UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Inteligência Artificial Explicável na Previsão de Reinternação Hospitalar: Uma Análise de Modelos de Aprendizado de Máquina

Matheus Mendes dos Santos

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Matheus Mendes dos Santos

Inteligência Artificial Explicável na Previsão de Reinternação Hospitalar: Uma Análise de Modelos de Aprendizado de Máquina

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Rodrigo Colnago Contreras

Versão original

São Carlos

2023

É possível elaborar a ficha catalográfica em LaTeX ou incluir a fornecida pela Biblioteca. Para tanto observe a programação contida nos arquivos USPSC-modelo.tex e fichacatalografica.tex e/ou gere o arquivo fichacatalografica.pdf.

A biblioteca da sua Unidade lhe fornecerá um arquivo PDF com a ficha catalográfica definitiva, que deverá ser salvo como fichacatalografica.pdf no diretório do seu projeto.

Matheus Mendes dos Santos

Explainable Artificial Intelligence in Hospital Readmission Prediction: An Analysis of Machine Learning Models

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Rodrigo Colnago Contreras

Original version

São Carlos

2023

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e Motivação

Nos últimos anos, houve um aumento dramático na quantidade de dados criados e compartilhados através da internet. Grande parte desse conteúdo está disponível publicamente e sem custo, representando um suprimento virtualmente ilimitado de dados de treinamento para várias tarefas de modelagem estatística e rotulagem. Essa "enxurrada de dados"apresenta desafios significativos, mas também oportunidades revolucionárias para o desenvolvimento de sistemas baseados em dados (SELTZER; ZHANG, 2009).

O crescente volume de dados gerados a partir dessas diversas fontes tem impulsionado uma popularização da inteligência artificial (IA), sobretudo com o uso de aprendizado de máquina, que, de forma sucinta, trata da criação de algoritmos que respondam e se adaptem automaticamente aos dados sem a necessidade de intervenção humana de forma contínua (FILHO, 2015). Aliado a ferramentas poderosas para o processamento massivo e paralelo de dados, esses avanços têm permitido tanto à indústria quanto à comunidade científica desenvolver modelos altamente eficazes e realizar análises mais sofisticadas. A adoção generalizada de técnicas de aprendizado de máquina tem se mostrado fundamental para desvendar o potencial dos dados disponíveis, resultando em descobertas inovadoras e aplicações transformadoras em diversas áreas, incluindo o setor de saúde (BATISTA; FILHO, 2019).

O número de estudos médicos de IA cresceu de forma exponencial no período de 2005 a 2019 (MESKO; GOROG, 2020), revelando o forte interesse da comunidade científica em buscar métodos cada vez mais eficazes para aprimorar o cuidado e a qualidade de vida dos pacientes. A aplicação de modelos preditivos de aprendizado de máquina possui um potencial significativo para auxiliar na tomada de decisões em diversas etapas do cuidado à saúde, principalmente no diagnóstico, intervenção e acompanhamento de problemas de saúde (OBERMEYER; LEE, 2017).

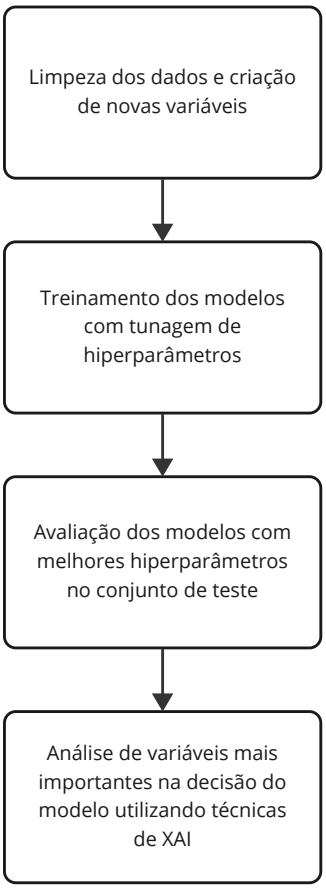
Na área da saúde, as decisões tomadas pelos sistemas de IA podem ter um impacto significativo na vida das pessoas. Se os usuários não conseguem compreender o processo de tomada de decisão, podem não confiar no sistema, chegando até a rejeitá-lo. Portanto, a capacidade de fornecer uma explicação sobre como e por que uma decisão específica foi feita tornou-se uma qualidade essencial em sistemas desse tipo (CONFALONIERI *et al.*, 2021). Além disso, a explicabilidade também pode auxiliar os desenvolvedores na identificação e correção de erros ou viés nos sistemas de IA, tornando-os mais justos e precisos.

Diante desse contexto, este trabalho se propõe a avaliar diversos modelos de clas-

sificação com o objetivo de prever a reinternação hospitalar de pacientes, utilizando um conjunto de dados de cuidados de saúde mental coletados pelo sistema de informação da Coordenação de Internações em Ribeirão Preto, Brasil, de julho de 2012 a dezembro de 2017. A relevância deste tipo de trabalho se dá, pois, por meio dessas previsões, medidas preventivas podem ser adotadas antecipadamente, evitando a necessidade de reinternação e, consequentemente, reduzindo os custos hospitalares, visto que o custo médio das internações é significativamente maior que o custo médio dos atendimentos ambulatoriais (CESCONETTO; LAPA; CALVO, 2008). Além da predição, o estudo também utilizará técnicas de inteligência artificial explicável (do inglês: *explainable artificial inteligence* - XAI) a fim de identificar quais variáveis estão mais associadas a casos de reinternação, fornecendo informações relevantes para o desenvolvimento de políticas e medidas preventivas mais eficazes.

A Figura 1 apresenta um fluxo geral deste trabalho:

Figura 1 – Fluxo geral do trabalho



Fonte: Autor (2023)

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo, serão apresentadas as fundamentações teóricas das técnicas de aprendizado de máquina utilizadas neste trabalho, juntamente com um levantamento de estudos correlatos na área da saúde que empregaram técnicas semelhantes.

2.1 Aprendizado de Máquina

A crescente complexidade nos desafios computacionais e o volume massivo de dados gerados por diversas fontes impulsionaram a necessidade de ferramentas computacionais mais avançadas. Neste contexto, o aprendizado de máquina define-se como um campo de estudo da Inteligência Artificial que dá aos computadores a habilidade de aprender sem serem explicitamente programados. Para tal, no aprendizado de máquina, os computadores são treinados para aprender com dados passados, utilizando técnicas que os possibilitem a derivar uma função ou hipótese capaz de solucionar problemas a partir de observações específicas, oferecendo soluções baseadas em dados históricos (FACELI *et al.*, 2011).

Os sistemas de aprendizado de máquina podem seguir vários paradigmas de aprendizado. Existem sistemas supervisionados, não supervisionados, semi-supervisionados, auto-supervisionados e de reforço. O aprendizado supervisionado é usado quando o modelo pode ser treinado com exemplos rotulados, enquanto o aprendizado não supervisionado é usado quando não há rótulos disponíveis. O aprendizado semi-supervisionado é uma combinação dos dois, enquanto o auto-supervisionado é usado quando o modelo pode ser treinado com exemplos gerados automaticamente. O aprendizado por reforço é usado quando o modelo pode aprender a tomar decisões com base em recompensas e punições (GÉRON, 2022).

Dentro do paradigma de aprendizado supervisionado, os algoritmos podem ser usados em tarefas de classificação ou regressão. A classificação é usada para prever classes, enquanto a regressão é usada para prever valores. Por exemplo, a classificação pode ser usada para prever se um e-mail é spam ou não, enquanto a regressão pode ser usada para prever o preço de uma casa com base em suas características (GÉRON, 2022).

A seguir será apresentada uma breve introdução aos modelos de classificação utilizados no desenvolvimento deste projeto.

2.1.1 K-Vizinhos Mais Próximos

O algoritmo K-Vizinhos Mais Próximos (ou KNN, do inglês *K-Nearest Neighbors*) é um método de classificação que opera com base na proximidade dos vizinhos mais próximos de um ponto de dados. Quando um novo ponto de dados precisa ser classificado, o algoritmo encontra os k pontos mais próximos a ele a partir do conjunto de dados

de treinamento. Em seguida, ele classifica o novo ponto de dados com base na classe mais frequente entre esses k vizinhos mais próximos. O algoritmo KNN assume que todas as observações correspondem a pontos em um espaço n-dimensional e não requer um processo de aprendizado. Ele apenas prevê a categoria do novo ponto de dados com base nas categorias dos pontos conhecidos no momento da classificação (WANG, 2019).

Os vizinhos mais próximos de uma instância são mensurados por métricas de distância, tais como a distância Euclidiana ou de Manhattan. A título de exemplo, será apresentada a distância Euclidiana, definida pela Equação 2.1:

$$d(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x'_i)^2} \tag{2.1}$$

2.1.2 Regressão Logística

A regressão logística é um modelo estatístico usado para modelar a probabilidade de um evento ocorrer em termos de variáveis independentes (KLEINBAUM; KLEIN, 2010). O modelo logístico é baseado na função logística, definida pela Equação 2.2. A Figura 2 apresenta um exemplo da curva logística.

$$f(z) = \frac{1}{1 + e^z} \tag{2.2}$$

De forma simplificada, o modelo logístico pode ser definido da seguinte maneira:

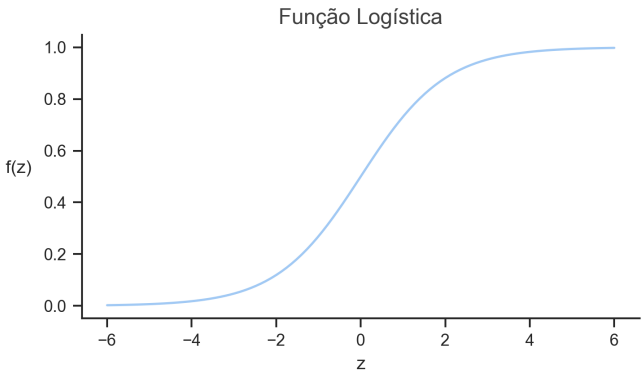
$$\hat{y} = f(Z) = P(y = 1|Z) = \frac{1}{1 + e^{-Z}} \tag{2.3}$$

onde $P(y = 1|Z)$ é a probabilidade de ocorrência da classe positiva dado Z. Sendo Z:

$$Z = ln(\frac{p}{1 - p}) = \alpha + \sum_{j=1}^{k} \beta_k X_{ik} \tag{2.4}$$

onde p representa a probabilidade de ocorrência do evento de interesse, enquanto X denota o conjunto de variáveis preditoras. Os parâmetros do modelo, α e β, são estimados por meio da técnica de máxima verossimilhança (FÁVERO *et al.*, 2009). Essa abordagem visa encontrar uma combinação de coeficientes que maximize a probabilidade de ocorrência do evento de interesse

Figura 2 – Exemplo de função logística



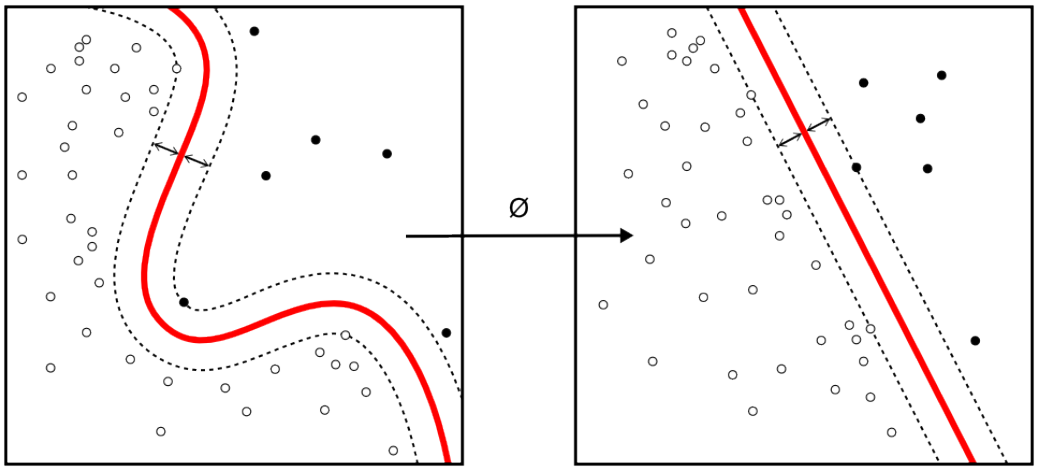
Fonte: Autor (2023)

2.1.3 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (ou SVM, do inglês *Support Vector Machines*) são um método de aprendizado supervisionado que busca encontrar o hiperplano de separação que maximiza a margem entre as classes. Elas são usadas em problemas de classificação para encontrar a melhor fronteira de decisão entre as classes, permitindo sobreposição entre elas, mas minimizando essa sobreposição. As SVMs são particularmente úteis em problemas de classificação com muitas características, onde outras técnicas podem sofrer de sobreajuste, o que as torna robustas em relação a ruídos e *outliers* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

As SVMs podem lidar com conjuntos de dados não lineares por meio do uso de *kernels*, que mapeiam os dados para um espaço de maior dimensão onde eles podem ser separados linearmente. A Figura 3 mostra um exemplo de um espaço de características que foi transformado.

Figura 3 – Exemplo de transformação utilizando *kernel*: Convertendo um problema de classificação não linear em um problema de classificação linear em uma dimensão superior.



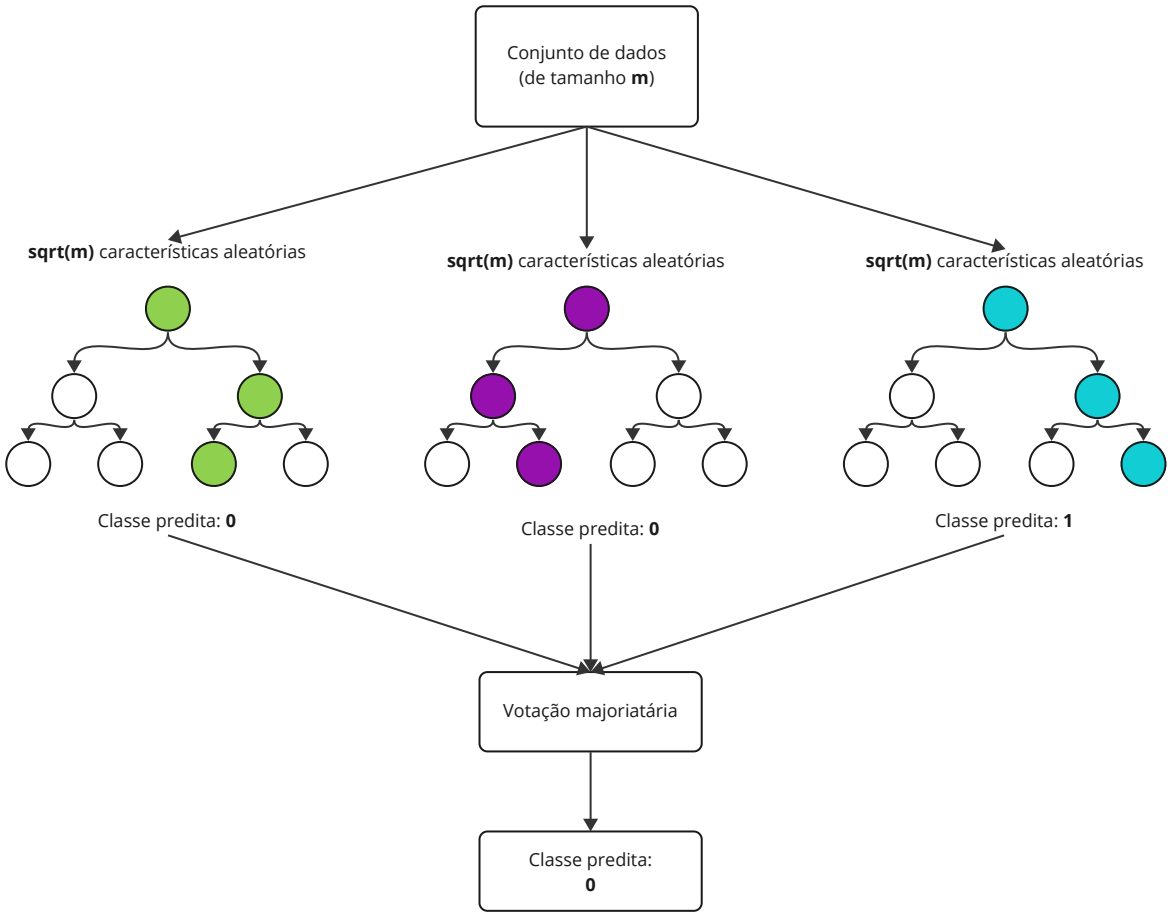
Fonte: (Wikipedia contributors, 2023)

Embora as SVMs sejam amplamente utilizadas e tenham várias vantagens, também apresentam algumas desvantagens. Uma das principais desvantagens das SVMs é a necessidade de ajuste cuidadoso dos parâmetros do modelo, como o parâmetro de regularização e a escolha do *kernel*, o que pode ser uma tarefa desafiadora e exigir conhecimento especializado. Além disso, as SVMs podem ser computacionalmente intensivas, especialmente em conjuntos de dados muito grandes, o que pode tornar seu treinamento demorado. Outra limitação das SVMs é a dificuldade em lidar com conjuntos de dados com muitas classes, uma vez que a abordagem de um-contra-um para problemas multiclasse pode levar a um grande número de classificadores binários (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.1.4 Florestas Aleatórias e *Gradient Boosted Trees*

As Florestas Aleatórias (do inglês *Random Forests*) e *Gradient Boosted Trees* são exemplos de comitês (*ensembles*) de algoritmos de aprendizado de máquina. Resumidamente, o método consiste na construção de múltiplas árvores de decisão em subespaços aleatórios do espaço de características, utilizando amostras aleatórias dos dados de treinamento e subconjuntos aleatórios das características disponíveis. Cada árvore de decisão é capaz de generalizar sua classificação de forma única, e a combinação das previsões de todas as árvores resulta em uma melhoria monotônica na precisão da classificação (HO, 1995). Estudos recentes têm indicado que a aplicação desses algoritmos tem apresentado desempenho superior na previsão em comparação com outros métodos de aprendizado de máquina (GÉRON, 2022; RASCHKA; MIRJALILI, 2017). A Figura 4 ilustra a construção de três árvores de decisão para um problema de classificação, construídas pelo algoritmo *Random Forest*.

Figura 4 – Ilustração do algoritmo *Random Forest*.



Fonte: Adaptado de Batista and Filho (2019)

Com o objetivo de melhorar a precisão das previsões das árvores de decisão, o algoritmo *Gradient Boosted Trees* (FRIEDMAN, 2000) utiliza a estratégia de *boosting*. Em termos mais detalhados, o processo de *boosting* geralmente começa com um modelo base simples, como uma árvore de decisão rasa. Este modelo é treinado no conjunto de dados e usado para fazer previsões. Em seguida, os exemplos que foram classificados incorretamente ou para os quais as previsões tiveram grandes erros são ponderados mais fortemente no próximo modelo a ser treinado. Este próximo modelo é treinado para se concentrar nos exemplos que foram mal classificados pelo modelo anterior, e o processo é repetido (FRIEDMAN, 2000).

Ainda segundo Friedman (2000), a técnica de *Gradient Boosting* proposta pelo autor permitiu avanços importantes na capacidade de lidar com problemas complexos de previsão e classificação. Essa abordagem influenciou diretamente o desenvolvimento de algoritmos populares, como *XGBoost* e *LightGBM*, que são amplamente utilizados em competições de ciência de dados e em aplicações do mundo real devido à sua eficácia e desempenho.

2.2 Inteligência Artificial Explicável

A Inteligência Artificial Explicável refere-se à capacidade de compreender e explicar as decisões e previsões dos modelos de inteligência artificial. A explicabilidade é crucial para estabelecer a confiança do usuário, fornecer insights sobre como melhorar o modelo e compreender o processo que está sendo modelado. A capacidade de interpretar corretamente a saída de um modelo de previsão é extremamente importante, especialmente em contextos nos quais modelos complexos e de alto desempenho são utilizados (LUNDBERG; LEE, 2017). Nesta seção, será fornecida uma breve exposição do método de XAI aplicado neste trabalho, o *SHapley Additive exPlanations* (SHAP).

2.2.1 *SHapley Additive exPlanations* (SHAP)

O SHAP (LUNDBERG; LEE, 2017) unifica seis métodos existentes de interpretação de predições: LIME, DeepLIFT, Layer-wise Relevance Propagation (LRP), Shapley sampling values, SmoothGrad e Integrated Gradients.

O LIME (Local Interpretable Model-Agnostic Explanations) (RIBEIRO; SINGH; GUESTRIN, 2016b) é um método que explica as previsões de um modelo complexo construindo um modelo localmente interpretável em torno da previsão. O DeepLIFT (SHRIKUMAR *et al.*, 2016) compara a ativação de cada variável na previsão atual com a ativação média da variável em todas as previsões. O LRP (BACH *et al.*, 2015) atribui a importância de cada variável para uma previsão propagando a relevância da saída do modelo de volta para as entradas. O Shapley sampling values (ŠTRUMBELJ; KONONENKO, 2014) amostra aleatoriamente subconjuntos de variáveis e calcula a contribuição média de cada variável para a previsão. O SmoothGrad (SMILKOV *et al.*, 2017) suaviza as entradas do modelo para reduzir o ruído e, em seguida, calcula a importância de cada variável para uma previsão. O Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) integra a contribuição de cada variável para uma previsão ao longo de um caminho suave entre a entrada atual e uma entrada de referência.

O SHAP unifica esses métodos por meio da introdução do conceito de "modelo de explicação", que representa uma aproximação interpretável do modelo original. Além disso, fornece resultados teóricos que asseguram a existência de uma solução única na classe de métodos aditivos, atendendo a um conjunto de propriedades desejáveis, incluindo consistência, localidade e completude. O SHAP propõe uma nova medida de importância de variáveis chamada *SHAP value*, a única medida aditiva que satisfaz essas propriedades. Esses valores, conforme demonstrado pelo SHAP, estão alinhados com a intuição humana e são eficazes na discriminação entre classes de saída do modelo (LUNDBERG; LEE, 2017).

2.3 Trabalhos de classificação na área médica

A utilização de algoritmos de aprendizado de máquina para análises preditivas é uma tarefa que envolve uma série de etapas que vão desde a preparação dos dados, até o lançamento e monitoramento da solução (GÉRON, 2022). O trabalho de Santos *et al.* (2019) se propõe a exemplificar as etapas envolvidas na utilização desses algoritmos para análises preditivas em saúde, com um exemplo de aplicação para predizer óbito em idosos de São Paulo. A metodologia empregada envolveu a utilização de diferentes algoritmos de aprendizado de máquina para treinar modelos preditivos a partir de dados do estudo SABE, e a avaliação desses modelos em termos de acurácia, sensibilidade e especificidade. Os resultados mostraram que os modelos tiveram desempenho razoável na predição de óbito em idosos, com destaque para o algoritmo Random Forest. No entanto, os autores ressaltam que a amostra utilizada foi relativamente pequena e que a performance dos modelos pode ser melhorada com mais dados e refinamento dos algoritmos.

O trabalho de Loreto, Lisboa and Moreira (2020) aborda a predição de reinternações em UTI usando algoritmos de classificação. A proposta é investigar se as características basais e informações coletadas no momento da admissão do paciente podem possibilitar previsões precisas de reinternação na UTI. A metodologia empregada envolveu a análise de um conjunto de dados de 11.805 pacientes adultos de três UTIs em um hospital universitário brasileiro, com a exclusão de pacientes falecidos durante a primeira admissão na UTI. Foram testados oito algoritmos de classificação e avaliados com base em seis métricas. Os resultados mostraram que as características do paciente registradas na admissão na UTI foram capazes de prever o risco de reinternação, superando modelos anteriores que utilizavam apenas dados disponíveis na alta da UTI.

Wang *et al.* (2021) apresenta um estudo sobre a estimativa da probabilidade de reinternação à UTI para pacientes transferidos da UTI para a enfermaria geral, abordando os desafios de dados desbalanceados e esparsos. A pesquisa visa melhorar a gestão médica e reduzir as taxas de reinternação à UTI, levando a melhores resultados para os pacientes e redução de custos de saúde. A metodologia empregada inclui a análise de valores ausentes, o teste de razão de verossimilhança e a utilização de um modelo de floresta aleatória com decaimento de peso. Os resultados mostram que o modelo proposto supera outros métodos tradicionais em sete indicadores de desempenho diferentes.

O artigo de Stiglic *et al.* (2014) propõe uma abordagem para a classificação de reinternações pediátricas usando modelos de diferentes hospitais, permitindo alta performance e compreensibilidade dos resultados. A metodologia empregada é baseada em *deep learning* e *stacked generalization*, e foi avaliada usando dados de 54 hospitais na Califórnia para demonstrar as possibilidades de implantação em larga escala. Os resultados mostraram melhorias significativas no desempenho de classificação e interpretabilidade dos resultados.

O trabalho de Yin *et al.* (2023) tem como objetivo avaliar se a classificação de feridas do *Centers for Disease Control and Prevention* (CDC) pode ser usada como um indicador de risco para reinternação hospitalar em 30 dias após a cirurgia. A metodologia empregada foi uma revisão retrospectiva de pacientes submetidos a cirurgias eletivas em um hospital universitário. Os resultados mostraram que a classificação de feridas do CDC foi um preditor significativo de reinternação hospitalar em 30 dias após a cirurgia. Os autores concluem que a classificação de feridas do CDC pode ser uma ferramenta útil para identificar pacientes com maior risco de reinternação hospitalar após a cirurgia.

Fernandes *et al.* (2021) apresenta uma abordagem multipropósito de aprendizado de máquina para prever o prognóstico negativo da COVID-19 em São Paulo, Brasil. O trabalho envolveu o treinamento de cinco algoritmos de aprendizado de máquina com dados estruturados de pacientes com COVID-19, utilizando técnicas de validação cruzada e otimização bayesiana. O objetivo era prever o risco de desenvolver condições críticas em pacientes com COVID-19. Os resultados mostraram que a abordagem proposta teve um desempenho satisfatório na previsão do prognóstico negativo da COVID-19, o que pode ajudar na tomada de decisões clínicas e na alocação efetiva de recursos de saúde.

O trabalho de Asif *et al.* (2018) propõe uma metodologia baseada em aprendizado de máquina para identificar genes relacionados a doenças complexas. O modelo de aprendizado de máquina proposto utiliza uma matriz de similaridade funcional entre genes para prever doenças genéticas complexas. Essa matriz é construída a partir de medidas de similaridade entre genes, como perfis de expressão gênica, redes de interação proteína-proteína e Gene Ontology. Em seguida, um classificador de aprendizado de máquina é treinado com essa matriz para identificar genes relevantes para a doença em questão. O modelo proposto foi avaliado em relação à sua capacidade de identificar genes relacionados ao Transtorno do Espectro Autista e mostrou melhorias significativas em relação a outros métodos de identificação de genes. Foi desenvolvido um fluxo de trabalho automatizado para tornar a metodologia acessível a pesquisadores sem conhecimento extenso em programação e aprendizado de máquina.

Rosa *et al.* (2023) apresenta um estudo que utiliza técnicas de aprendizado de máquina para classificar fatores que influenciam a ocorrência de dermatites ocupacionais. A metodologia empregada envolveu a avaliação de diferentes algoritmos de aprendizado de máquina, a comparação dos resultados obtidos por cada técnica e a identificação dos fatores mais influentes na ocorrência da lesão ocupacional. Os resultados indicaram que todas as técnicas apresentaram acuracidade entre 55% e 69,4%, sensitividade entre 49,1% e 80,7% e especificidade entre 50% e 66,7%. Os fatores mais influentes identificados foram a exposição a produtos químicos, o uso de equipamentos de proteção individual inadequados e a falta de treinamento adequado.

2.4 Técnicas de explicação aplicadas a trabalhos de classificação na área médica

A explicabilidade já foi identificada como um fator chave para a adoção de sistemas de IA numa vasta gama de contextos (DOSHI-VELEZ; KIM, 2017; LIPTON, 2017; RIBEIRO; SINGH; GUESTRIN, 2016a). No estudo conduzido por Abououf *et al.* (2023) é apresentado um sistema de monitoramento de saúde que utiliza inteligência artificial para detectar e classificar eventos e anomalias médicas. A metodologia empregada envolve a utilização de um modelo de detecção de anomalias e eventos, um modelo de classificação e uma técnica de explicabilidade chamada *KernelSHAP*. O sistema foi avaliado em um conjunto de dados de pacientes e obteve resultados promissores na detecção e classificação de eventos e anomalias médicas. Além disso, a técnica de explicabilidade permitiu que os médicos entendessem melhor as decisões tomadas pelo sistema de inteligência artificial.

Com o uso de técnicas de aprendizado profundo e transferência de aprendizado para detectar a COVID-19 a partir de radiografias de tórax, o estudo de Brunese *et al.* (2020) utiliza as camadas de ativação da rede, ou seja, as áreas da radiografia de tórax que o modelo considerou para gerar a predição, para fornecer explicabilidade sobre a predição. Ainda segundo os autores, isso pode representar uma sugestão para o radiologista localizar imediatamente as áreas da radiografia de tórax que podem ser de interesse. Os resultados mostraram que o modelo proposto alcançou uma acurácia de 98,08% na detecção de COVID-19 em radiografias de tórax. O estudo também sugere que essa tecnologia pode ser usada como uma ferramenta de triagem para ajudar no diagnóstico da COVID-19 em configurações clínicas do mundo real.

Em (KIM; KIM, 2022) foi desenvolvida uma técnica de previsão de mortalidade relacionada ao calor em uma unidade espacial detalhada dentro de uma cidade, utilizando um modelo baseado em floresta aleatória e a técnica de explicação SHAP. A metodologia empregada incluiu a coleta de dados meteorológicos, demográficos e socioeconômicos, pré-processamento de dados, divisão em conjuntos de treinamento e teste, construção do modelo, otimização de hiperparâmetros, avaliação de desempenho e interpretação dos resultados. Os resultados mostraram que o modelo proposto é capaz de prever com precisão a mortalidade relacionada ao calor em uma área específica da cidade de Daegu, na Coreia do Sul. A técnica SHAP permitiu uma interpretação global e local dos resultados, identificando as variáveis mais importantes para a previsão.

Também utilizando modelos SHAP, o trabalho de GhoshRoy, Alvi and Santosh (2023) utiliza técnicas para prever a fertilidade masculina, com o intuito de melhorar a transparência, responsabilidade e explicabilidade de sete modelos de IA padrão da indústria. A metodologia empregada consistiu em coletar dados de pacientes com problemas de fertilidade masculina e aplicar os modelos de IA para prever a fertilidade. Em seguida, foram utilizadas técnicas de amostragem e validação cruzada para avaliar a robustez e estabilidade de cada modelo. Por fim, a técnica de XAI foi utilizada para explicar o desem-

penho de cada modelo e os resultados mostraram que o SHAP foi capaz de interpretar as previsões dos sistemas de IA e fornecer explicações claras e compreensíveis para os usuários.

3 DESENVOLVIMENTO

Neste capítulo, será apresentada uma exposição ordenada e detalhada do desenvolvimento do presente trabalho. O início deste capítulo se dá com uma explanação da metodologia empregada, delineando os passos e abordagens utilizados para alcançar os objetivos propostos.

3.1 Metodologia

3.1.1 Definição de domínio do problema

O problema abordado neste estudo refere-se à previsão de reinternação hospitalar de pacientes que receberam cuidados de saúde mental no sistema de informação da Coordenação de Internações em Ribeirão Preto, Brasil, no período de julho de 2012 a dezembro de 2017. Este domínio envolve a aplicação de modelos de classificação de aprendizado de máquina em um conjunto de dados específico, com o propósito de antecipar eventos de reinternação hospitalar. Além disso, o estudo incorpora técnicas de inteligência artificial explicável, notadamente o SHapley Additive exPlanations (SHAP), para identificar e quantificar as variáveis que apresentam maior influência nos casos de reinternação, contribuindo assim para uma compreensão mais profunda dos fatores associados a esses eventos. A análise deste domínio visa aprimorar a capacidade de tomada de decisões e políticas de saúde, com ênfase na prevenção eficaz e na otimização dos recursos hospitalares.

3.1.2 Conjunto de dados

O conjunto de dados empregado neste estudo já foi explorado em outras pesquisas (BARROS *et al.*, 2016; MIYOSHI *et al.*, 2018; KHARRAT *et al.*, 2020). Ele abrange informações de 8.755 pacientes, com uma média de idade de 37,6 anos. As características do conjunto de dados englobam aspectos sociodemográficos, dados sobre internações hospitalares, diagnósticos, utilização de serviços médicos, informações de alta hospitalar e registros temporais, como datas de admissão, alta e óbito. A Tabela 1 descreve as variáveis utilizadas neste estudo.

Tabela 1 – Variáveis utilizadas do conjunto de dados

Nome da variável	**Tipo da variável**
Arranjo docimiliar	Categórica
AVC	Booleana
Convulsão	Booleana
Dia da semana na 1ª internação	Numérica (inteira)
Diabetes	Booleana
Diagnóstico (CID10)	Categórica
Doença infecto	Booleana
Estado civil	Categórica
Etnia	Categórica
HAS	Booleana
Idade na 1ª internação	Numérica (inteira)
Mês da 1ª internação	Numérica (inteira)
Problemas respiratórios	Booleana
Quantidade de problemas na 1ª internação	Numérico (inteira)
Sexo	Categórica
Situação profissão	Categórica
Tempo de internação (em horas)	Numérica (contínua)
Traumatismo	Booleana

Fonte: Autor (2023)

É importante ressaltar que essas não são todas as variáveis disponíveis no conjunto de dados. Mais detalhes sobre como a seleção de atributos foi feita estão disponíveis na subseção 3.1.4.

3.1.3 Ferramentas utilizadas

Todo o desenvolvimento deste trabalho foi feito utilizando a linguagem de programação Python. Essa linguagem não apenas oferece uma sintaxe clara e concisa, mas também proporciona um vasto ecossistema de bibliotecas especializadas que simplificam e aprimoram cada etapa de um projeto de ciência de dados.

As bibliotecas Pandas e Numpy foram utilizadas na limpeza e preparação dos dados para a seleção de modelos. A escolha dessas bibliotecas deve-se à eficiência e flexibilidade que oferecem em termos de manipulação de dados, permitindo operações rápidas e simplificadas. Pandas facilita a manipulação de estruturas de dados tabulares, enquanto

a Numpy fornece suporte para operações numéricas eficientes, ambos fundamentais em projetos de ciência de dados.

A biblioteca Scikit-Learn (BUITINCK *et al.*, 2013) foi utilizada para o pré-processamento, seleção e validação cruzada dos modelos. A escolha dessa biblioteca deve-se à sua abrangência na implementação de algoritmos de aprendizado de máquina, proporcionando uma interface consistente para a construção, avaliação e ajuste de modelos. Isso facilita a experimentação e comparação entre diferentes abordagens, agilizando o desenvolvimento do projeto.

Matplotlib e Seaborn foram as bibliotecas utilizadas na criação de visualizações em todas as etapas do projeto. A escolha dessas bibliotecas deve-se à capacidade de produzir gráficos informativos e visualmente atraentes. Isso não apenas facilita a compreensão rápida dos resultados, mas também fortalece a comunicação efetiva dos achados para diferentes públicos, desde especialistas em dados até partes interessadas não técnicas.

Por fim, a bilioteca Shap foi utilizada por sua capacidade de explicar as predições dos modelos de aprendizado de máquina. Isso não apenas torna as previsões mais interpretáveis, mas também fornece insights profundos sobre como cada variável contribui para as decisões do modelo, melhorando a confiabilidade e compreensão do sistema.

3.1.4 Limpeza e prepação dos dados

O processo de limpeza de dados foi executado em diversas etapas para garantir a qualidade e a relevância do conjunto de dados para as análises e modelagem subsequentes. Inicialmente, foram identificadas e eliminadas variáveis redundantes para o escopo específico da tarefa em pauta, que consistia na previsão de reinternação de pacientes. Além disso, foram eliminadas variáveis que poderiam potencialmente introduzir *data leakage* nos modelos a serem desenvolvidos posteriormente. Um exemplo ilustrativo desse cenário seria a presença de variáveis que forneciam informações sobre a segunda internação de um paciente. Utilizar tais informações representaria um vazamento de dados, pois indicaria ao modelo que o paciente já passou por uma reinternação, informação que não estaria disponível no momento da primeira internação. Este cuidado durante a limpeza dos dados foi essencial para preservar a integridade do processo de modelagem e garantir resultados mais precisos e confiáveis.

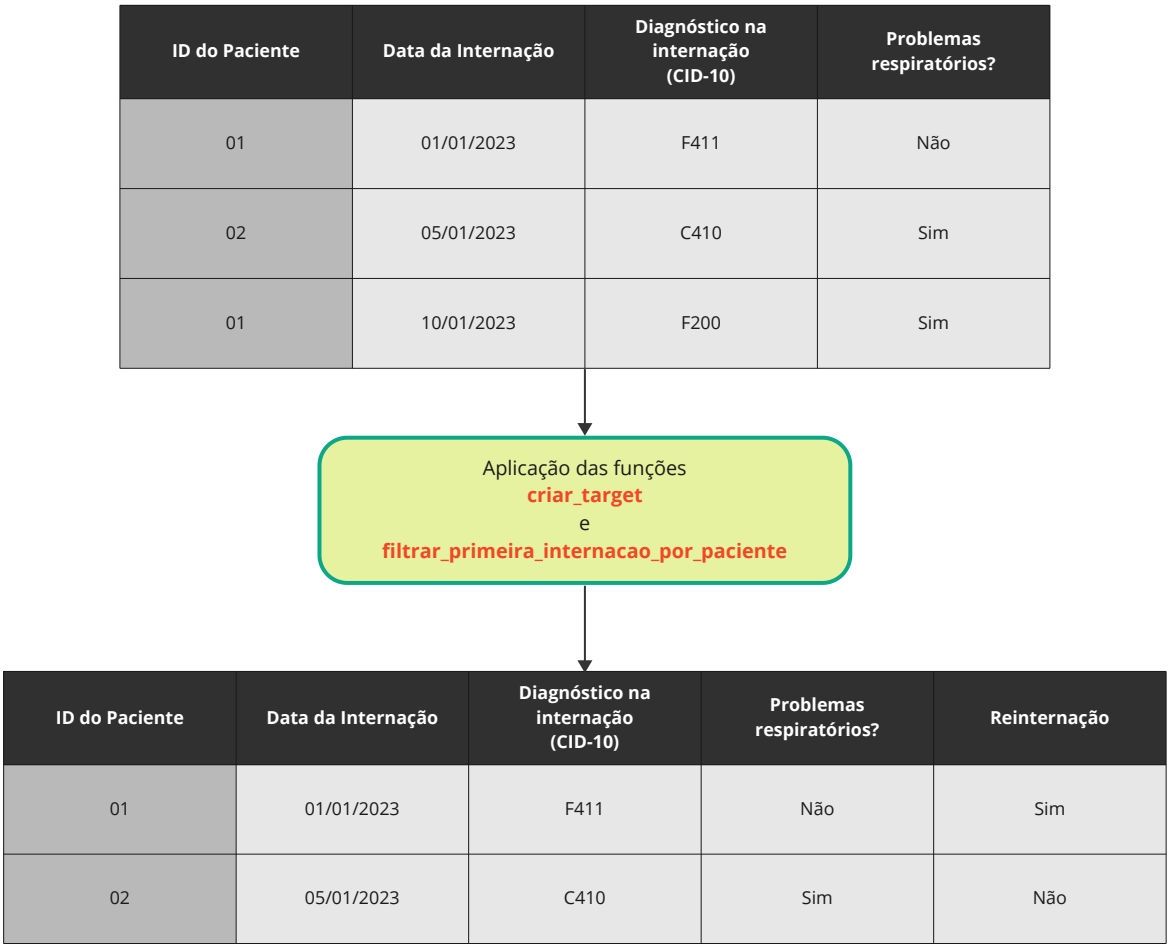
Dado que a natureza dos dados envolve preenchimentos manuais, tornou-se essencial padronizar e simplificar as categorias para garantir consistência e facilitar a interpretação. Variáveis como estado civil, etnia e arranjo domiciliar foram submetidas a procedimentos específicos para aprimorar a qualidade das informações. O arranjo domiciliar, por exemplo, representado por uma variedade de termos como “sozinho”, “com companheiro(a)”, “com filhos(as)”, entre outros, foi tratado por uma função que simplificou as categorias, mapeando-as para as opções mais abrangentes como “acompanhado” ou

"sozinho". Esse tipo de simplificação não apenas facilitou a análise, mas também reduziu a complexidade do conjunto de dados.

As variáves relacionadas a data (data de internação, data de alta e data de nascimento) foram tratadas e serviram de base para a criação de novas variáveis que buscam aprimorar a compreensão do comportamento dos pacientes e, consequentemente, melhorar as previsões dos modelos. Por exemplo, a variável "**mes_data_internacao**" indica o mês da internação, enquanto "**dia_semana_data_internacao**" representa o dia da semana. Esses atributos têm o potencial de capturar padrões sazonais ou comportamentais que podem influenciar as chances de reinternação. Além disso, a idade do paciente na data da internação foi calculada para investigar se ela tem algum impacto nas taxas de reinternação. Uma métrica de tempo, "**tempo_internacao_horas**", foi introduzida para avaliar a duração da internação em horas, proporcionando uma perspectiva mais granular do tempo de permanência. Essas novas variáveis foram concebidas com a intenção de enriquecer o conjunto de dados e investigar se fatores temporais específicos estão relacionados à probabilidade de reinternação.

Como o conjunto de dados trata de informações coletadas na internação, ele não tem por definição uma variável que pode ser usada como variável alvo para reinternação, portanto, ela precisou ser criada. Para estabelecer a variável-alvo "reinternacao", foi desenvolvida uma função que utiliza a contagem de ocorrências de cada paciente para determinar se houve reinternação, considerando como critério a presença de mais de uma internação para um mesmo paciente. Além disso, uma outra função foi aplicada para extrair as informações exclusivas da primeira internação de cada paciente. Essa estratégia não apenas evita o vazamento de dados, mas também contribui para a integridade da modelagem, garantindo que as previsões sejam baseadas unicamente nas informações disponíveis até o momento da primeira internação. A Figura 5 exemplifica o funcionamento dessas funções e quais dados são considerados no conjunto final.

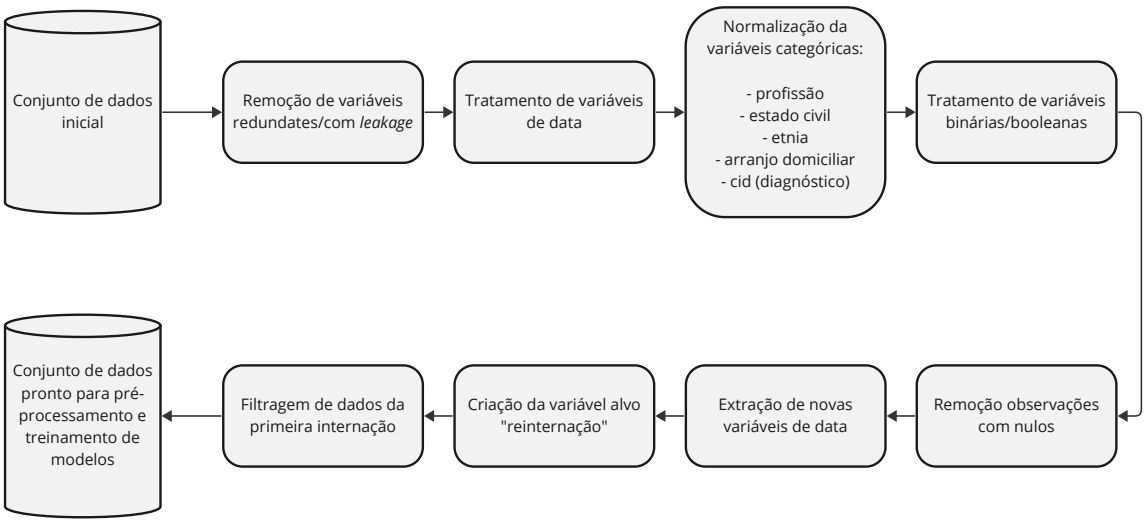
Figura 5 – Fluxograma ilustrando o processo de criação da variável-alvo “reinternação” e a filtragem para extrair informações da primeira internação de cada paciente. Na parte superior, o estado inicial do conjunto de dados com múltiplas internações. Na parte inferior, o conjunto final, otimizado para as próximas fases de pré-processamento e treinamento dos modelos, após a aplicação das funções.



Fonte: Autor (2023)

Em resumo, a fase de limpeza e preparação dos dados foi meticulosa, visando a qualidade e relevância necessárias para análises e modelagem. Redundâncias e potenciais fontes de vazamento foram eliminadas, garantindo a integridade do processo de modelagem. A padronização de categorias simplificou a interpretação e a manipulação cuidadosa de variáveis relacionadas a datas introduziu novas características, buscando capturar padrões temporais relevantes para a previsão de reinternação. A Figura 6 apresenta um fluxograma visualizando todas as transformações aplicadas ao conjunto inicial de dados, delineando as etapas que culminaram no conjunto final, pronto para as próximas fases de pré-processamento e treinamento dos modelos.

Figura 6 – Fluxograma representando as transformações aplicadas ao conjunto inicial de dados durante o processo de limpeza e preparação.



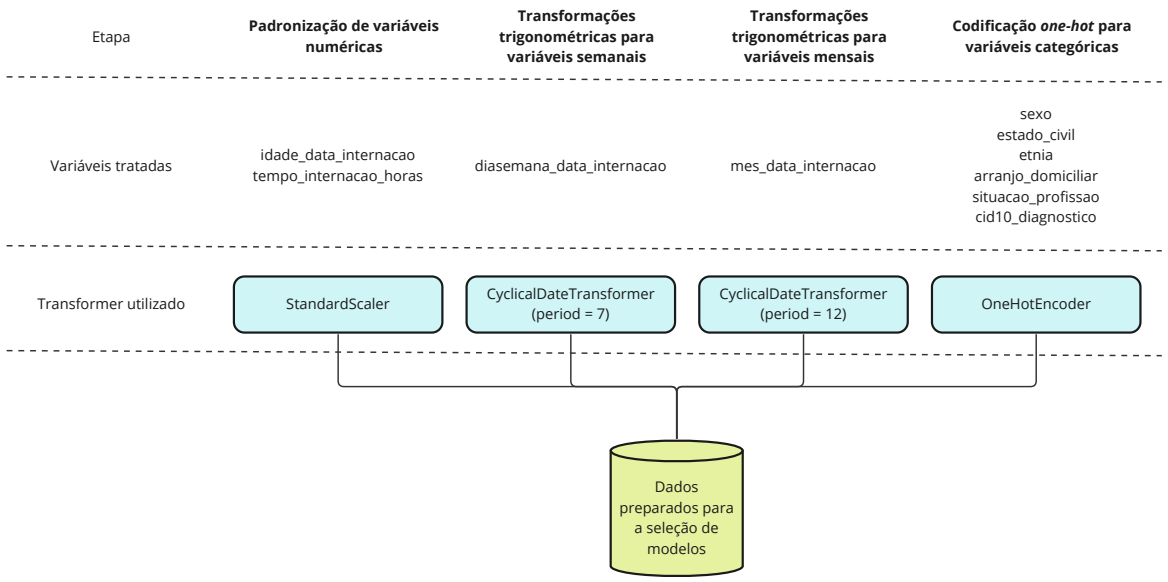
Fonte: Autor (2023)

3.1.5 Treinamento dos classificadores

A fase de treinamento e seleção dos classificadores tem início com a subdivisão do conjunto de dados em conjuntos distintos para treinamento e teste (o conjunto de teste foi empregado exclusivamente para validar os modelos já treinados, e nunca para a calibração destes modelos). Em seguida, utilizando as interfaces *BaseEstimator* e *TransformerMixin* do Scikit-Learn (BUITINCK *et al.*, 2013), foi definido um *Transformer* que foi especialmente projetado para lidar com variáveis de data. Sua funcionalidade é de converter uma única coluna de datas em duas novas, aplicando as funções trigonométricas seno e cosseno sobre os valores originais da data. Essa abordagem traz uma vantagem, pois captura o padrão cíclico inerente às datas. Por exemplo, ao considerar o mês do ano em uma escala de 1 a 12, a distância entre os meses 12 e 1 é intuitivamente curta, representando uma transição direta de dezembro para janeiro. No entanto, essa distância não é adequadamente expressa por uma simples diferença aritmética de 11 unidades. As transformações seno e cosseno, ao incorporar a posição angular da data no ciclo anual, são capazes de preservar essa relação cíclica de maneira mais precisa.

As variáveis numéricas foram padronizadas, enquanto as categóricas foram tratadas por meio de codificação *one-hot*. Adicionalmente, para equilibrar a variável resposta, foi aplicado undersampling, reduzindo a predominância da classe majoritária. A Figura 7 apresenta o *pipeline* completo de pré-processamento de dados.

Figura 7 – *Pipeline* de pré-processamento dos dados.



Fonte: Autor (2023)

3.1.5.1 Treinamento sem tunagem de hiperparâmetros

A etapa inicial do treinamento dos modelos foi conduzida sem a tunagem de hiperparâmetros. A estratégia adotada envolveu a aplicação de sete algoritmos de aprendizado de máquina, com o objetivo de avaliar o desempenho inicial de cada modelo antes de otimizar seus hiperparâmetros.

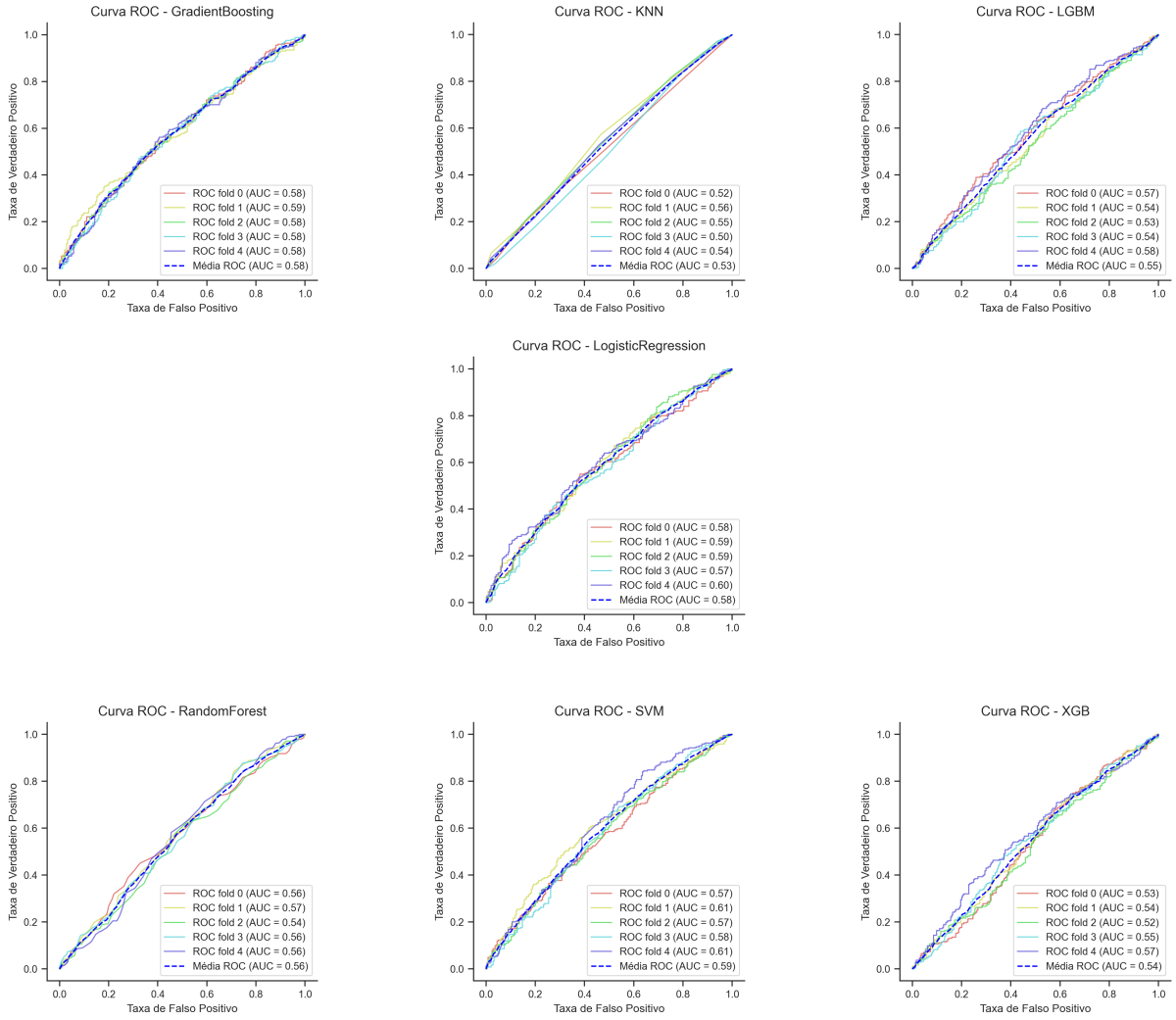
Inicialmente, foi desenvolvido um *Pipeline* que abrangeu os dois estágios necessários: o pré-processamento dos dados e o ajuste do modelo. A avaliação do desempenho dos modelos ocorreu por meio de validação cruzada estratificada com cinco *folds* (*Stratified K-Fold Cross-Validation*). Métricas fundamentais para a tarefa de classificação binária foram empregadas, abrangendo acurácia, precisão, *recall* e a área sob a curva ROC (AUC-ROC).

É relevante ressaltar a importância de cada uma dessas métricas para esse tipo de análise, que visa prever reinternações de pacientes. A acurácia reflete a precisão global do modelo, enquanto a precisão destaca a proporção de instâncias positivas corretamente classificadas. O *recall*, por sua vez, destaca a habilidade do modelo em capturar a totalidade das instâncias positivas, sendo crucial quando o foco está na minimização de falsos negativos, ou seja, quando é crucial identificar todos os casos de reinternação. Já a área sob a curva ROC proporciona uma métrica abrangente da capacidade discriminativa do modelo em diferentes limiares de probabilidade, sendo particularmente útil em tarefas de classificação desbalanceada (GÉRON, 2022).

Cada modelo foi treinado e avaliado em cada uma das cinco partições do conjunto de treinamento. A Figura 8 apresenta as curvas ROC e os valores da área para cada um

dos *folds*.

Figura 8 – Curvas ROC para cada um dos classificadores e cada *fold* da validação cruzada



Fonte: Autor (2023)

De acordo com as curvas ROC, é possível observar que nenhum dos modelos se destacou em relação aos outros no treinamento sem a tunagem de hiperparâmetros. O classificador que obteve o melhor AUC durante o treinamento foi o SVM, com uma área no valor de 0.59.

A Tabela 2 resume os resultados obtidos durante o treinamento e a análise revela insights importantes sobre o desempenho dos modelos. Observando as métricas fornecidas, pode-se observar que o modelo SVM alcançou a maior acurácia, precisão e AUC entre todos os classificadores, indicando uma melhor capacidade geral de fazer previsões corretas e distinguir entre as classes positiva e negativa. Por outro lado, o modelo KNN demonstrou o desempenho mais baixo em todas essas métricas, sugerindo que sua capacidade preditiva pode ser menos confiável em comparação com os outros modelos.

Tabela 2 – Resultados obtidos após o treinamento dos modelos sem tunagem de hiperparâmetros

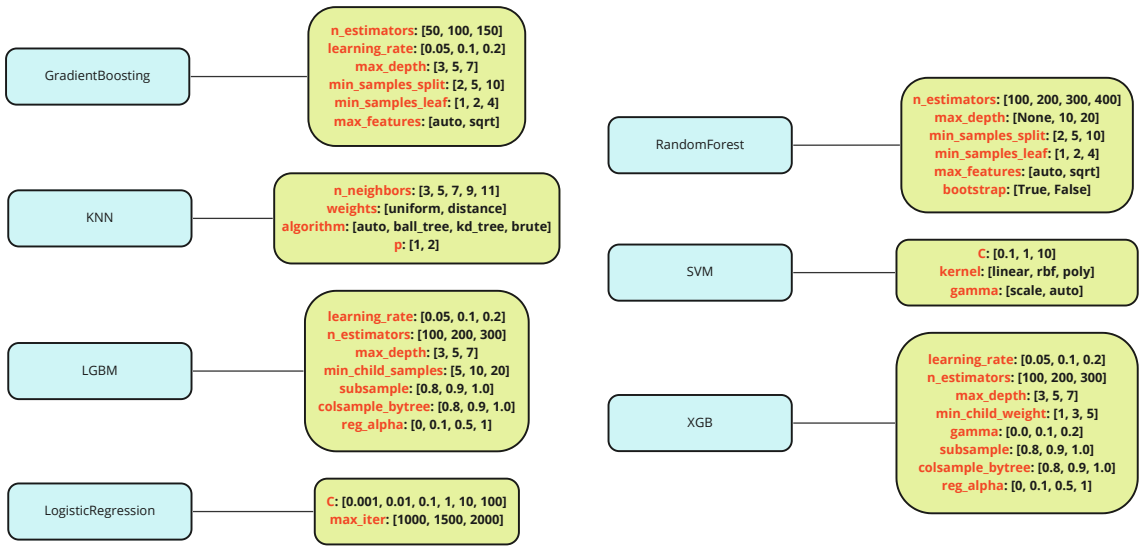
Modelo	Acurácia	Precisão	Recall	AUC
GradientBoosting	0.555738	0.555502	0.562295	0.581794
KNN	0.525410	0.525395	0.519672	0.534181
LGBM	0.539754	0.539814	0.533607	0.552842
LogisticRegression	0.559426	0.559606	0.557377	0.585189
RandomForest	0.534426	0.533356	0.550820	0.550509
SVM	0.563115	0.563822	0.555738	0.586848
XGB	0.534426	0.534750	0.527869	0.542001

Fonte: Autor (2023)

3.1.5.2 Treinamento com tunagem de hiperparâmetros

A etapa subsequente do treinamento dos modelos envolveu a tunagem de hiperparâmetros para otimizar o desempenho de cada classificador. A Figura 9 mostra quais conjuntos de hiperparâmetros foram considerados para cada um dos modelos testados. A estratégia visava encontrar a combinação ideal de hiperparâmetros que maximizasse o desempenho preditivo de cada modelo.

Figura 9 – Classificadores e seus respectivos conjuntos de hiperparâmetros testados na etapa de treinamento com tunagem.



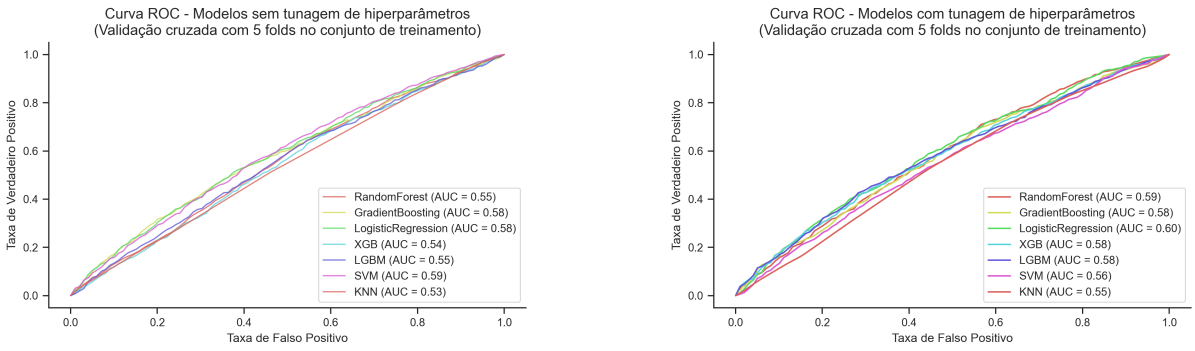
Fonte: Autor (2023)

O treinamento desta etapa seguiu exatamente a mesma metodologia explicada na subseção 3.1.5.1, com a única diferença sendo que, dentro do *loop* que percorre cada

modelo, o *GridSearchCV* foi configurado para explorar o espaço de hiperparâmetros definido para cada classificador. O conjunto de hiperparâmetros que proporcionou o melhor desempenho, medido pela métrica da área sob a curva ROC, foi selecionado como a configuração ótima para aquele modelo específico. Cabe ressaltar essa foi a métrica escolhida como critério principal para a tunagem de hiperparâmetros, uma vez que fornece uma medida robusta da capacidade discriminativa dos modelos.

A Figura 10 compara as curvas ROC de cada um dos modelos nos dois métodos de treinamento (com e sem tunagem de hiperparâmetros). Nota-se que, no treinamento sem tunagem, os classificadores GradientBoosting, LogisticRegression e SVM demonstram um desempenho superior em relação aos demais modelos, evidenciado pelo posicionamento dessas três curvas acima das demais. No entanto, no treinamento com tunagem, percebe-se uma maior proximidade entre as curvas, indicando que os modelos conseguiram aprimorar o desempenho, mesmo que marginalmente, com a tunagem.

Figura 10 – Comparação das curvas ROC para cada classificador nas etapas de treinamento com e sem tunagem de hiperparâmetros



Fonte: Autor (2023)

A Tabela 3 apresenta os resultados obtidos no treinamento com tunagem. Observando os resultados, é possível destacar que o modelo LogisticRegression se sobressai em várias métricas, apresentando a mais alta acurácia, precisão e AUC entre todos os classificadores. Esse desempenho superior sugere que o modelo LogisticRegression é mais capaz de fazer previsões corretas e distinguir entre as classes positiva e negativa.

No entanto, é crucial considerar que o desempenho de um modelo pode variar dependendo do contexto específico e dos objetivos da aplicação. Por exemplo, se o foco principal for minimizar falsos negativos (identificar corretamente casos de reinternação), pode-se observar que o modelo SVM possui um recall notavelmente alto. Isso indica que o SVM é mais eficaz em capturar a totalidade das instâncias positivas em comparação com outros modelos. Dessa forma, a escolha do "melhor" modelo depende dos requisitos específicos do problema em questão. O LogisticRegression destaca-se em termos gerais,

enquanto o SVM parece ter uma capacidade superior em identificar casos de reinternação.

Tabela 3 – Resultados obtidos após o treinamento dos modelos com tunagem de hiperparâmetros

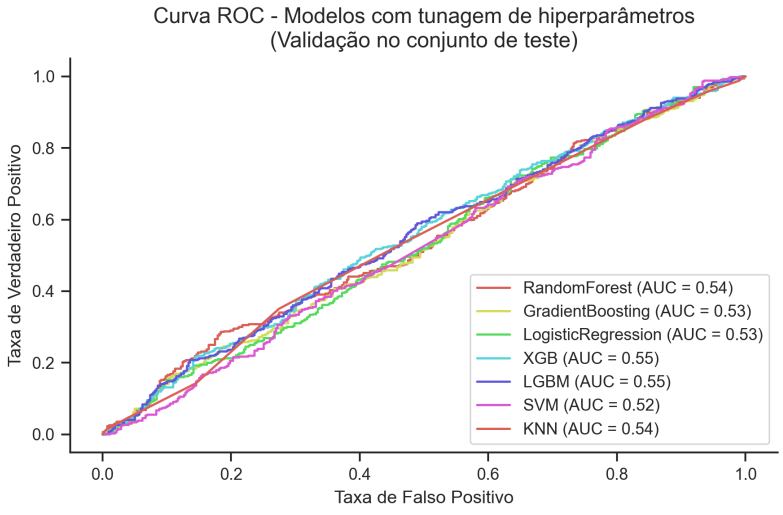
Modelo	**Acurácia**	**Precisão**	**Recall**	**AUC**
RandomForest	0.554917	0.552547	0.577719	0.579562
GradientBoosting	0.550615	0.550843	0.549116	0.571107
LogisticRegression	0.564208	0.564933	0.559426	0.587308
XGB	0.544608	0.544739	0.541992	0.560123
LGBM	0.543404	0.543352	0.543370	0.558586
SVM	0.547154	0.543553	0.663115	0.581624
KNN	0.528811	0.528966	0.524836	0.537458

Fonte: Autor (2023)

3.1.5.3 Validação no conjunto de teste

A fase de validação dos modelos, agora equipados com os melhores conjuntos de hiperparâmetros, representa um passo importante para avaliar a capacidade preditiva no mundo real. A Figura 11 exibe as curvas ROC de cada modelo, revelando que os classificadores XGB e LGBM destacaram-se pela capacidade de generalização. Esta capacidade demonstra a habilidade desses modelos em prever com maior precisão instâncias que não foram previamente observadas durante o treinamento.

Figura 11 – Curvas ROC para cada classificador na etapa de validação no conjunto de teste



Fonte: Autor (2023)

A Tabela 4 apresenta os resultados da validação no conjunto de teste, fornecendo uma análise abrangente das métricas-chave. Notavelmente, o modelo LGBM liderou em acurácia e precisão, enquanto o XGB liderou em AUC, mesmo que a discrepância entre as métricas de ambos os classificadores esteja na ordem da terceira casa decimal. O modelo SVM se destaca, novamente, exibindo um valor notavelmente alto de *recall*.

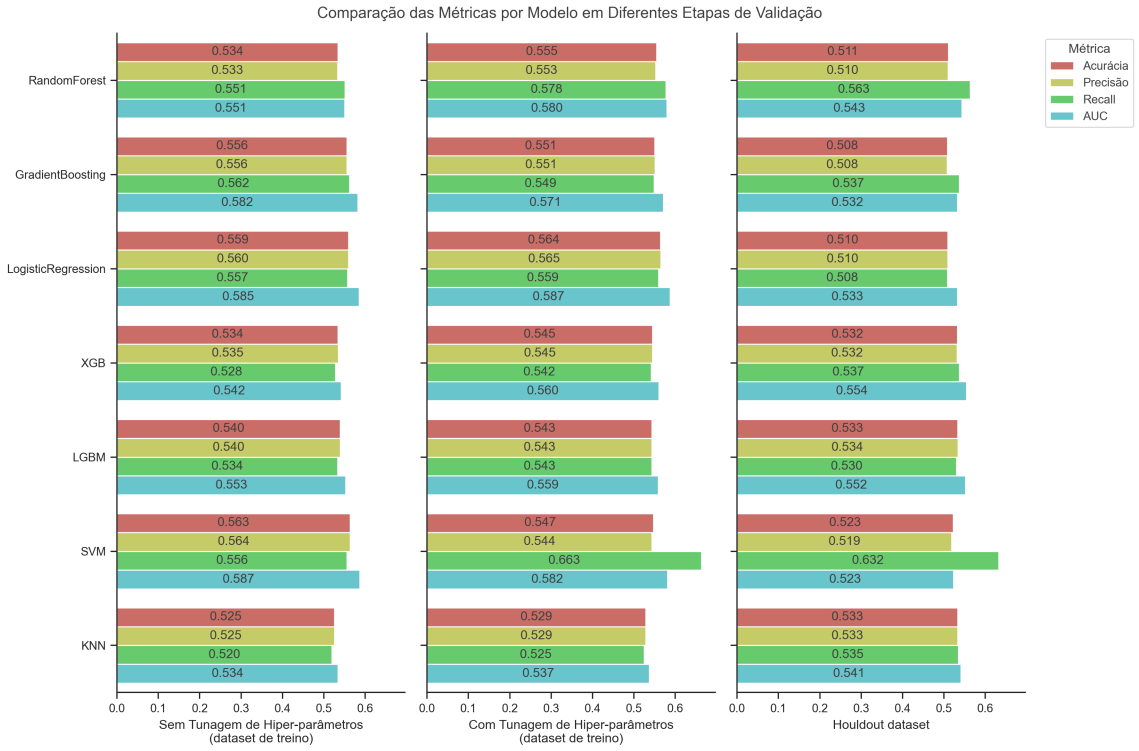
Tabela 4 – Resultados obtidos na validação no conjunto de teste

Modelo	**Acurácia**	**Precisão**	**Recall**	**AUC**
RandomForest	0.510740	0.509890	0.553699	0.541202
GradientBoosting	0.522673	0.521542	0.548926	0.545679
LogisticRegression	0.509547	0.509569	0.508353	0.532521
XGB	0.532220	0.531915	0.536993	0.554218
LGBM	0.533413	0.533654	0.529833	0.551911
SVM	0.522673	0.518591	0.632458	0.520830
KNN	0.533413	0.533333	0.534606	0.541322

Fonte: Autor (2023)

A Figura 12 apresenta um resumo abrangente, comparando todas as métricas em todas as fases de validação, incluindo treinamento sem e com tunagem, bem como no conjunto de teste, para todos os modelos.

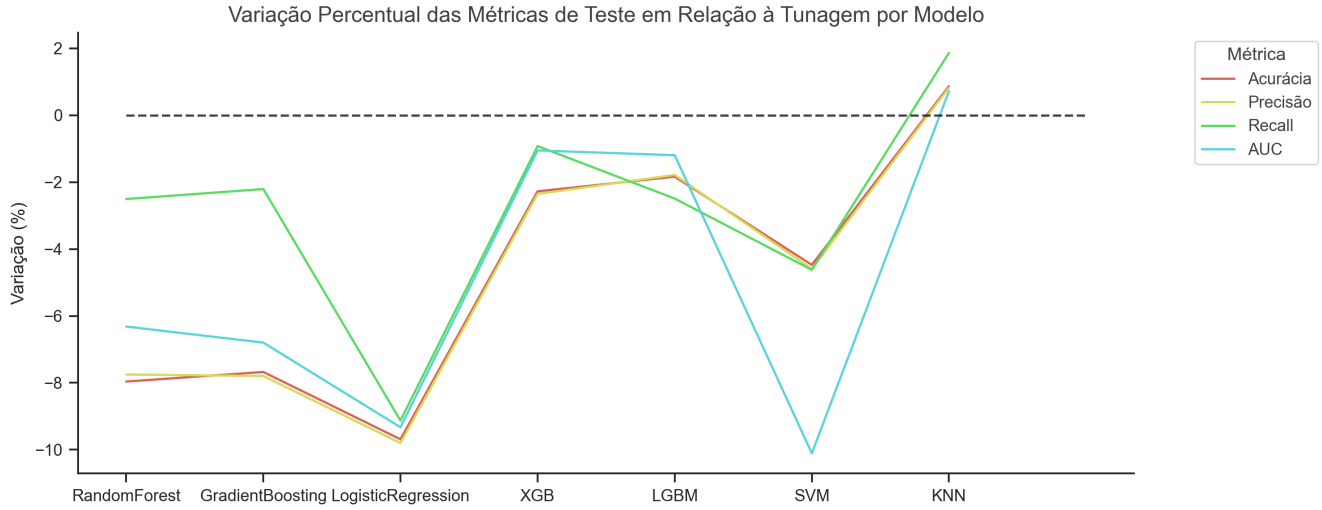
Figura 12 – Comparação das métricas de desempenho em diferentes etapas de validação (treinamento sem e com tunagem, e teste) para os modelos



Fonte: Autor (2023)

Também é possível observar, na Figura 13, a variação percentual das métricas no conjunto de teste em comparação com as métricas na etapa de tunagem. Com exceção do classificador KNN, todos os modelos apresentaram uma degradação nas métricas. Notavelmente, o LogisticRegression foi o modelo com a pior capacidade de generalização, registrando uma variação de aproximadamente -8% em todas as métricas. Em contrapartida, os modelos XGB e LGBM sofreram as menores quedas, mantendo-se próximos do valor de -4%. Considerando esses resultados, ambos os modelos se destacam como candidatos viáveis para a escolha do melhor modelo. Optou-se por selecionar o modelo LGBM, que será utilizado na análise subsequente das variáveis mais importantes, empregando a abordagem SHAP.

Figura 13 – Variação percentual das métricas no conjunto de teste em comparação com as métricas na etapa de tunagem



Fonte: Autor (2023)

3.1.6 Análise de variáveis mais importantes

Por fim, será utilizada a biblioteca `shap` para avaliar o impacto das variáveis no processo de tomada de decisões dos modelos. Isso permitirá a identificação de quais características estão mais fortemente associadas a casos de reinternação hospitalar, fornecendo uma visão detalhada das relações entre variáveis e resultados.

REFERÊNCIAS

ABOUOUF, M. *et al.* Explainable ai for event and anomaly detection and classification in healthcare monitoring systems. **IEEE Internet of Things Journal**, p. 1–1, 2023.

ASIF, M. *et al.* Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. **PLOS ONE**, Public Library of Science, v. 13, n. 12, p. 1–15, 12 2018. Available at: https://doi.org/10.1371/journal.pone.0208626.

BACH, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. **PLOS ONE**, Public Library of Science, v. 10, n. 7, p. 1–46, 07 2015. Available at: https://doi.org/10.1371/journal.pone.0130140.

BARROS, R. E. M. *et al.* Impact of length of stay for first psychiatric admissions on the ratio of readmissions in subsequent years in a large Brazilian catchment area. **Social Psychiatry and Psychiatric Epidemiology**, v. 51, n. 4, p. 575–587, abr. 2016. ISSN 1433-9285. Available at: https://doi.org/10.1007/s00127-016-1175-x.

BATISTA, A. F. de M.; FILHO, A. D. P. C. Machine learning aplicado à saúde. Sociedade Brasileira de Computação, 2019. Available at: https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/29/95/245-1?inline=1.

BRUNESE, L. *et al.* Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. **Computer Methods and Programs in Biomedicine**, v. 196, p. 105608, 2020. ISSN 0169-2607. Available at: https://www.sciencedirect.com/science/article/pii/S0169260720314413.

BUITINCK, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. **CoRR**, abs/1309.0238, 2013. Available at: http://arxiv.org/abs/1309.0238.

CESCONETTO, A.; LAPA, J. dos S.; CALVO, M. C. M. Avaliação da eficiência produtiva de hospitais do sus de santa catarina, brasil. **Revista Saúde Pública**, 2008.

CONFALONIERI, R. *et al.* A historical perspective of explainable artificial intelligence. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, John Wiley & Sons Ltd., v. 11, n. 1, p. 21, 2021. ISSN 1942-4787.

DOSHI-VELEZ, F.; KIM, B. **Towards A Rigorous Science of Interpretable Machine Learning**. 2017.

FACELI, K. *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina**. [*S.l.: s.n.*]: LTC, 2011.

FÁVERO, L. P. L. *et al.* **Análise de dados: modelagem multivariada para tomada de decisões**. [*S.l.: s.n.*]: Elsevier, 2009.

FERNANDES, F. T. *et al.* A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil. **Scientific reports**, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2021.

FILHO, A. D. P. gatto. Uso de big data em saúde no brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, SciELO Public Health, v. 24, p. 325–332, 2015.

FRIEDMAN, J. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, 11 2000.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition**. [*S.l.: s.n.*]: O'Reilly Media, Inc., 2022. ISBN 9781098125974.

GHOSHROY, D.; ALVI, P. A.; SANTOSH, K. Unboxing industry-standard ai models for male fertility prediction with shap. **Healthcare**, v. 11, n. 7, 2023. ISSN 2227-9032. Available at: https://www.mdpi.com/2227-9032/11/7/929.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2. ed. [*S.l.: s.n.*]: Springer New York Inc., 2009. ISBN 9780387848587.

HO, T. K. Random decision forests. *In*: **Proceedings of 3rd International Conference on Document Analysis and Recognition**. [*S.l.: s.n.*], 1995. v. 1, p. 278–282 vol.1.

KHARRAT, F. G. Z. *et al.* Feature sensitivity criterion-based sampling strategy from the optimization based on phylogram analysis (fs-opa) and cox regression applied to mental disorder datasets. **PLOS ONE**, Public Library of Science, v. 15, n. 7, p. 1–25, 07 2020. Available at: https://doi.org/10.1371/journal.pone.0235147.

KIM, Y.; KIM, Y. Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. **Sustainable Cities and Society**, v. 79, p. 103677, 2022. ISSN 2210-6707. Available at: https://www.sciencedirect.com/science/article/pii/S2210670722000117.

KLEINBAUM, D. G.; KLEIN, M. **Logistic Regression**. 3. ed. [*S.l.: s.n.*]: Springer New York, NY, 2010. ISBN 9781441917416.

LIPTON, Z. C. **The Mythos of Model Interpretability**. 2017.

LORETO, M.; LISBOA, T.; MOREIRA, V. P. Early prediction of icu readmissions using classification algorithms. **Computers in Biology and Medicine**, v. 118, p. 103636, 2020. ISSN 0010-4825. Available at: https://www.sciencedirect.com/science/article/pii/S0010482520300329.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. Curran Associates, Inc., p. 4765–4774, 2017. Available at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

MESKO, B.; GOROG, M. A short guide for medical professionals in the era of artificial intelligence. **npj Digital Medicine**, Nature Publishing Group, v. 3, p. 126, 2020.

MIYOSHI, N. S. B. *et al.* An ehealth platform for the support of a brazilian regional network of mental health care (ehealth-interop): development of an interoperability platform for mental care integration. **JMIR Mental Health**, 2018.

OBERMEYER, Z.; LEE, T. H. Lost in thoughtthe limits of the human mind and the future of medicine. **New England Journal of Medicine**, Mass Medical Soc, v. 377, n. 13, p. 1209–1211, 2017.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**. 2. ed. [*S.l.: s.n.*]: Packt, 2017. ISBN 9781787125933.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **Model-Agnostic Interpretability of Machine Learning**. 2016.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. **CoRR**, abs/1602.04938, 2016. Available at: http://arxiv.org/abs/1602.04938.

ROSA, A. C. F. d. *et al.* Uso de técnicas de aprendizado de máquina para classificação de fatores que influenciam a ocorrência de dermatites ocupacionais. **Revista Brasileira de Saúde Ocupacional**, Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho - FUNDACENTRO, v. 48, 2023. ISSN 0303-7657. Available at: https://doi.org/10.1590/2317-6369/31620pt2023v48e4.

SANTOS, H. G. d. *et al.* Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de são paulo, brasil. **Cadernos de saúde pública**, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 35, n. 7, 2019. ISSN 0102-311X.

SELTZER, M. L.; ZHANG, L. The data deluge: Challenges and opportunities of unlimited data in statistical signal processing. IEEE, p. 3701–3704, 2009. Available at: https://doi.org/10.1109/ICASSP.2009.4960430.

SHRIKUMAR, A. *et al.* Not just a black box: Learning important features through propagating activation differences. **CoRR**, abs/1605.01713, 2016. Available at: http://arxiv.org/abs/1605.01713.

SMILKOV, D. *et al.* Smoothgrad: removing noise by adding noise. **CoRR**, abs/1706.03825, 2017. Available at: http://arxiv.org/abs/1706.03825.

STIGLIC, G. *et al.* Pediatric readmission classification using stacked regularized logistic regression models. **AMIA Annu Symp Proc**, p. 1072–1081, 2014.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and Information Systems**, v. 41, n. 3, p. 647–665, Dec 2014. ISSN 0219-3116. Available at: https://doi.org/10.1007/s10115-013-0679-x.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. **CoRR**, abs/1703.01365, 2017. Available at: http://arxiv.org/abs/1703.01365.

WANG, B. *et al.* Predictive classification of icu readmission using weight decay random forest. **Future Generation Computer Systems**, v. 124, p. 351–360, 2021. ISSN 0167-739X. Available at: https://www.sciencedirect.com/science/article/pii/S0167739X21002065.

WANG, L. Research and implementation of machine learning classifier based on knn. **IOP Conference Series: Materials Science and Engineering**, IOP Publishing, v. 677, n. 5, p. 052038, dec 2019. Available at: https://dx.doi.org/10.1088/1757-899X/677/5/052038.

Wikipedia contributors. **Support vector machine — Wikipedia, The Free Encyclopedia**. 2023. [Online; accessed 6-January-2024]. Available at: https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1190739318.

YIN, V. *et al.* Centers for disease control (cdc) wound classification is prognostic of 30-day readmission following surgery. **World Journal of Surgery**, v. 47, n. 10, p. 2392–2400, 2023. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1007/s00268-023-07093-3.