,76]	,42 [,33-,52]	,73 [,67-,79]	,44 [,31-,56]
		RF	,57 [,41-,73]	**,95 [,92-,97]**	**,59 [,50-,73]**	,94 [,92-,97]	,58 [,49-,66]	,76 [,71-,81]	,52 [,42-,62]	**,93 [,90-,95]**	**,61 [,50-,71]**
		AB	,78 [,62-,93]	,85 [,80-,91]	,41 [,32-,50]	,97 [,94-,99]	,54 [,45-,61]	,72 [,67-,77]	,45 [,36-,54]	,89 [,86-,92]	,52 [,40-,64]
		MLP	,62 [,46-,83]	,92 [,86-,95]	,51 [,41-,63]	,95 [,92-,97]	,56 [,45-,65]	,74 [,69-,79]	,49 [,38-,59]	,87 [,80-,92]	,57 [,47-,70]
		XGB	,72 [,60-,84]	,91 [,88-,94]	,52 [,44-,63]	,96 [,94-,98]	**,60 [,52-,68]**	**,77 [,72-,81]**	**,54 [,53-,55]**	,92 [,90-,94]	,60 [,50-,69]
$C2$	$HCPA$	LR	**,68 [,63-,72]**	,73 [,69-,77]	,60 [,56-,64]	,79 [,76-,82]	,64 [,60-,66]	,70 [,67-,72]	,39 [,34-,45]	,77 [,74-,79]	,65 [,60-,68]
		J48	,56 [,48-,62]	,70 [,65-,74]	,53 [,48-,57]	,73 [,69-,76]	,54 [,49-,58]	,63 [,60-,65]	,25 [,19-,31]	,62 [,58-,67]	,53 [,48-,59]
		RF	,62 [,57-,67]	**,77 [,74-,82]**	**,62 [,58-,68]**	,77 [,74-,80]	,62 [,59-,65]	,70 [,67-,72]	,40 [,34-,45]	,78 [,75-,81]	,66 [,62-,70]
		AB	,67 [,61-,75]	,71 [,63-,76]	,58 [,54-,62]	,78 [,75-,82]	,62 [,59-,66]	,69 [,66-,71]	,37 [,32-,42]	,76 [,74-,79]	,64 [,59-,69]
		MLP	,67 [,52-,83]	,70 [,54-,81]	,57 [,50-,63]	,78 [,72-,85]	,61 [,55-,66]	,67 [,64-,70]	,35 [,29-,41]	,75 [,71-,78]	,61 [,57-,67]
		XGB	**,68 [,63-,74]**	,74 [,70-,78]	,61 [,56-,65]	**,79 [,76-,83]**	**,64 [,60-,68]**	**,70 [,68-,73]**	**,40 [,39-,41]**	**,78 [,75-,81]**	**,67 [,61-,71]**
$C2$	HMV_{CTI}	LR	**,80 [,67-,89]**	,82 [,75-,88]	,66 [,55-,76]	**,90 [,84-,94]**	**,72 [,64-,80]**	**,79 [,73-,84]**	**,58 [,46-,68]**	**,87 [,83-,92]**	**,73 [,63-,84]**
		J48	,68 [,53-,81]	,80 [,70-,86]	,60 [,50-,71]	,84 [,77-,92]	,63 [,53-,73]	,73 [,67-,79]	,46 [,33-,59]	,73 [,62-,83]	,58 [,42-,72]
		RF	,76 [,65-,88]	**,83 [,75-,90]**	**,67 [,56-,78]**	,88 [,82-,94]	,71 [,63-,79]	,78 [,73-,84]	,57 [,45-,68]	,87 [,81-,92]	,69 [,53-,81]
		AB	,77 [,59-,93]	,80 [,72-,88]	,64 [,52-,78]	,89 [,81-,96]	,70 [,59-,80]	,77 [,70-,84]	,54 [,39-,68]	,86 [,80-,91]	,69 [,53-,82]
		MLP	,69 [,53-,82]	**,83 [,70-,89]**	,64 [,51-,75]	,86 [,78-,92]	,66 [,55-,76]	,75 [,69-,82]	,50 [,37-,63]	,83 [,75-,89]	,69 [,59-,79]
		XGB	,73 [,61-,84]	**,83 [,76-,88]**	,66 [,53-,76]	,87 [,81-,92]	,69 [,58-,77]	,77 [,70-,83]	,54 [,53-,56]	,87 [,81-,92]	,71 [,57-,83]

Tabela D.2 – Resultado de classificação para a tarefa de predição de necessidade de CTI

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**,66 [,00-,00]**	,77 [,00-,00]	,55 [,00-,00]	**,84 [,00-,00]**	**,60 [,00-,00]**	**,70 [,00-,00]**	,40 [,00-,00]	,78 [,00-,00]	**,63 [,00-,00]**
		J48	,52 [,00-,00]	,74 [,00-,00]	,46 [,00-,00]	,78 [,00-,00]	,49 [,00-,00]	,62 [,00-,00]	,25 [,00-,00]	,62 [,00-,00]	,48 [,00-,00]
		RF	,53 [,00-,00]	**,86 [,00-,00]**	**,62 [,00-,00]**	,81 [,00-,00]	,57 [,00-,00]	**,70 [,00-,00]**	**,41 [,00-,00]**	**,79 [,00-,00]**	**,63 [,00-,00]**
		AB	,61 [,00-,00]	,75 [,00-,00]	,51 [,00-,00]	,82 [,00-,00]	,55 [,00-,00]	,67 [,00-,00]	,34 [,00-,00]	,74 [,00-,00]	,57 [,00-,00]
		MLP	,58 [,00-,00]	,77 [,00-,00]	,52 [,00-,00]	,81 [,00-,00]	,55 [,00-,00]	,67 [,00-,00]	,34 [,00-,00]	,73 [,00-,00]	,59 [,00-,00]
		XGB	,60 [,00-,00]	,80 [,00-,00]	,57 [,00-,00]	,83 [,00-,00]	,59 [,00-,00]	**,70 [,00-,00]**	,40 [,00-,00]	,78 [,00-,00]	,62 [,00-,00]

Tabela D.3 – Resultado de classificação para a tarefa de predição de necessidade de VMI

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C1$	HMV	LR	**,67 [,57-,78]**	,77 [,72-,81]	,36 [,30-,43]	**,92 [,89-,95]**	**,47 [,41-,55]**	,65 [,62-,70]	**,33 [,26-,41]**	**,80 [,76-,84]**	**,47 [,39-,58]**
		J48	,40 [,29-,48]	,83 [,80-,87]	,32 [,24-,39]	,88 [,85-,90]	,35 [,27-,42]	,60 [,56-,64]	,21 [,12-,29]	,61 [,51-,67]	,32 [,25-,40]
		RF	,35 [,25-,44]	**,93 [,90-,96]**	**,50 [,41-,62]**	,88 [,85-,91]	,41 [,33-,47]	**,66 [,61-,69]**	,32 [,23-,39]	,79 [,74-,82]	,45 [,39-,54]
		AB	,61 [,43-,77]	,74 [,62-,84]	,32 [,24-,41]	,91 [,87-,94]	,42 [,34-,50]	,62 [,55-,68]	,26 [,17-,37]	,74 [,69-,80]	,39 [,31-,48]
		MLP	,54 [,40-,71]	,82 [,70-,89]	,37 [,27-,47]	,90 [,87-,93]	,43 [,35-,50]	,64 [,59-,69]	,30 [,19-,37]	,74 [,69-,78]	,43 [,34-,50]
		XGB	,53 [,44-,64]	,85 [,81-,89]	,40 [,35-,48]	,90 [,87-,93]	,46 [,39-,51]	**,66 [,63-,70]**	**,33 [,32-,34]**	,78 [,74-,82]	,44 [,37-,54]
$C2$	$HCPA$	LR	**,89 [,81-,95]**	,24 [,10-,36]	,63 [,59-,68]	,60 [,52-,68]	,74 [,72-,76]	,53 [,46-,60]	,14 [,07-,22]	,70 [,67-,72]	,78 [,75-,81]
		J48	,70 [,64-,76]	**,49 [,41-,56]**	,66 [,62-,70]	,53 [,47-,58]	,68 [,65-,71]	,59 [,56-,63]	,19 [,12-,27]	,61 [,57-,64]	,69 [,64-,74]
		RF	,81 [,76-,86]	,46 [,38-,54]	**,69 [,64-,73]**	,63 [,58-,68]	,74 [,72-,76]	**,64 [,60-,66]**	**,28 [,22-,33]**	,73 [,71-,76]	,81 [,78-,83]
		AB	,87 [,74-,98]	,33 [,04-,52]	,65 [,59-,70]	**,64 [,52-,74]**	,74 [,71-,77]	,57 [,40-,64]	,21 [,02-,30]	,71 [,68-,74]	,80 [,77-,82]
		MLP	,80 [,58-,98]	,38 [,04-,73]	,66 [,59-,74]	,58 [,50-,71]	,72 [,65-,76]	,57 [,41-,65]	,19 [,02-,30]	,69 [,64-,72]	,77 [,73-,81]
		XGB	,83 [,78-,88]	,44 [,36-,52]	,68 [,64-,72]	**,64 [,59-,71]**	**,75 [,73-,77]**	,63 [,60-,66]	,28 [,26-,29]	**,74 [,71-,76]**	**,82 [,79-,84]**
$C2$	HMV_{CTI}	LR	**,73 [,52-,93]**	,40 [,15-,63]	**,61 [,51-,70]**	**,55 [,42-,74]**	**,65 [,57-,72]**	,55 [,46-,61]	**,13 [,01-,23]**	**,62 [,55-,68]**	**,68 [,60-,75]**
		J48	,64 [,46-,82]	,44 [,23-,63]	,59 [,50-,69]	,49 [,37-,62]	,61 [,52-,68]	,53 [,44-,61]	,09 [-,03-,22]	,55 [,46-,65]	,64 [,52-,73]
		RF	,70 [,59-,83]	,43 [,28-,62]	**,61 [,53-,71]**	,53 [,42-,65]	**,65 [,59-,70]**	**,56 [,50-,63]**	**,13 [,02-,26]**	,61 [,53-,67]	,66 [,56-,75]
		AB	,72 [,53-,92]	,40 [,12-,61]	,60 [,52-,70]	,53 [,38-,67]	**,65 [,55-,71]**	,54 [,43-,62]	,12 [-,01-,24]	,60 [,53-,66]	,66 [,55-,75]
		MLP	,63 [,38-,85]	**,47 [,18-,76]**	**,61 [,51-,72]**	,51 [,42-,61]	,61 [,48-,69]	,54 [,45-,62]	,11 [-,02-,26]	,59 [,51-,66]	,66 [,56-,75]
		XGB	,68 [,56-,80]	,43 [,29-,58]	,60 [,52-,69]	,52 [,41-,62]	,63 [,57-,69]	,55 [,49-,61]	,10 [,08-,12]	,59 [,52-,65]	,65 [,56-,74]

Tabela D.4 – Resultado de classificação para a tarefa de predição de mortalidade– validação cruzada entre os conjuntos de dados HMV e HCPA

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C2$	$HCPA(treino)$ $HMV_{CTI}(teste)$	LR	**,91 [,84-,97]**	,66 [,59-,73]	,55 [,48-,63]	**,94 [,90-,98]**	**,68 [,62-,74]**	**,73 [,69-,77]**	**,48 [,41-,56]**	**,88 [,84-,92]**	**,77 [,69-,85]**
		J48	,59 [,42-,75]	,66 [,55-,76]	,44 [,34-,54]	,78 [,70-,85]	,50 [,39-,61]	,61 [,53-,68]	,22 [,07-,37]	,63 [,52-,73]	,49 [,37-,60]
		RF	,73 [,60-,87]	**,76 [,67-,84]**	**,58 [,49-,67]**	,86 [,80-,93]	,64 [,57-,72]	,72 [,67-,78]	,45 [,35-,56]	,82 [,77-,87]	,62 [,52-,73]
		AB	,83 [,60-1,05]	,63 [,42-,83]	,51 [,41-,61]	,90 [,80-,99]	,62 [,54-,71]	,68 [,60-,76]	,39 [,27-,51]	,80 [,74-,86]	,60 [,48-,71]
		MLP	,79 [,60-,98]	,67 [,50-,83]	,52 [,42-,63]	,88 [,79-,97]	,63 [,54-,71]	,69 [,61-,77]	,40 [,27-,53]	,80 [,74-,87]	,64 [,53-,75]
		XGB	,86 [,75-,96]	,62 [,55-,70]	,51 [,44-,58]	,91 [,84-,97]	,64 [,57-,71]	,69 [,64-,74]	,40 [,38-,42]	,81 [,76-,87]	,63 [,52-,74]
$C2$	$HMV_{CTI}(treino)$ $HCPA(teste)$	LR	**,48 [,36-,59]**	,83 [,75-,91]	,63 [,56-,70]	**,73 [,69-,77]**	**,54 [,47-,61]**	**,66 [,62-,70]**	**,32 [,25-,39]**	,73 [,69-,77]	**,60 [,54-,65]**
		J48	,40 [,20-,60]	,78 [,52-1,03]	,54 [,40-,67]	,68 [,63-,74]	,45 [,35-,55]	,58 [,50-,67]	,19 [,04-,33]	,58 [,49-,67]	,52 [,44-,60]
		RF	,39 [,27-,51]	,87 [,80-,93]	,64 [,58-,70]	,71 [,67-,74]	,48 [,39-,56]	,63 [,59-,67]	,28 [,21-,34]	**,74 [,71-,77]**	**,60 [,56-,65]**
		AB	,34 [,20-,48]	**,89 [,83-,95]**	**,65 [,58-,72]**	,69 [,66-,73]	,44 [,33-,55]	,61 [,55-,67]	,25 [,17-,34]	,72 [,68-,76]	**,60 [,55-,65]**
		MLP	,39 [,28-,50]	,85 [,78-,91]	,60 [,53-,67]	,70 [,67-,73]	,47 [,38-,55]	,62 [,57-,66]	,25 [,18-,33]	,69 [,65-,72]	,58 [,53-,64]
		XGB	,38 [,24-,52]	,87 [,79-,95]	,64 [,56-,71]	,70 [,67-,74]	,47 [,37-,57]	,62 [,58-,67]	,27 [,25-,28]	,73 [,69-,76]	**,60 [,55-,65]**

Tabela D.5 – Resultado de classificação para a tarefa de predição de necessidade de VMI– validação cruzada entre os conjuntos de dados HMV e HCPA

Cenário	Conjunto de Dados	Algoritmo	Sen	Esp	VPP	VPN	$F1_+$	ma-F1	Kappa	AUROC	AUPRC
$C2$	$HCPA(treino)$ $HMV_{CTI}(teste)$	LR	,75 [,61-,89]	,32 [,18-,46]	,57 [,52-,64]	**,51 [,41-,62]**	**,65 [,57-,72]**	**,52 [,46-,58]**	**,07 [-,03-,17]**	**,58 [,52-,64]**	**,66 [,59-,73]**
		J48	,60 [,49-,76]	**,44 [,31-,55]**	,57 [,49-,63]	,47 [,37-,56]	,58 [,50-,67]	**,52 [,45-,57]**	,04 [-,09-,15]	,53 [,45-,59]	,62 [,53-,69]
		RF	,72 [,62-,81]	,31 [,21-,45]	,56 [,50-,64]	,47 [,39-,57]	,63 [,57-,69]	,50 [,45-,56]	,03 [-,07-,12]	,56 [,50-,62]	,64 [,56-,70]
		AB	**,79 [,56-,95]**	,23 [,04-,48]	,56 [,50-,61]	,47 [,31-,61]	**,65 [,56-,72]**	,47 [,38-,56]	,03 [-,06-,13]	,56 [,49-,61]	,63 [,56-,69]
		MLP	,69 [,47-,90]	,39 [,17-,66]	**,58 [,51-,68]**	,50 [,41-,64]	,62 [,53-,70]	**,52 [,45-,61]**	**,07 [-,04-,24]**	**,58 [,51-,65]**	,65 [,58-,72]
		XGB	,67 [,56-,75]	,35 [,23-,48]	,56 [,50-,64]	,46 [,37-,57]	,61 [,54-,66]	,50 [,44-,56]	,01 [-,02-,02]	,54 [,48-,61]	,63 [,55-,69]
$C2$	$HMV_{CTI}(treino)$ $HCPA(teste)$	LR	**,76 [,52-,90]**	,33 [,13-,55]	,62 [,59-,66]	,49 [,42-,57]	**,68 [,56-,73]**	,53 [,47-,58]	,09 [,02-,18]	,60 [,52-,66]	,69 [,63-,75]
		J48	,58 [,42-,77]	**,49 [,29-,65]**	,62 [,57-,68]	,45 [,39-,52]	,60 [,50-,69]	,53 [,48-,59]	,07 [-,03-,19]	,54 [,47-,61]	,67 [,60-,73]
		RF	,67 [,54-,83]	,47 [,29-,61]	**,64 [,60-,68]**	**,50 [,44-,55]**	,65 [,59-,71]	**,57 [,51-,60]**	**,14 [,06-,20]**	**,60 [,55-,66]**	,68 [,63-,73]
		AB	,67 [,40-,92]	,42 [,15-,67]	,62 [,58-,68]	,46 [,38-,55]	,64 [,49-,73]	,53 [,46-,59]	,08 [-,02-,18]	,57 [,51-,65]	,66 [,60-,73]
		MLP	,65 [,50-,77]	,46 [,34-,61]	,63 [,60-,69]	,48 [,43-,54]	,64 [,56-,70]	,55 [,52-,59]	,11 [,04-,19]	,58 [,53-,64]	**,70 [,64-,74]**
		XGB	,62 [,48-,75]	**,49 [,32-,66]**	,63 [,58-,70]	,47 [,41-,53]	,62 [,54-,68]	,55 [,50-,60]	,10 [,07-,11]	,58 [,52-,65]	,67 [,61-,73]

APÊNDICE E — TESTES DE SIGNIFICÂNCIA ESTATÍSTICA

Tabela E.1 – Resultados dos testes de significância estatística

Abordagens	Algoritmo	*Sensibilidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Especificidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Precisão (VPP)* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior
	LR	1.47E-13	sim	sim	4.49E-67	sim		3.66E-66	sim	
	AB	6.83E-02			8.10E-10	sim		1.52E-25	sim	
$HCPA - mortalidade$:	J48	6.57E-16	sim	sim	3.87E-09	sim		5.13E-02		sim
múltiplas variáveis vs idade	MLP	7.39E-01		sim	1.53E-09	sim		1.44E-21	sim	
	RF	2.09E-01		sim	3.16E-54	sim		8.37E-68	sim	
	XGB	9.38E-08	sim		7.27E-18	sim		3.48E-58	sim	
	LR	2.60E-01			2.26E-66	sim		3.90E-18	sim	
	AB	3.60E-01		sim	1.41E-06	sim		6.43E-08	sim	
$HMV - mortalidade$:	J48	8.33E-18	sim	sim	4.14E-18	sim		8.27E-36	sim	
múltiplas variáveis vs idade	MLP	8.73E-18	sim	sim	4.02E-18	sim		1.39E-45	sim	
	RF	1.77E-38	sim	sim	4.76E-71	sim		3.90E-18	sim	
	XGB	2.11E-11	sim	sim	2.03E-54	sim		3.89E-65	sim	
	LR	3.31E-03	sim	sim	1.43E-23	sim		5.13E-21	sim	
	AB	2.14E-01			3.65E-01			2.90E-01		
$HMV_{CTI} - mortalidade$:	J48	9.15E-13	sim	sim	1.08E-02	sim		1.32E-01		sim
múltiplas variáveis vs idade	MLP	4.16E-14	sim	sim	2.12E-11	sim		5.01E-05	sim	
	RF	2.71E-03	sim		1.16E-32	sim		3.47E-42	sim	
	XGB	1.86E-05	sim	sim	4.04E-23	sim		6.57E-25	sim	

Tabela E.2 – (**Continuação**) Resultados dos testes de significância estatística

Abordagens	Algoritmo	*Sensibilidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Especificidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Precisão (VPP)* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior
	LR	4.14E-04	sim		1.02E-78	sim		5.81E-79	sim	
	AB	1.76E-02	sim	sim	5.18E-17	sim		1.33E-41	sim	
$HMV - CTI$: múltiplas	J48	2.24E-09	sim	sim	6.87E-24	sim		7.83E-34	sim	
variáveis vs idade	MLP	5.06E-02		sim	1.44E-16	sim		3.90E-18	sim	
	RF	5.22E-04	sim	sim	1.70E-75	sim		2.46E-81	sim	
	XGB	4.07E-05	sim		1.39E-57	sim		1.38E-77	sim	
	LR	8.46E-18	sim		3.90E-18	sim		1.97E-72	sim	
	AB	7.08E-12	sim		8.99E-01			5.52E-08	sim	
$HMV - CTI$: múltiplas	J48	2.73E-07	sim		3.81E-02	sim	sim	4.53E-01		sim
variáveis vs saturação O2	MLP	5.08E-09	sim		9.64E-01			1.72E-07	sim	
	RF	1.05E-08	sim		1.08E-17	sim		6.42E-55	sim	
	XGB	6.38E-28	sim		4.71E-10	sim		1.26E-40	sim	
	LR	7.06E-51	sim	sim	1.98E-56	sim		3.90E-18	sim	
	AB	4.00E-06	sim	sim	1.85E-08	sim		4.70E-08	sim	
$HCPA - vmi$: múltiplas	J48	3.90E-18	sim	sim	9.73E-08	sim		8.62E-03	sim	sim
variáveis vs saturação O2	MLP	3.90E-18	sim	sim	3.90E-18	sim		8.14E-32	sim	
	RF	4.91E-15	sim		2.57E-05	sim	sim	1.56E-03	sim	
	XGB	4.02E-11	sim		3.59E-05	sim	sim	8.69E-01		

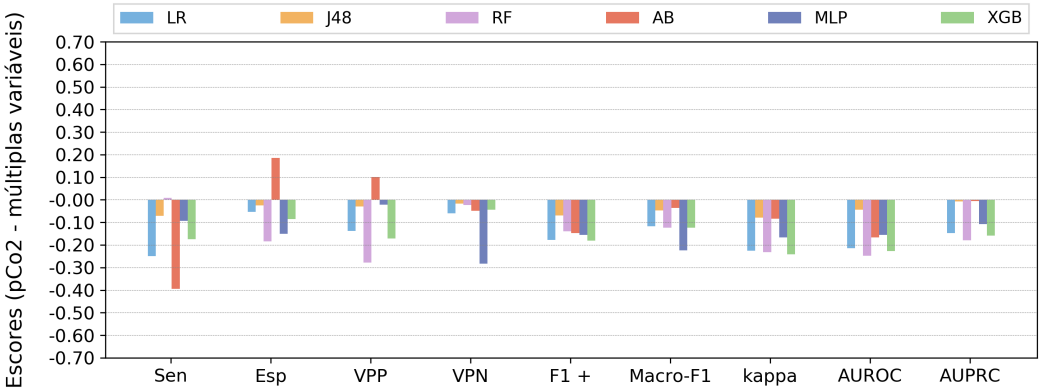
Tabela E.3 – (**Continuação**) Resultados dos testes de significância estatística

Abordagens	Algoritmo	*Sensibilidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Especificidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Precisão (VPP)* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior
	LR	3.90E-18	sim		9.00E-18	sim		3.90E-18	sim	
	AB	7.12E-18	sim		9.24E-10	sim	sim	1.21E-01		
$HMV - vmi$: múltiplas	J48	6.17E-01			7.96E-06	sim		1.26E-07	sim	
variáveis vs saturação O2	MLP	2.47E-04	sim		1.72E-01			2.26E-12	sim	
	RF	3.03E-10	sim	sim	3.90E-18	sim		3.90E-18	sim	
	XGB	4.02E-17	sim		5.31E-16	sim		4.27E-18	sim	
	LR	3.90E-18	sim	sim	9.00E-18	sim		3.90E-18	sim	
	AB	7.12E-18	sim	sim	9.24E-10	sim		1.21E-01		
$HMV_{CTI} - vmi$: múltiplas	J48	6.17E-01		sim	7.96E-06	sim		1.26E-07	sim	
variáveis vs saturação O2	MLP	2.47E-04	sim	sim	1.72E-01			2.26E-12	sim	
	RF	3.03E-10	sim	sim	3.90E-18	sim		3.90E-18	sim	
	XGB	4.02E-17	sim	sim	5.31E-16	sim		4.27E-18	sim	
	LR	3.90E-18	sim	sim	3.90E-18	sim		3.90E-18	sim	
	AB	4.11E-13	sim	sim	1.92E-15	sim		1.92E-15	sim	
$HCPA - vmi$: múltiplas	J48	3.90E-18	sim	sim	1.18E-06	sim		4.23E-03	sim	sim
variáveis vs PCO2	MLP	1.85E-16	sim	sim	9.04E-18	sim		1.01E-16	sim	
	RF	1.37E-17	sim		9.74E-09	sim	sim	1.21E-07	sim	sim
	XGB	8.09E-11	sim		6.81E-08	sim	sim	4.09E-03	sim	sim

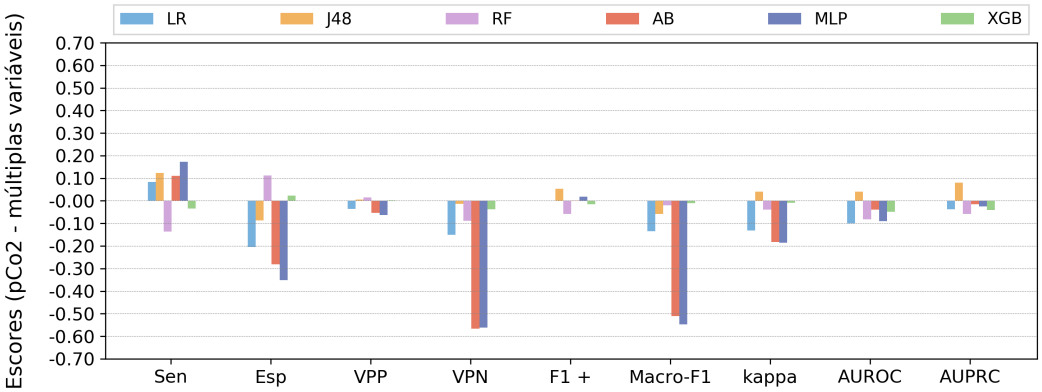
Tabela E.4 – (**Continuação**) Resultados dos testes de significância estatística

Abordagens	Algoritmo	*Sensibilidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Especificidade* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior	*Precisão (VPP)* P-value	Rejeição de H0 ($\alpha = 0.05$)	Abordagem alternativa superior
	LR	8.84E-68	sim		1.50E-17	sim		3.90E-18	sim	
	AB	3.12E-64	sim		9.32E-18	sim	sim	2.87E-10	sim	sim
$HMVvmi$: múltiplas	J48	1.20E-08	sim		1.14E-01			2.42E-05	sim	
variáveis vs PCO2	MLP	2.57E-01			8.82E-01			1.16E-01		
	RF	2.85E-01		sim	3.90E-18	sim		3.90E-18	sim	
	XGB	3.05E-36	sim		4.18E-20	sim		3.90E-18	sim	
	LR	3.98E-17	sim	sim	5.11E-18	sim		4.67E-18	sim	
	AB	1.42E-14	sim	sim	3.18E-16	sim		3.69E-15	sim	
$HMV_{CTI} - vmi$: múltiplas	J48	1.65E-14	sim	sim	2.07E-15	sim		3.05E-12	sim	
variáveis vs PCO2	MLP	5.55E-10	sim	sim	3.67E-13	sim		2.70E-08	sim	
	RF	1.11E-01		sim	2.72E-28	sim		3.18E-31	sim	
	XGB	4.36E-06	sim	sim	1.33E-16	sim		1.11E-17	sim	

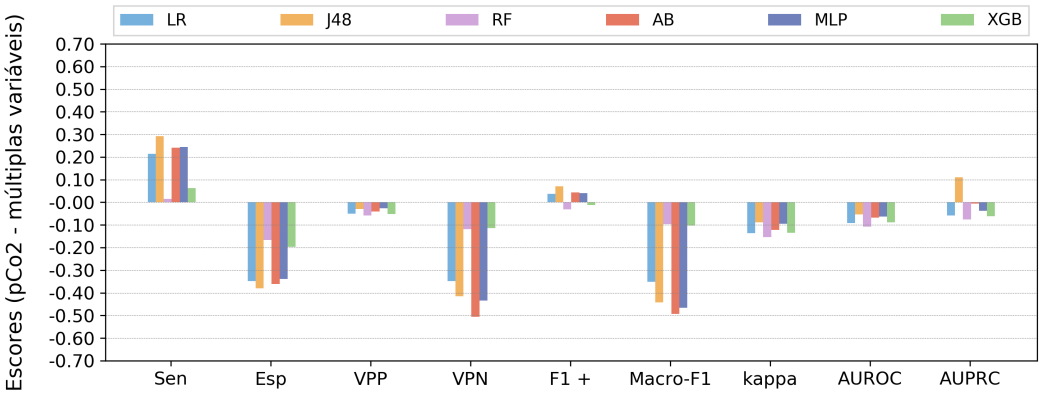
APÊNDICE F — COMPARAÇÃO ENTRE OS CLASSIFICADORES DE MÚLTIPLAS VARIÁVEIS E OS DE VARIÁVEL ÚNICA - VMI



(a) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no HMV



(b) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no $HCPA$



(c) Diferença entre os classificadores de múltiplas variáveis e os de variável única para a predição de VMI no HMV_{CTI}

Figura F.1 – Comparação entre os classificadores de múltiplas variáveis e os de variável única (pCO2) para a predição de VMI
Fonte: O Autor