UNIVERSIDADE FEDERAL DE SÃO PAULO
INSTITUTO DE CIÊNCIA E TECNOLOGIA
GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

Comparação de algoritmos de árvore de decisão para a identificação de risco cardiovascular: abordagem baseada em dados gerais de saúde coletados por pesquisas transversais

Aluno: Pedro Henrique Crisp Modesto
Orientador: Prof. Dr. Adenauer Girardi Casali

Trabalho apresentada ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos, SP.
2023

FOLHA DE APROVAÇÃO

PEDRO HENRIQUE CRISP MODESTO

Comparação de algoritmos de árvore de decisão para a identificação de risco cardiovascular: abordagem baseada em dados gerais de saúde coletados por pesquisas transversais

Trabalho apresentada ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos, 11 de dezembro de 2023.

Banca Avaliadora

Prof. Dr. Frederico Aletti
(Presidente - Unifesp)

Prof. Dr. Henrique Mohallem Paiva
(Membro 1 - Unifesp)

Prof. Dr. Henrique Alves de Amorim
(Membro 2 - Unifesp)

Elaborado por sistema de geração automática com os dados fornecidos pelo(a) autor(a).

Henrique Crisp Modesto, Pedro
Comparação de algoritmos de árvore de decisão para a identificação de risco cardiovascular: abordagem baseada em dados gerais de saúde coletados por pesquisas transversais/ Pedro Henrique Crisp Modesto
Orientador(a) Adenauer Girardi Casali-São José dos Campos, 2023.
8 p.

Trabalho de Conclusão de Curso-Engenharia Biomédica-Universidade Federal de São Paulo-Instituto de Ciência e Tecnologia, 2023.

1. Aprendizado de Máquina. 2. Infarto do Miocárdio. 3. Doença Arterial Coronariana. 4. Árvores de Decisão. 5. XGBoosting. I. Girardi Casali, Adenauer , orientador(a). II. Título.

Trabalho apresentada ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

Comparação de algoritmos de árvore de decisão para a identificação de risco cardiovascular: abordagem baseada em dados gerais de saúde coletados por pesquisas transversais

Pedro Henrique Crisp Modesto[1], Adenauer Girardi Casali[2]

[1] Instituto de Ciência e Tecnologia da Unifesp / Engenharia Biomédica

Resumo - **O infarto do miocárdio (IM) e a doença arterial coronariana (DAC) estão entre as condições mais custosas aos cofres públicos do Sistema Único de Saúde (SUS). A prevenção de tais condições depende de exames como a angiografia, que possuem alto custo e são de limitado acesso a pessoas de baixa renda. É neste contexto que diversos estudos têm buscado empregar estratégias de aprendizado de máquina para aprimorar o reconhecimento automatizado de pacientes em risco de IM e DAC. Com o presente trabalho pretendemos investigar a possibilidade de se empregar algoritmos baseados em árvores para classificar e interpretar informações gerais sobre o estado de saúde de pacientes em risco de doenças cardiovasculares. Três algoritmos baseados em árvores de decisão foram testados empregando-se um banco de dados coletados pelo Centro de Controle de Doença dos Estados Unidos (CDC). Os resultados indicam eficácia dos métodos em termos tanto de acurácia quanto de detecção de verdadeiros positivos e sugerem que tais algoritmos, sobretudo quando aliados a uma análise de interpretação dos fatores de risco, possuem potencial de serem empregados para guiar estratégias de baixo custo na prevenção de doenças cardiovasculares.**

Abstract — **Myocardial infarction (MI) and coronary artery disease (CAD) rank among the most financially burdensome conditions for public healthcare systems such as Brazil's Unified Health System (SUS). Preventing these conditions often relies on diagnostic procedures such as angiography, which are not only expensive but also scarcely accessible to low-income individuals. It is within this context that numerous studies have sought to employ machine learning strategies to enhance automated recognition of patients at risk of MI and CAD. Here we aim to investigate the feasibility of using tree-based algorithms to classify and interpret general health data of patients potentially at risk of cardiovascular diseases. We tested three algorithms based on decision trees using a database from the United States Centers for Disease Control and Prevention (CDC). The results demonstrate the effectiveness of these methods, in terms of both accuracy and detection of true positives, suggesting that such algorithms - especially when combined with an interpretability analysis of risk factors – hold promise to be employed in the development of low-cost strategies to prevent cardiovascular diseases.**

Keywords—**Aprendizado de Máquina, Infarto do Miocárdio, Doença Arterial Coronariana, Árvore de Decisão, XGBoosting**

I. INTRODUÇÃO

Doenças cardiovasculares são as principais causas de morte no Brasil e no mundo. Estudos apontam que pelo menos 400 mil brasileiros morrem todos os anos em decorrência de problemas cardíacos [1]. Dentre as doenças cardiovasculares mais conhecidas pela sociedade estão a doença arterial coronariana (DAC), o acidente vascular cerebral (AVC) e o infarto do miocárdio (IM). Embora haja diversos estudos mostrando significativa associação dessas doenças com fatores genéticos, hábitos alimentares, raça e outros fatores de risco, devido à ampla quantidade de diagnósticos diferenciais nem sempre é fácil detectar clinicamente pacientes com maior risco [2]. Além disso, estima-se que um em cada cinco ataques cardíacos seja assintomático, ou seja, ocorra sem que a pessoa esteja ciente do seu potencial risco até que o dano venha a ser causado [3]. Esses casos frequentemente resultam em complicações de longo prazo ou até mesmo na morte do indivíduo. No ano de 2015, IM foi a doença que mais custou aos cofres públicos do Sistema Único de Saúde (SUS), cerca de 22,2 bilhões de reais [4]. Atualmente, a angiografia é o exame mais preciso para prever doença arterial cardíaca, porém o alto custo deste exame impede o seu acesso à população mais pobre. Neste contexto, técnicas de aprendizado de máquina têm sido empregadas para o desenvolvimento de métodos mais econômicos e acessíveis de identificação de pacientes com maior risco de eventos cardíacos [4,5].

É possível encontrar na literatura diversos estudos que utilizam diferente técnicas e modelagens de dados clínicos para classificar o risco de se desenvolver IM ou DAC. Em particular, os algoritmos de aprendizado de máquina baseados em árvores têm se mostrado eficientes na

identificação de pacientes em risco. Princy et al. (2020) utilizou dados antropométricos, clínicos e de exames laboratoriais para testar diferentes algoritmos de aprendizado de máquina e classificar o risco de o paciente desenvolver doenças cardíacas. Neste trabalho, o modelo que apresentou a melhor performance foi o *Decision Tree*, atingindo uma acurácia de 73% [2]. Já Ishaq et al. (2021) utilizou diferentes modelos em dados de exames laboratoriais, reportando o algoritmo baseado em árvore *Extra Tree Classifier* como tendo atingido uma acurácia de 92,6%, com 93% de precisão e 93% de F1-Score [5]. Por fim, Muntasir Nishat et al. (2022) também obteve uma performance de 90% de acurácia, desta vez utilizando o modelo *Random Forest*, em dados de exames laboratoriais e estado clínico dos pacientes [6].

Todas estas referências anteriores empregaram exames laboratoriais na identificação do risco de IM ou DAC. Neste trabalho pretendemos testar o potencial de tais algoritmos para prever risco de eventos cardíacos a partir de dados clínicos de fácil acesso, que descrevem hábitos e a saúde geral dos pacientes e que podem ser obtidos por chamadas telefônicas. Para tal, usaremos um banco de dados do Centro de Prevenção e Controle de Doença dos Estados Unidos (CDC) e confrontaremos algoritmos baseados em árvores convencionais, como *Random Forest* e *Decision Tree*, com o XGBoosting, um algoritmo mais recente e ainda não testado em tal tipo de problema, em que as árvores são construídas de uma maneira sequencial para gradualmente minimizar erros de forma eficiente [12].

Por fim, através de uma análise dos marcadores empregados nos modelos aqui testados, pretendemos usar estas estratégias de aprendizado de máquina não só para classificação dos padrões, mas também para identificação de quais são os principais fatores de risco para IM e DAC no banco de dados utilizado, de forma a tornar o resultado deste processo interpretável para o seu uso clínico.

II. MATERIAL E MÉTODOS

Utilizamos uma base de dados pública com 18 características clínicas extraídas em mais de 300 mil indivíduos em um estudo transversal realizado pelo Centro de Prevenção e Controle de Doença dos Estados Unidos (CDC), onde 8,6% dos pacientes afirmaram já terem sido diagnosticados com DAC ou IM.

Anualmente, o centro coleta dados por telefone sobre o estado de saúde da população norte-americana. Essa base de dados se encontra disponível no site Kaggle e já previamente tratada com a remoção de argumentos nulos [7].

As características que estão disponíveis no banco de dados são elencadas na Tabela 1.

Tabela 1: Características disponíveis na base de dados

Característica	Descrição	Tipo
HeartDisease	Flag de identificação de pacientes que tiveram infarto do miocárdio ou doença arterial coronariana	Booleana
BMI	Índice de Massa Muscular	Numérica
Smoking	Flag de identificação de fumantes	Booleana
AlcoholDrinking	Flag de identificação de alcoólatras	Booleana
Stroke	Flag de identificação de pacientes que já tiveram derrame cerebral	Booleana
PhysicalHealth	Quantidade de dias que os pacientes tiveram problemas físicos no último mês	Numérica
MentalHealth	Quantidade de dias que os pacientes tiveram problemas emocionais no último mês	Numérica
DiffWalking	Flag de identificação de pacientes que possuem dificuldade de andar e/ou problemas ao subir escadas	Booleana
Sex	Gênero	Texto
AgeCategory	Faixa etária	Texto
Race	Etnia	Texto
Diabetic	Identificação de pacientes diabéticos e pré-diabéticos	Texto
PhysicalActivity	Flag de identificação de pacientes que praticam atividade física	Booleana
GenHealth	Avaliação do paciente com seu estado de saúde	Texto
SleepTime	Média diária de horas dormidas	Numérica
Asthma	Flag de identificação de pacientes que possuem asma	Booleana
KidneyDisease	Flag de identificação de pacientes que possuem doença renal	Booleana
SkinCancer	Flag de identificação de pacientes que possuem ou tiveram câncer de pele	Booleana

Após a extração dos dados, executamos uma etapa de pré-processamento para transformar as variáveis booleanas e categóricas de texto em formatos numéricos, com o objetivo de facilitar a aplicação de técnicas de aprendizado de máquina supervisionado.

A partir dos dados pré-processados, foi realizada uma análise com o objetivo de identificar a frequência e a identificação de *outliers*. Nesta etapa, observando-se a distribuição de valores de *BMI* e *SleepTime*, foi notado a presença de valores que divergem significativamente de padrões esperados para adultos. Foram removidos, portanto, todos os padrões que reportavam:

1) Índices de massa corporal inferior a 14 kg/m^2 ou superior a 60 kg/m^2;

2) Média de horas dormidas por dia inferior a 4 horas ou superior a 14 horas.

A análise quantitativa das relações entre a característica alvo e as demais características foi realizada quantificando-se as probabilidades condicionais das características, dada a variável alvo.

Para avaliar se as probabilidades determinadas pela análise condicional eram estatisticamente significativas, o teste qui-quadrado (χ^2) foi empregado. O teste χ^2 é amplamente utilizado para examinar hipóteses entre variáveis independentes categóricas e as frequências observadas e esperadas entre grupos [9].

Os dados empregados no estudo possuem um grande desbalanceamento de classes, já que majoritariamente os pacientes não tiveram diagnóstico de doenças cardíacas (91,4%). Este desbalanço de classes deve ser levado em consideração durante o treinamento de uma técnica de aprendizado de máquina, já que a falta de exemplos da classe minoritária faz com que o algoritmo possa naturalmente se enviesar em direção à classe majoritária [19]. Para tratar este problema, os algoritmos de árvores foram empregados em conjunto a três metodologias de balanceamento de classes: sobreamostragem, subamostragem e combinação de ambas. O conjunto de dados foi dividido, utilizando uma proporção de 75% para o treinamento após o balanceamento, deixando 25% para a fase de teste. Os algoritmos foram submetidos a um procedimento de validação cruzada 5-fold e avaliados no conjunto de teste, junto com as características mais importantes no modelo.

A. Algoritmos de Aprendizado de Máquina Empregados:

Decision Tree (DT): Trata-se de um algoritmo recursivo que se baseia em ramos e nós, onde ramos são os caminhos das decisões tomadas e nós as subdivisões do conjunto de dados. As regras de cada nó são definidas de acordo com uma definição de peso de cada característica, que por sua vez é calculado pelo índice de Gini ou entropia, formando uma estrutura de árvore em que cada padrão passa pelos nós e ramos até a definição de qual classe ele pertence [10].

Random Forest (RF)*:* Consiste em um conjunto de árvores aleatórias paralelas, onde o classificador utiliza as características de maneira aleatória para criar N árvores. O modelo mescla diferentes árvores aleatórias independentes, em que a classificação é definida de acordo com procedimento de votação, onde é selecionada a classe que obteve maior número de votos entre as árvores [11].

XGBoosting (XGB): Algoritmo baseado em árvores de decisão e em aumento de gradiente aprimorados. Assim como o RF, este algoritmo também trabalha no desenvolvimento de diferentes árvores onde cada uma contribui para o modelo geral. Porém, aqui as árvores são desenvolvidas durante o treinamento de maneira sequencial, tal que as próximas árvores são ajustadas para minimizar o erro cometido pelas anteriores [20].

B. Técnicas de reamostragem dos dados utilizadas:

Sobreamostragem: Foi utilizado a técnica de SMOTE (*Syntetic Minority Oversampling Technique*) que cria novos argumentos sintéticos da classe minoritária a partir de instâncias próximas no espaço do hiperplano até que possua uma proporção equilibrada para o treinamento do algoritmo [13].

Subamostragem: Aplicou-se a técnica de *Random Undersampling*, no qual ocorre a redução do conjunto de treinamento aleatório da classe majoritária, onde são eliminados aleatoriamente argumentos até atingir uma proporção equilibrada entre as classes [13].

Combinação entre Sobre e Subamostragem: Para equilibrar as distribuições das classes, evitar perda de dados relevantes e complicações de sobreajuste, testamos a técnica de SMOTE-ENN (*Synthetic Minority Oversampling Technique and Edited Nearest Neighbor*), que combina as técnicas de sobreamostragem e subamostragem até atingir um equilíbrio entre as classes [6].

C. Hiper-parametrização dos algoritmos:

Com a finalidade de atingir a melhor performance do algoritmo classificador aplicado no conjunto de dados trabalhados, utilizamos o procedimento de busca aleatória de parâmetros como implementado na biblioteca *sklearn* em Python (*RandomSearchCV*, [14]).

D. Métricas de avaliação de performance:

Para a avaliação dos modelos foram utilizadas as seguintes métricas de avaliação: acurácia, precisão, revocação e *f1-score*. As fórmulas destas métricas são calculadas a partir dos números de padrões verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos Negativos (FN) de acordo com a tabela abaixo [15]:

Tabela 2: Métrica de Avaliação do algoritmo classificador

Métrica	Fórmula
Acurácia	$\frac{VP + VN}{VP + VN + FP + FN}$
Precisão	$\frac{VP}{VP + FP}$
Revocação	$\frac{VP}{VP + FN}$

F1 - Score	$2 * \frac{Precisão * Revocação}{Precisão + Revocação}$

E. Avaliação das características mais relevantes:

Após a hiper parametrização e avaliação do algoritmo, aplicamos o índice de Gini para avaliar quais são as características mais relevantes que o modelo utilizou para gerar as previsões. Este método consiste em avaliar cada característica e calcular a "pureza" que a ela atinge na distinção das classes. É a partir das melhores características assim determinadas que são construídos os "nós" das árvores de decisão [18].

F. Materiais:

Todo o desenvolvimento do projeto foi executado em Python 3.11.4, na IDLE *Visual Studio Code*, em uma máquina HP 246 G6 Notebook PC, com processador Intel ® Core ™ I5-7200u CPU @ 2.500Ghz (4CPUs) ~2.7Ghz

III. RESULTADOS

A. Análise das Características

As figuras 3, 4 e 5 exemplificam algumas das características avaliadas na etapa de análise dos dados e suas proporções entre as classes definidas pela variável alvo (HeartDisease=1 correspondente à população com DAC e IM e HeartDisease=0 correspondente à população sem as doenças cardiovasculares):

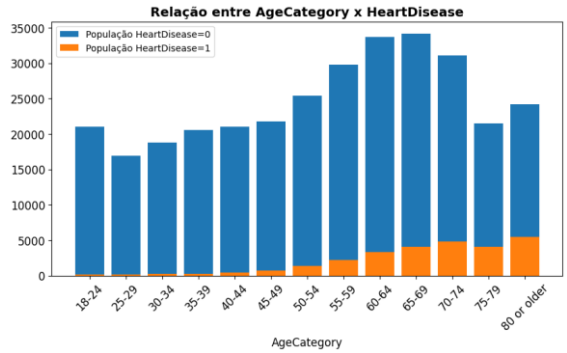

Figura 3: Faixa etária das pessoas sem doenças cardíacas (azul) e com doenças cardíacas (laranja).

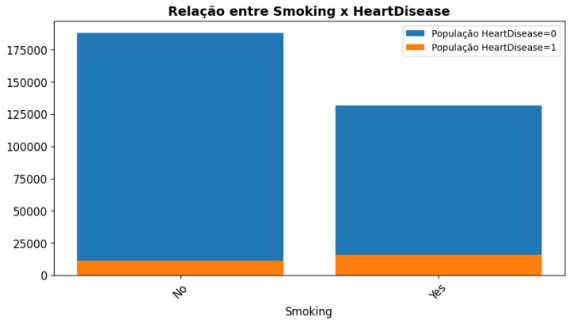

Figura 4: Proporção de fumantes das pessoas sem doenças cardíacas (azul) e com doenças cardíacas (laranja)

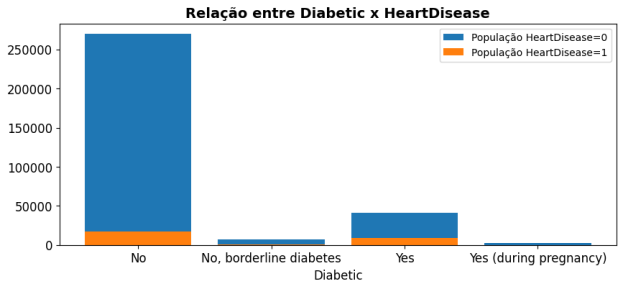

Figura 5: Proporções de diabéticos das pessoas sem doenças cardíacas (azul) e com doenças cardíacas (laranja).

Abaixo reportamos os resultados da análise das probabilidades condicionais, juntamente com o teste qui-quadrado que indicou diferenças significativas entre as classes para todas as características medidas:

Tabela 3: Resultado das probabilidades e p-valor

Condição	P(X | HeartDisease=1)	P(X | HeartDisease=0)	p-valor
IMC>25	75,4%	67,1%	$<10^{-5}$
Smooking = Yes	58,5%	39,5%	$<10^{-5}$
AlcoholDrinking = Yes	4,1%	7,0%	$<10^{-5}$
Stroke = Yes	15,7%	2,5%	$<10^{-5}$
PhysicalHealth > 15 dias	21,0%	6,6%	$<10^{-5}$
MentalHealth > 15	10,9%	7,9%	$<10^{-5}$
DiffWalking = Yes	35,9%	11,4%	$<10^{-5}$
Sex = Male	59,1%	46,5%	$<10^{-5}$
AgeCategory > 40 anos	97,2%	73,8%	$<10^{-5}$
Diabetic = Yes	32,5%	10,8%	$<10^{-5}$

PhysicalActivity = Yes	64,3%	79,0%	0.01
GenHealth = Good	35,3%	28,6%	$<10^{-5}$
SleepTime < 8 horas	56,0%	61,0%	$<10^{-5}$
Asthma = Yes	17,6%	12,9%	$<10^{-5}$
KidneyDisease = Yes	12,4%	2,8%	$<10^{-5}$
SkinCancer = Yes	18,3%	8,5%	$<10^{-5}$

B. **Comparação dos Modelos e técnicas de balanceamento**

Após a etapa de avaliação das características, os algoritmos classificadores foram testados junto às diferentes metodologias utilizadas para balanceamento de classes. Os resultados do desempenho dos modelos testados estão exibidos nas tabelas 3, 4 e 5, com a indicação em verde dos melhores desempenhos obtidos para cada métrica:

Tabela 3: Resultados dos modelos com sobreamostragem

Métrica	**DT**	**RF**	**XGB**
Acurácia	82,9%	85,7%	82,9%
Precisão	20,5%	25,0%	24,6%
F1 - Score	25,8%	28,7%	32,6%
Revocação	34,9%	33,8%	48,4%

Tabela 4: Resultados dos modelos com subamostragem

Métrica	**DT**	**RF**	**XGB**
Acurácia	67,6%	71,9%	73,2%
Precisão	16,0%	20,0%	21,3%
F1 - Score	25,8%	31,8%	33,6%
Revocação	66,2%	76,8%	79,4%

Tabela 5: Resultados dos modelos com a combinação de métodos de subamostragem e sobreamostragem

Métrica	**DT**	**RF**	**XGB**
Acurácia	80,5%	82,6%	83,4%
Precisão	21,7%	25,7%	27,5%
F1 - Score	30,2%	35,0%	37,3%
Revocação	49,5%	55,0%	58,0%

C. **Hiper-parametrização do melhor modelo**

Considerando que o modelo XGB destacou-se em comparação às estratégias tradicionais, exibimos na figura abaixo os resultados da hiper-parametrização deste modelo. para cada método de balanceamento de classes empregado

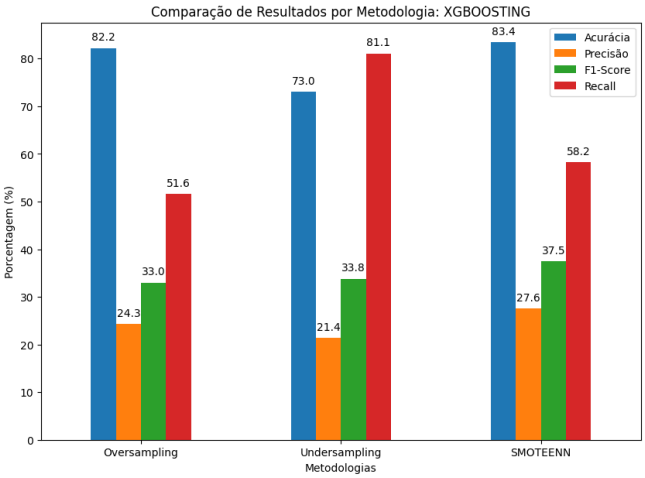

Figura 7: Resultados do modelo *XGBoosting* hiper-parametrizado

D. **Resultados finais do modelo com maior acurácia**

A figura 8 exibe a matriz de confusão do modelo hiper-parametrizado que atingiu maior acurácia, empregando a combinação de balanceamento de classes:

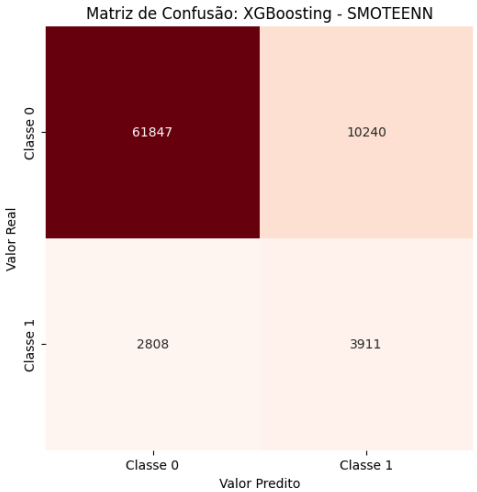

Figura 8: Matriz de confusão dos resultados preditos pelo modelo hiper-parametrizado *XGBoosting* com combinação de balanceamento de classes

As características mais relevantes obtidas a partir do índice de Gini para este modelo estão exibidas na figura 9.

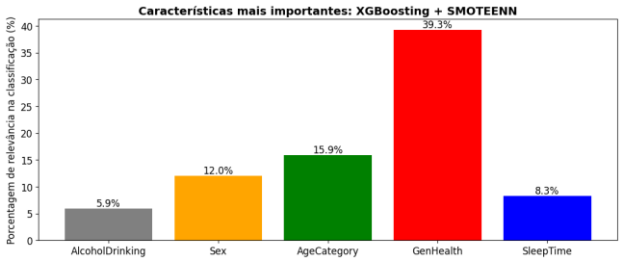

Figura 9: Características mais importantes que o modelo levou em consideração para a classificação de maior acurácia

E. Resultados finais do modelo com maior revocação

A figura 10 exibe a matriz de confusão do modelo hiper-parametrizado com a técnica de subamostragem, que atingiu maior valor de revocação.

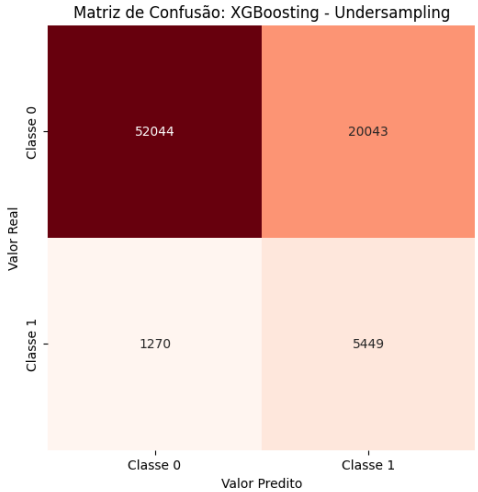

Figura 10: Matriz de confusão dos resultados preditos pelo modelo hiper-parametrizado *XGBoosting* com subamostragem para balanceamento das classes

As características mais importantes deste modelo de acordo com a análise com o índice de Gini estão exibidas na figura 11

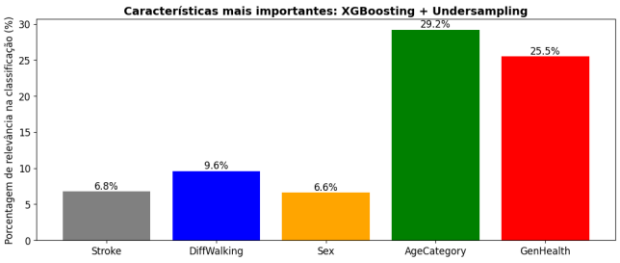

Figura 11: Características mais importantes que o modelo levou em consideração para a classificação de maior revocação

IV. DISCUSSÃO E CONCLUSÃO

Técnicas de aprendizado de máquina baseadas em árvores têm sido recentemente investigadas como potenciais ferramentas clínicas de previsão de doenças cardiovasculares. Normalmente estas técnicas são aplicadas a bancos de dados contendo resultados de exames laboratoriais específicos. Neste trabalho, investigamos o uso destes métodos em dados transversais relacionados ao estado de saúde geral dos pacientes. Nossos resultados indicam que tais algoritmos possuem potencial de apoiar estratégias de prevenção de infarto do miocárdio e doenças arterial coronariana mesmo a partir de tais dados de mais fácil acesso.

De maneira geral, o modelo *XGBoosting* destacou-se em quase todas as métricas avaliadas (Tabelas 3, 4 e 5). Este algoritmo envolve uma metodologia recente ainda pouco investigada na área biomédica. Por ser baseado em um treinamento sequencial das árvores, o método foi capaz de atingir um treinamento mais eficaz e mostrou-se robusto em termos de acurácia, independentemente da metodologia de balanceamento aplicada.

Com relação às estratégias de balanceamento, o emprego das metodologias de sobreamostragem e SMOTE-ENN resultaram em acurácias superiores a 80% (Figura 7). Porém, grande parte desta acurácia refletiu-se em uma baixa taxa de revocação, indicando baixa habilidade do modelo de identificar os casos positivos. Do ponto de vista da prevenção de condições como as doenças cardiovasculares, que podem levar a morte, a ênfase do treinamento de tais algoritmos deve residir em se minimizar o índice de falsos negativos e neste caso a revocação passa a ser claramente um indicador de maior importância clínica. Neste contexto, cabe ressaltar que muitos dos estudos existentes na literatura focam o desempenho dos métodos apenas na acurácia total dos modelos, desconsiderando assim os reais objetivos clínicos de tais técnicas.

O treinamento empregando a técnica de subamostragem revelou-se menos enviesado em direção à acurácia e mais adequado do ponto de vista da revocação. O algoritmo

XGBoosting com subamostragem seguido de híper-parametrização atingiu taxas superiores de 80% de revocação, com acurácia total superior a 70% (Figura 7). Tais valores são significativos, sobretudo considerando-se um desempenho que resulta de dados tabulados obtidos por entrevistas telefônicas. Estratégias como esta mostram-se, portanto, potencialmente relevantes para o desenvolvimento de sistemas de prevenção de mais amplo uso, como por exemplo aqueles desenvolvidos no contexto de planejamento de políticas públicas de saúde.

O presente trabalho não se limitou, porém, à análise de desempenho dos classificadores. Através do índice de Gini pudemos quantificar quais características do banco de dados melhor contribuíram para os resultados do modelo. Esta análise demonstrou que, apesar de todas as características mostrarem-se significativas entre os grupos estudados (Tabela 3), o desempenho obtido pelo nosso melhor modelo ancorou-se, sobretudo, na idade, autopercepção do paciente em relação ao seu estado de saúde e dificuldades para caminhar, seguido de perto por existência de AVC prévio (Figura 11). Esses fatores, juntamente com outros, já foram comprovados como fatores de risco para o desenvolvimento de doenças cardiovasculares por estudos na literatura científica e pela Organização Mundial da Saúde (OMS) [16, 17].

A possibilidade de se interpretar modelos de aprendizado de máquina como exemplificado nesta análise tornam tais técnicas não apenas mais transparentes e confiáveis do ponto de vista médico como também podem servir para revelar variáveis de interesse no acompanhamento clínico.

Tendo em vista os resultados obtidos, foi possível concluir que modelos baseados em árvore de decisão como o XGBoosting possuem potencial para auxiliar na identificação do risco de infarto do miocárdio ou doença arterial coronariana a partir de dados tabulados transversais e coletados por questionário. Neste sentido, os aspectos metodológicos ilustrados e destacados neste trabalho, como o foco em métricas de avaliação de maior interesse clínico, o uso de técnicas adequadas de controle de desbalanceamento de classes e o emprego de estratégias de identificação de características relevantes poderão se mostrar essenciais à utilização de tais métodos no desenvolvimento de estratégias de baixo custo para a prevenção de doenças cardiovasculares.

AGRADECIMENTOS

Gostaria, primeiramente, de expressar minha gratidão a Deus, que sempre me abençoou com inúmeras oportunidades ao longo da minha trajetória acadêmica, desde o momento de ingresso até a conclusão do curso. Aos meus pais, que constantemente investiram em meu sonho de cursar Engenharia Biomédica, assim como à minha namorada, que esteve ao meu lado durante praticamente todo o meu percurso acadêmico. Expresso meu reconhecimento ao professor Adenauer Girardi Casali por sua orientação e contribuições significativas para meu interesse e aprendizado em algoritmos de aprendizado de máquina.

Agradeço também à Unifesp por toda a sua estrutura e aos profissionais comprometidos que trabalham para proporcionar a melhor experiência de aprendizado ao estudante. Por fim, dedico uma parte deste agradecimento a mim mesmo, pois sempre batalhei para alcançar meus objetivos, nunca deixei de sonhar, acreditar e manter minha determinação ao longo dessa jornada acadêmica.

REFERÊNCIAS

1) No Brasil, mais de 230 mil pessoas morreram por doenças cardiovasculares em 2021. ([s.d.]). Recuperado 25 de junho de 2023, de https://www.cnnbrasil.com.br/saude/no-brasil-mais-de-230-mil-pessoas-morreram-por-doencas-cardiovasculares-em-2021/amp/
2) Princy, R. J. P., Parthasarathy, S., Hency Jose, P. S., Raj Lakshminarayanan, A., & Jeganathan, S. (2020). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 570–575. https://doi.org/10.1109/ICICCS48265.2020.9121169
3) Heart Disease Facts | cdc.gov. (2023, maio 15). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm
4) Stevens, B., Pezzullo, L., Verdian, L., Tomlinson, J., George, A., & Bacal, F. (2018). The Economic Burden of Heart Conditions in Brazil. Arquivos Brasileiros de Cardiologia, 111, 29–36. https://doi.org/10.5935/abc.20180104
5) Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. IEEE Access, 9, 39707–39716. https://doi.org/10.1109/ACCESS.2021.3064084
6) Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., Reza, M. T., & Khan, M. R. H. (2022). A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. Scientific Programming, 2022, e3649406. https://doi.org/10.1155/2022/3649406
7) Indicators of Heart Disease (2022 UPDATE). ([s.d.]). Recuperado 9 de novembro de 2023, de https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
8) Toprak, M. (2020, agosto 29). Bayes' Theorem. Medium. https://medium.com/@toprak.mhmt/bayes-theorem-89daf9f11769
9) Chi-Square ($\chi 2$) Statistic: What It Is, Examples, How and When to Use the Test. ([s.d.]). Investopedia. Recuperado 31 de outubro de 2023, de https://www.investopedia.com/terms/c/chi-square-statistic.asp

10) Kotsiantis, S. B. (2013). Decision trees: A recent overview. Artificial Intelligence Review, 39(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4
11) Bashar, S. S., Miah, M. S., Karim, A. Z., Al Mahmud, M. A., & Hasan, Z. (2019, February). A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.
12) What is XGBoost? ([s.d.]). NVIDIA Data Science Glossary. Recuperado 25 de junho de 2023, de https://www.nvidia.com/en-us/glossary/data-science/xgboost/
13) Liu, A. Y. C. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets (Doctoral dissertation, University of Texas at Austin).
14) Sklearn.model_selection.RandomizedSearchCV. ([s.d.]). Scikit-Learn. Recuperado 9 de novembro de 2023, de https://scikit-learn/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
15) Kunumi. (2021, março 16). Métricas de Avaliação em Machine Learning: Classificação. Kunumi Blog. https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198
16) Mattila, K. J., Asikainen, S., Wolf, J., Jousimies-Somer, H., Valtonen, V., & Nieminen, M. (2000). Age, Dental Infections, and Coronary Heart Disease. Journal of Dental Research, 79(2), 756–760. https://doi.org/10.1177/00220345000790020901
17) Doenças cardiovasculares—OPAS/OMS | Organização Pan-Americana da Saúde. ([s.d.]). Recuperado 9 de novembro de 2023, de https://www.paho.org/pt/topicos/doencas-cardiovasculares
18) Lopes, L. (2023, junho 12). Árvore de Decisão em ML: Guia completo. Medium. https://medium.com/@lorranloopes13/%C3%A1rvore-de-decis%C3%A3o-em-ml-guia-completo-abd288c7ec0b
19) Dados Desbalanceados—O que são e como lidar com eles | by Felipe Azank | Turing Talks | Medium. ([s.d.]). Recuperado 10 de novembro de 2023, de https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b
20) XGBoost versus Random Forest | Qwak. ([s.d.]). Recuperado 10 de novembro de 2023, de https://www.qwak.com/post/xgboost-versus-random-forest