**Estudo comparativo de algoritmos de árvore de decisão para a identificação de risco de infarto ou doença arterial coronariana**

Pedro Henrique Crisp Modesto1, Adenauer Girardi Casali2

1 Instituto de Ciência e Tecnologia da Unifesp / Engenharia Biomédica

*Resumo* **— Doenças cardiovasculares são as principais causas de morte no Brasil e no mundo, e 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 trabalhos têm buscado empregar estratégias de aprendizado de máquina para aprimorar o reconhecimento automatizado de pacientes em risco de IM e DAC. Algoritmos baseados em árvores tem se mostrado bastante eficientes para a classificação do risco do paciente, logo, este projeto tem como objetivo testar diferentes algoritmos baseados em árvores com diferentes metodologias de treinamento do modelo para avaliar o melhor empregado.**

*Abstract* **— Cardiovascular diseases are the leading causes of death in Brazil and worldwide. Myocardial infarction (MI) and coronary artery disease (CAD) are among the most costly conditions for the public healthcare system (SUS). Preventing these conditions relies on examinations such as angiography, which are expensive and often inaccessible to individuals with low income. In this context, several studies have sought to employ machine learning strategies to enhance the automated recognition of patients at risk of MI and CAD. Tree-based algorithms have proven to be highly effective in patient risk classification. Therefore, this project aims to test different tree-based algorithms with various model training methodologies to assess the most effective approach.**

*Keywords***— Classificação, Aprendizado de Máquina, Doenças Cardíacas**

# 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), acidente vascular cerebral (AVC) e infarto do miocárdio (IM). Embora haja diversos estudos que mostram uma grande associação dessas doenças com fatores genéticos, hábitos alimentares, raça e outros fatores de risco, nem sempre é fácil detectar clinicamente pacientes com maior risco devido à ampla quantidade de diagnósticos diferenciais [2]. Além disso, de acordo com o Centro de Prevenção e Controle de Doenças dos EUA, estima-se que um em cada cinco ataques cardíacos seja assintomático, ou seja, a pessoa não está ciente do ocorrido até que o dano seja 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 impede o acesso a população mais pobre, logo, é necessário buscar métodos mais econômicos e acessíveis para o auxílio do diagnóstico médico, como por exemplo, o uso do aprendizado de máquina para detectar padrões e reconhecer pacientes com maior risco de acontecer um ataque cardíaco [4,5].

É possível encontrar na literatura diversos estudos que utilizam diferente técnicas e modelagens de dados clínicos para classificar o risco de o paciente desenvolver IM ou DAC. Os algoritmos de aprendizado de máquina baseado em árvores têm mostrado ser bastante eficientes na identificação de pacientes com risco de desenvolver infarto ou doenças cardíacas em cima de dados tabulados de condições clínicas.

Em Princy et al. (2020), foram testados diferentes algoritmos de aprendizado de máquina em dados clínicos para classificar o risco de o paciente desenvolver doenças cardíacas, onde o modelo que apresentou a melhor performance foi o de árvore de decisão com uma acurácia de 73% [2].Assim como Ishaq et al. (2021), utilizou diferentes modelos associados a técnica de balanceamento de classes, SMOTE, onde o algoritmo *Extra Tree Classifier* alcançou uma acurácia de 92,6%, 93% de precisão e 93% de F1-Score [5]. E em Muntasir Nishat et al. (2022), que obteve uma performance de 90% de acurácia utilizando o modelo *Random Forest* associado com o balanceamento de classes por SMOTE-ENN [6].

Nesse contexto, o objetivo deste trabalho é comparar os algoritmos de aprendizado de máquina convencionais baseados em árvores, como *Random Forest* e *Decision Tree*, com os algoritmos mais recentes baseados em árvores, como o *XGBoosting*. Pretende-se avaliar se o desempenho e a capacidade de ajuste dos dados são aprimorados nesse algoritmo mais atualizado. Ademais, busca-se identificar a técnica de aprendizado de máquina mais eficaz como uma ferramenta de apoio aos médicos no processo diagnóstico.

# MATERIAL E MÉTODOS

Foi utilizado 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 (CPCD). Anualmente, o centro coleta dados por telefone do 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:

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, foi executada 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 o preenchimento das variáveis. Além disso, foi aplicado o Teorema de Bayes (TB) para comparação das características com a característica alvo (HeartDisease). O TB é uma fórmula estatística capaz de calcular probabilidade condicional entre dois eventos, A e B, serem verdadeiros dado que B é verdadeiro, ou seja, para este estudo foi aplicado para determinar a probabilidade de uma pessoa satisfazer uma condição, levando em consideração se ela teve ou não um infarto. [8]

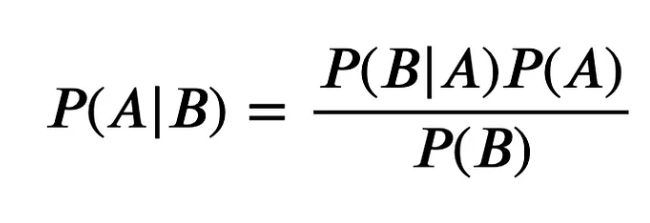


Figura 1: Equação do Teorema de Bayes

Para avaliar se as probabilidades determinadas pelo teste t eram estatisticamente significativas, o teste qui-quadrado (X²) foi empregado. O teste X² é amplamente utilizado para examinar hipóteses entre variáveis independentes e as frequências observadas e esperadas entre grupos [9]. Assim, ele possibilita a avaliação de características binárias presentes nos conjuntos de dados.

Foi aplicado três algoritmos de árvores distintos em conjunto com três metodologias de balanceamento de classes, seguido da hiper parametrização de cada modelo, a fim de determinar o método mais eficaz. O conjunto de dados foi dividido, utilizando uma proporção de 75% para o treinamento após o balanceamento, deixando 25% para a fase de teste. As técnicas de balanceamento utilizadas incluíram sobre amostragem, sub amostragem e uma combinação de ambas. 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.

1. **Algoritmos de Aprendizado de Máquina:**

***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 calculado por 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, onde o classificador utiliza as características de maneira aleatória para criar N árvores, logo, o modelo mescla diferentes árvores aleatórias independentes, em que a classificação é definida de acordo com procedimento de votação, onde seleciona 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. Onde as árvores são feitas de uma maneira paralela para corrigir erros da anterior, desse modo, gradualmente encontra um modelo que minimiza erros de classificação de maneira eficiente e otimizada [12].

1. **Técnicas de reamostragem dos dados utilizadas:**

***Oversampling:*** 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].

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

***Combinação entre Oversampling + Undersampling:*** Para equilibrar as distribuições das classes, evitar perder dados relevantes e complicações de *overfitting,* foi utilizado a técnica de SMOTE-ENN (Synthetic Minority Oversampling Technique and Edited Nearest Neighbor), que combina as técnicas de oversampling e undersampling até atingir um equilíbrio entre as classes [6].

1. **Hiper parametrização dos algoritmos:**

Com a finalidade de atingir a melhor performance do algoritmo classificador aplicado no conjunto de dados trabalhados, foi utilizado o algoritmo *RandomSearchCV*, com o objetivo de testar diferentes parâmetros de maneira aleatória para otimizar o modelo no ajuste dos dados de variáveis discretas [14].

1. **Métricas de avaliação de performance:**

Para a avaliação do modelo do modelo, foram utilizadas as seguintes métricas de avaliação: matriz de confusão, acurácia, precisão, revocação, *f1-score*, especificidade, erro tipo 1 e erro tipo2. As fórmulas são calculadas a partir dos Verdadeiro Positivos (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) [15].

Tabela 2: Métrica de Avaliação do algoritmo classificador

|  |  |
| --- | --- |
| Métrica | Fórmula |
| Acurácia |  |
| Precisão |  |
| Revocação |  |
| F1 - Score |  |

1. **Materiais:**

Todo o desenvolvimento do projeto foi executado em conjunto com o software 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

1. RESULTADOS

Após a extração dos dados, durante uma etapa de análise, foi possível observar que a proporção entre as pessoas que participaram da pesquisa CPCD era de 91,4% não tiveram diagnóstico de DAC ou IM, enquanto 8,6% tiveram.

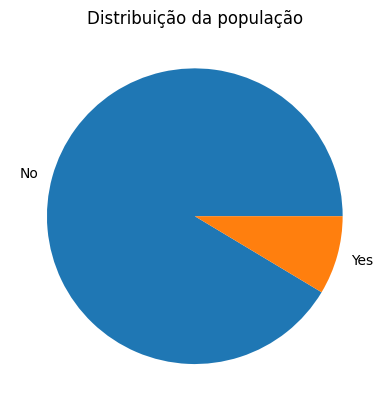


Figura 2: Distribuição da população em relação a variável alvo

Nesta etapa, também foi possível identificar *outliers* em algumas características que foram removidas. Observando a distribuição de valores de *BMI* e *SleepTime*, foi notado a presença de valores que divergem significativamente de padrões esperados para adultos. Portanto, foram removidos os seguintes valores:

1. Quando o índice de massa corporal é inferior a 14 kg/m2 e superior a 60 kg/m2;
2. Quando a média de horas dormidas por dia é inferior a 4 horas e superior a 14 horas.

Ao aplicar o Teorema de Bayes para determinar a probabilidade de uma condição ser satisfeita com base no evento da característica alvo, juntamente com o teste qui-quadrado para verificar a significância das diferenças, os resultados obtidos foram:

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 | |

A relação entre as condições que mais se mostraram diferentes nos dois grupos foram: *DiffWalking = Yes*, AgeCategory > 40 anos e *Diabetic = Yes*.

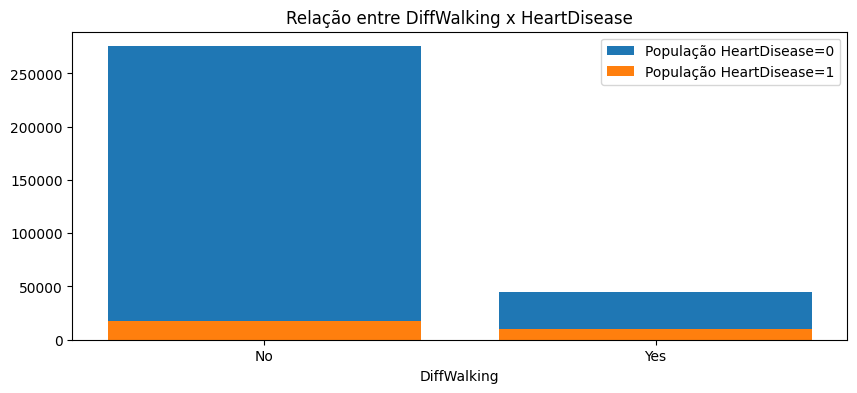


Figura 3: Relação entre a variável de dificuldade de andar com pessoas com DAC e IM.

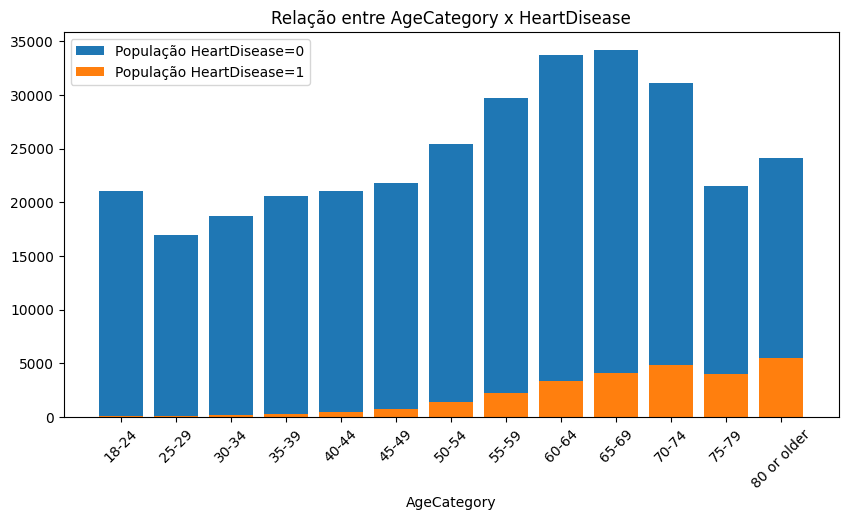


Figura 4: Relação entre a variável de faixa etária com pessoas com DAC e IM

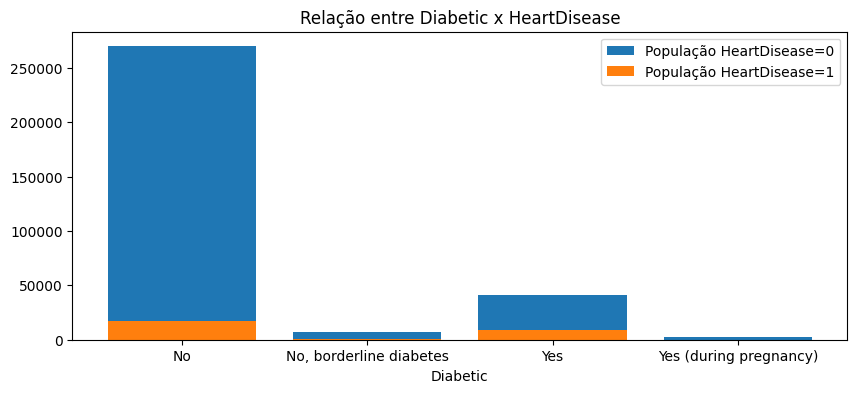


Figura 5: Relação entre a variável de identificação de diabéticos com pessoas com DAC e IM

Os resultados de performance dos modelos testados associados com as diferentes metodologias de balanceamento de classes foram:

1. *Oversampling*:

Tabela 3: Resultados dos modelos com sobre amostragem

|  |  |  |  |
| --- | --- | --- | --- |
| 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% |

1. *Undersampling*:

Tabela 4: Resultados dos modelos com sub amostragem

|  |  |  |  |
| --- | --- | --- | --- |
| 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% |

1. *Oversampling* + *Undersampling*:

Tabela 5: Resultados dos modelos com a combinação de métodos

|  |  |  |  |
| --- | --- | --- | --- |
| 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% |

A matriz de confusão dos dados previstos pelo modelo XGB com o balanceamento de classe combinado por sobre amostragem e sub amostragem foram:

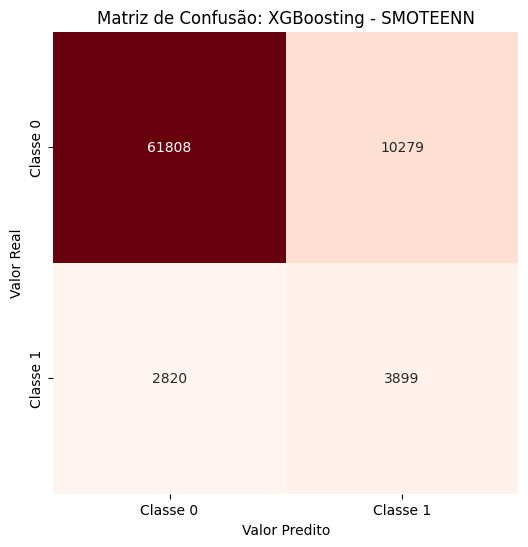


Figura 6: Matriz de confusão dos resultados preditos pelo modelo *XGBoosting*

As métricas obtidas a partir do modelo *XGBoosting* hiper parametrizado, combinado com os métodos de balanceamento de classes, foram as seguintes:

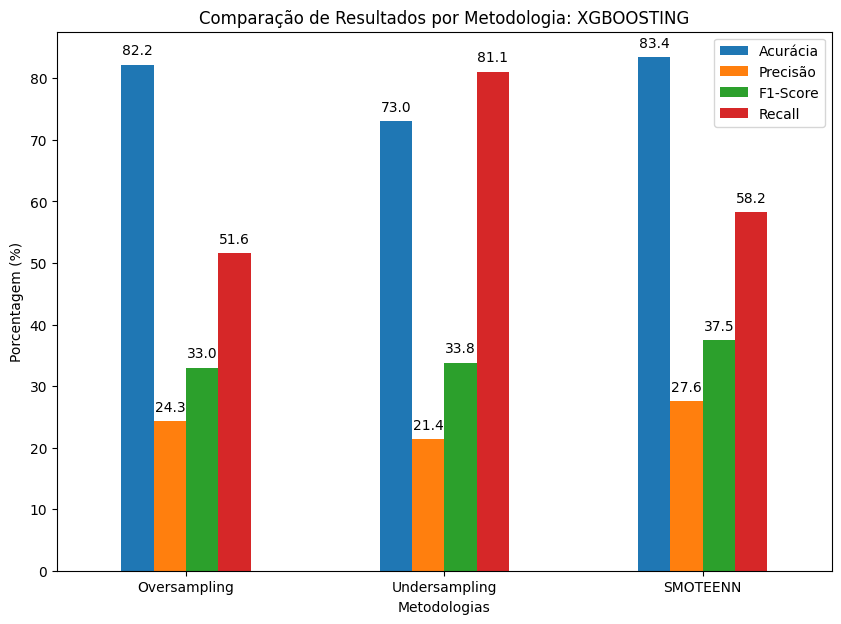


Figura 7: Resultados do modelo *XGBoosting* hiper parametrizado

A matriz de confusão dos dados previstos do modelo hiper parametrizado com a combinação de balanceamento de classes:

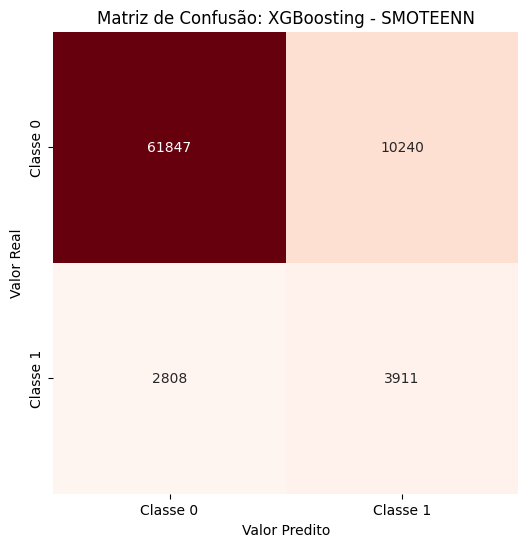


Figura 8: Matriz de confusão dos resultados preditos pelo modelo hiper parametrizado *XGBoosting*

A partir do modelo treinado, foi possível identificar as características mais importantes, calculadas pelo índice de Gini, para classificar uma pessoa com o risco de desenvolver DAC ou IM.

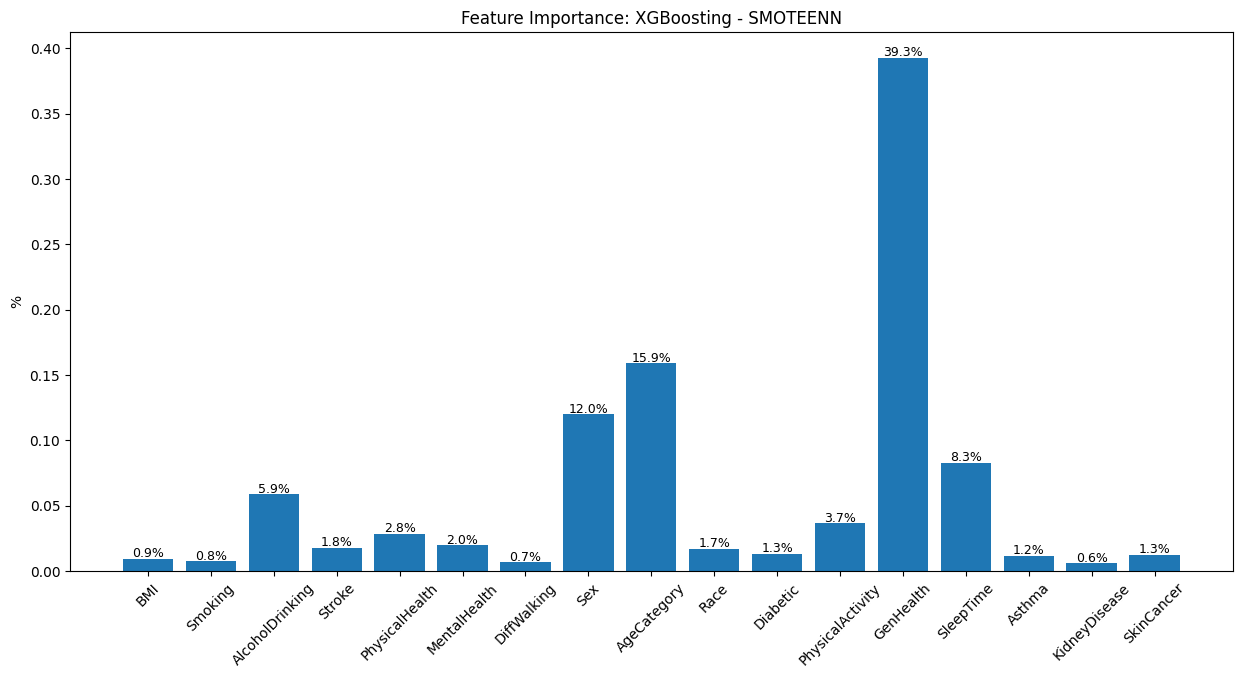


Figura 9: Características mais importantes que o modelo levou em consideração para a classificação

1. DISCUSSÃO E CONCLUSÃO

A partir das probabilidades calculadas pelo TB, foi possível concluir que várias condições de saúde estão associadas a pessoas que tiveram doenças cardíacas. Além disso, todas as probabilidades apresentaram diferenças significativas quando o teste do qui-quadrado foi aplicado. Como mencionado anteriormente, a faixa etária, a presença de diabetes e a dificuldade de locomoção, incluindo subir escadas, demonstraram as maiores diferenças entre o grupo que sofreu um infarto do miocárdio (IM) e doença arterial coronariana (DAC), em comparação com o grupo que não teve essas condições. Também é importante ressaltar que a identificação de fumantes revelou diferenças significativas entre os grupos.

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].

Ao considerar os resultados dos modelos juntamente com as metodologias de balanceamento, pode-se concluir que as métricas obtidas foram satisfatórias. Utilizando a técnica de sobre amostragem, os modelos demonstraram uma acurácia superior a 80%. No entanto, a precisão, que reflete a capacidade do modelo de não classificar exemplos negativos como casos de DAC e IM positivos, não foi tão alta, resultando em um percentual considerável de falsos positivos (FP). Em relação à revocação, que indica a habilidade do modelo de identificar todos os exemplos positivos, foi observado um desempenho mediano, com destaque para o algoritmo *XGBoosting*, que se sobressaiu nesse aspecto [15].

A utilização da metodologia de sub amostragem resultou em um impacto negativo na acurácia e na precisão do modelo. O desempenho na classificação de exemplos negativos foi um pouco comprometido, embora tenha havido uma melhoria na identificação dos exemplos positivos. Apesar de aumentar a revocação, ou seja, a precisão na identificação dos casos positivos [15], essa abordagem também resultou em um aumento nas falsas sugestões de risco para pacientes que na realidade não o apresentam.

Quanto à combinação das técnicas de sobre amostragem e sub amostragem, os modelos demonstraram um desempenho mais equilibrado. As acurácias dos modelos DT e RF não superaram aquelas obtidas apenas com a sobre amostragem, e as métricas de revocação não ultrapassaram as obtidas com a sub amostragem. No entanto, ao considerar o *f1-score,* que combina os valores de precisão e revocação em uma única métrica [15], todos os modelos alcançaram valores mais altos.

De maneira geral, o modelo *XGBoosting* destacou-se em quase todas as métricas, independentemente da metodologia de balanceamento aplicada. Portanto, esse novo algoritmo baseado em árvores revelou-se um forte candidato para a aplicação na classificação de riscos de doenças cardíacas em dados tabulados. Quanto à metodologia de balanceamento de classes, a combinação de sobre amostragem com sub amostragem, utilizando o algoritmo SMOTE-ENN, mostrou-se a mais equilibrada, apresentando as melhores métricas. Essa abordagem não resultou na perda de padrões importantes nem levou ao *overfitting* do modelo [6].

Avaliando a matriz de confusão do Modelo XGB com o SMOTE-ENN foi possível verificar que o modelo obteve 85,7% de VN, 14,3% de FP, 58% de VP e 42% de FN. Tendo sua sensibilidade e especificidade maior que seu erro tipo 1 e erro tipo 2.

Embora as métricas do algoritmo XGB tenham sido boas, com o objetivo de otimizar os resultados, utilizou-se a biblioteca RandomSearchCV para testar diversas combinações de parâmetros do algoritmo, a fim de garantir o melhor ajuste aos dados. Onde os novos parâmetros conseguiram melhorar o *f1-score* e a revocação de todas as formas de balanceamento, sendo a combinação de *oversampling* + *undersampling* ainda a mais equilibrada. Com base nos resultados da matriz de confusão do modelo com a melhor metodologia foram: 85,8% de VN, 14,2% de FP, 58,2% de VP e 41,8% de FN.

Por meio do cálculo do índice de Gini, foi possível identificar as características mais relevantes no modelo de XGB em conjunto com a combinação de amostragens. Esse algoritmo avalia cada característica para determinar a "pureza" na classificação dos grupos que apresentaram ou não doenças cardíacas, ou seja, as características mais eficazes na distinção entre os grupos [18]. Logo as características mais relevantes para este modelo foram as seguintes: GenHealth (39,3%), que consiste na autoavaliação da saúde pelo paciente, AgeCategory (15,9%), representando a faixa etária do paciente, Sex (12,0%), indicando o sexo masculino ou feminino, e SleepTime (8,3%), que é a média de horas de sono diárias.

Tendo em vista os resultados dos modelos, foi possível concluir que o uso de algoritmos de aprendizado de máquina pode representar ferramentas de extrema importância para auxiliar os médicos na identificação do risco de um paciente desenvolver infarto do miocárdio ou doença arterial coronariana, considerando certos fatores, como as características mais importantes, que atuam como indicadores de possíveis alertas.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, que sempre me abençoou com diversas oportunidades ao longo da minha trajetória acadêmica, desde a oportunidade de ingresso até na conclusão do curso. Aos meus pais, que sempre investiram e me apoiaram no meu sonho de cursar Engenharia Biomédica, a minha namorada, que esteve junto comigo durante quase todo meu período acadêmico me apoiando, me ajudando e sido minha companhia. Ao professor Adenauer Girardi Casali, por ter me orientado e contribuído no meu interesse e aprendizado do uso de algoritmos de aprendizado de máquina. A Unifesp por toda sua estrutura e profissionais que trabalham para o melhor aprendizado ao estudante. E a mim mesmo, que sempre batalhou para alcançar os objetivos, que nunca deixou de sonhar, acreditar e ser determinado.

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 (χ2) 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