Aplicação de Estratégias de Mineração de Dados para Identificar Indivíduos Hipertensos com Doenças Cardiovasculares e Saudáveis no Brasil

Gustavo Costa

G. Costa

Pontifícia Universidade Católica de Minas Gerais

Belo Horizonte, Minas Gerais

Resumo

Este estudo explorou a aplicação de estratégias de mineração de dados para identificar indivíduos hipertensos com doenças cardiovasculares (HA + DCV) e saudáveis/sem diagnóstico no Brasil, utilizando a base de dados da Pesquisa Nacional de Saúde de 2019. Foram testados diferentes algoritmos de aprendizado de máquina, como Árvore de Decisão, Floresta Aleatória e Naive Bayes, com o objetivo de classificar as condições de saúde das pessoas.

Os resultados mostraram que os modelos apresentaram um desempenho semelhante, com destaque para a Árvore de Decisão e a Floresta Aleatória, que alcançaram 97% de precisão e sensibilidade para identificar corretamente os saudáveis. No entanto, a classificação dos indivíduos com HA + DCV foi mais desafiadora, com a sensibilidade apresentando valores mais baixos, o que pode ser explicado pela base de dados descritiva utilizada, que não forneceu diagnósticos médicos formais. Além disso, fatores como mudanças no estilo de vida e a falta de diagnóstico precoce podem ter influenciado o desempenho dos modelos.

O estudo evidenciou a importância de incorporar dados mais detalhados e longitudinais para aprimorar a precisão dos modelos e permitir uma identificação mais eficaz das pessoas com doenças crônicas.

Palavras-chave: mineração de dados, pré-processamento, aprendizado de máquina, hipertensão, doenças cardiovasculares.

1. Introdução

As Doenças Cardiovasculares (DCV) são, atualmente, configuram-se como a principal causa de morte ao redor do mundo. Durante o ano de 2008, estima-se que essas doenças causaram 17.3

milhões de mortes sendo 7.3 milhões por ataques cardíacos e 6.2 milhões por derrames. A Organização Mundial da Saúde (OMS) projeta que, até 2030, mais de 23 milhões de pessoas morrerão dessas doenças que afetam o sistema cardiovascular [1].

A Hipertensão Arterial (HA) é também uma doença crônica não transmissível e também afeta negativamente o sistema cardíaco do indivíduo, sendo um dos principais problemas de saúde pública no mundo inteiro. A OMS estima que há cerca de 600 milhões de pessoas que possuem HA com um crescimento global de 60% dos casos até 2025, além de um número de 7.1 milhões de mortes por ano [2]. No Brasil, dados da Pesquisa Nacional de Saúde (PNS) de 2019 revelam que 23,9% dos adultos reportaram diagnóstico médico de (HA) [3] o que indica que há uma necessidade de precaver a progressão dessa condição e suas complicações, uma vez que ela é um dos principais fatores de risco para doenças cardiovasculares.

A relevância do estudo da HA e das DCV decorre não apenas da alta prevalência dessas doenças na população, mas também do impacto significativo dessa condição na qualidade de vida, nas taxas de mortalidade e nos custos financeiros associados ao seu manejo. A HA é um fator de risco central para as DCV, que representam uma carga econômica representativa aos sistemas de saúde devido a hospitalizações, tratamentos prolongados e complicações evitáveis.

No Brasil, as DCV, frequentemente associadas à HA, geraram um custo de aproximadamente R$ 50 bilhões entre 2010 e 2020, considerando gastos diretos com hospitalizações e tratamentos pelo Sistema Único de Saúde (SUS). Além disso, estima-se que custos indiretos, como perda de produtividade e mortalidade prematura, aumentem ainda mais essa carga econômica no país [4, 5].

A importância do estudo desse tema está no impacto significativo da hipertensão e das doenças cardiovasculares tanto na saúde pública quanto nos custos econômicos no Brasil. Prevenir e diagnosticar precocemente essas condições é fundamental para reduzir a mortalidade e as complicações associadas por meio de investimentos públicos, adoção de estilos de vida mais saudáveis e o uso de medicamentos. Estudos recentes têm explorado o uso de mineração de dados e algoritmos de machine learning para melhorar o diagnóstico e a predição dessas doenças. Por exemplo, o artigo de Latifa A. AlKaabi e colaboradores [6] utilizou dados de 987 pessoas do Qatar Biobank para aplicar técnicas como regressão logística e árvores de decisão para prever hipertensão com dados não invasivos, demonstrando potencial para reduzir custos e otimizar a triagem de risco em populações vulneráveis.

Além disso, outro trabalho que explora essa aplicação da mineração de dados e do uso de algoritmos de machine learning na predição de doenças cardiovasculares é apresentado no estudo [7]. Este trabalho utilizou um conjunto de dados real contendo 70.000 instâncias para desenvolver modelos preditivos que classificam a ocorrência de doenças cardiovasculares. Os pesquisadores empregaram algoritmos como Random Forest, Decision Tree, XGBoost e Multilayer Perceptron (MLP), fatores de risco como dieta inadequada, obesidade e tabagismo foram identificados como variáveis relevantes para a predição, evidenciando como abordagens baseadas em dados podem apoiar sistemas de triagem.

Portanto, o objetivo desse trabalho é a exploração da base de dados PNS 2019 (Pesquisa Nacional de Saúde) para aprender quais são os fatores que descrevem o perfil de uma pessoa com hipertensão e alguma doença cardiovascular e as que não possuem nenhuma dessas duas comorbidades. Para atingir esse objetivo, foi aplicado um processo de descoberta de conhecimento em base de dados e, somado à isso, foi aplicado as técnicas de aprendizado de máquina como floresta aleatória e árvore de decisão para explicar, por meio dos resultados obtidos, os perfis em análise.

2. Trabalhos Relacionados

Nesta seção, apresentamos estudos que utilizam técnicas de aprendizado de máquina para predição de hipertensão e doenças cardiovasculares. Foram analisados os objetivos, métodos, resultados e principais insights de cada um dos trabalhos.

O trabalho de L. A. AlKaabi et al in [6], analisou 987 participantes acima de 18 anos e utilizou algoritmos como regressão logística e árvores de decisão para prever hipertensão. Identificou fatores de risco como idade, IMC, nível de educação, atividades físicas, obesidade, tabagismo, alcoolismo e consumo de sal. O modelo demonstrou boa aplicabilidade para triagem de risco com dados não invasivos, reduzindo custos em sistemas de saúde .

O artigo dos autores C. M. Bhatt et al in [7], investiga o uso de técnicas de aprendizado de máquina e aprendizado profundo para prever doenças cardíacas, com foco no desempenho do modelo Multilayer Perceptron (MLP). Utilizando uma base de dados do Kaggle sobre doenças cardiovasculares, com 70.000 instâncias, o estudo comparou vários modelos, incluindo Decision Trees, XGBoost, Random Forest e MLP, alcançando precisão superior a 86% para todos os modelos. O MLP, especificamente, demonstrou a maior acurácia, atingindo 87,28% com validação cruzada. Os autores também ressaltaram a falta de dados sobre fatores relacionados à essas doenças que poderiam ampliar os resultados dos modelos e também ressaltaram que poderiam ter utilizado de técnicas mais robustas de imputação de dados ausentes para melhorar a qualidade das informações.

De acordo com o autor N. M. de Carvalho et al in [8], eles realizaram um estudo aplicando técnicas de mineração de dados à Pesquisa Nacional de Saúde 2019, com o objetivo de identificar padrões relacionados ao diagnóstico de hipertensão. Os métodos incluíram o uso de algoritmos de classificação como Random Forest para analisar variáveis sociodemográficas e comportamentais. O estudo destacou a importância do tabagismo e do IMC como fatores de risco e obteve um F1-Score médio de 75%, trazendo como fatores de risco para a hipertensão a ingestão excessiva de sal, sobrepeso, a não prática de atividades físicas e o tabagismo.

Este estudo de Anna Karen Gárate-Escamila et al [9] explorou a predição de doenças cardíacas com a aplicação de técnicas de redução de dimensionalidade, combinando análise qui-quadrado (CHI) e análise de componentes principais (PCA). O objetivo foi encontrar os atributos mais relevantes para a predição e melhorar a eficiência dos modelos de aprendizado de máquina. Utilizando o conjunto de dados do UCI Heart Disease, composto por 74 características, os

autores implementaram seis classificadores de aprendizado de máquina, incluindo Random Forest (RF), que apresentou o melhor desempenho. Os resultados indicaram que o método CHI-PCA, aliado ao RF, alcançou precisão de 98,7% no conjunto Cleveland, 99,0% no conjunto Hungarian e 99,4% no conjunto combinado Cleveland-Hungarian. Entre os atributos mais significativos estavam colesterol, frequência cardíaca máxima, dor no peito e características relacionadas à depressão do segmento ST e vasos cardíacos.

De acordo com A. Saboor [10] et al in, o estudo investigou a aplicação de nove algoritmos de aprendizado de máquina para a predição de doenças cardíacas, utilizando um conjunto de dados padrão de doenças do coração. O objetivo foi melhorar a precisão na identificação precoce de doenças cardíacas, reduzindo os possíveis erros de abordagens manuais. Os autores realizaram etapas de pré-processamento e padronização dos dados, além de ajustar os hiperparâmetros dos modelos, incluindo algoritmos como SVM, Random Forest, e XGBoost. Os experimentos usaram a validação cruzada para avaliar os modelos. O algoritmo SVM destacou-se com uma precisão de 96,72% após ajustar os hiperparâmetros e a padronização dos dados. O estudo também revelou limitações, como uma eficiência menor em datasets muito grandes, e propôs melhorias futuras, incluindo o uso de XGBoost para predição. Os resultados reforçam a importância da padronização e seleção eficiente de atributos para melhorar o desempenho dos classificadores. Ademais, a pesquisa enfatiza que técnicas avançadas de aprendizado de máquina podem minimizar custos de diagnóstico e contribuir para sistemas de triagem mais eficientes.

No trabalho de A. D'Souza et al in [11], os autores propuseram o uso de técnicas de mineração de dados para a predição de doenças cardíacas, aplicando métodos como Redes Neurais Artificiais (ANN), K-Means Clustering e Geração de Itemsets Frequentes. O objetivo foi identificar padrões e relações entre fatores, como idade, sexo, pressão arterial e níveis de glicose, para prever a probabilidade de um paciente desenvolver doenças cardíacas. Os métodos foram avaliados com base em três métricas de desempenho: Recall, Precision e Acurácia. Os resultados indicaram que as ANN superaram o K-Means Clustering em todos os parâmetros, mostrando melhor desempenho em termos de precisão e capacidade de detectar casos verdadeiros. O estudo também revelou que, embora as técnicas de mineração de dados possam proporcionar novos conhecimentos para a área da saúde, o projeto sofreu com limitações devido à falta de conjuntos de dados atualizados e completos, o que pode afetar a eficácia dos modelos. O trabalho sugere que a coleta de dados reais de organizações de saúde poderia melhorar significativamente os resultados do sistema preditivo.

De acordo com Gangadhar et al in [12], investiga como técnicas de aprendizado profundo (Deep Learning) e aprendizado de máquina podem ser aplicadas para prever doenças cardíacas coronárias em estágio inicial. O estudo utiliza redes neurais artificiais (ANNs) como modelo principal, destacando que estas obtiveram a maior precisão (84,44%) em comparação com outros algoritmos, como SVM (83,33%), Random Forest (81,67%), Decision Tree (73,33%) e KNN (61,67%). Ademais, os autores também trazem a importância das métricas de Recall,

Sensibilidade e F1-Score como resultados fundamentais para entender o desempenho desses modelos no problema abordado.

Kavitha et al. [13] apresentaram um modelo híbrido que combina uma árvore de decisão para selecionar as características mais relevantes com um algoritmo de random forest, utilizado para agregar essas informações e calcular probabilidades. Essa abordagem permite que o modelo seja treinado e testado de forma eficaz, resultando em previsões finais mais precisas.

Ayatollahi [14] destaca que, utilizando técnicas de mineração de dados, foram selecionados 25 fatores relevantes relacionados à Doença Arterial Coronariana (CAD). Esses fatores serviram como variáveis de entrada para um modelo de Redes Neurais Artificiais (ANN), permitindo que seus pesos fossem calculados e analisados. Entre as variáveis mais significativas identificadas estão o nível de glicose no sangue, o colesterol, a presença de desconforto no peito, o hábito de fumar e o gênero. Esse método possibilitou uma abordagem mais estruturada para avaliar os principais indicadores associados à CAD.

O estudo dos pesquisadores S. Kaur et al in [15] explora a aplicação de aprendizado de máquina na predição de hipertensão, utilizando dados de pacientes, como idade, IMC, histórico familiar e hábitos de vida. A pesquisa mostra como algoritmos podem identificar fatores de risco relevantes, ajudando na detecção precoce e no gerenciamento da doença. Embora os resultados sejam promissores, o estudo ressalta a necessidade de validação em ambientes clínicos e a atualização contínua dos modelos com novos dados. Limitações como a dependência da qualidade dos dados e a possibilidade de erros de classificação também são discutidas.

3. Metodologia

No âmbito da metodologia aplicada, o estudo proposto seguiu uma série de passos descritos no Fluxograma contido na figura 1.

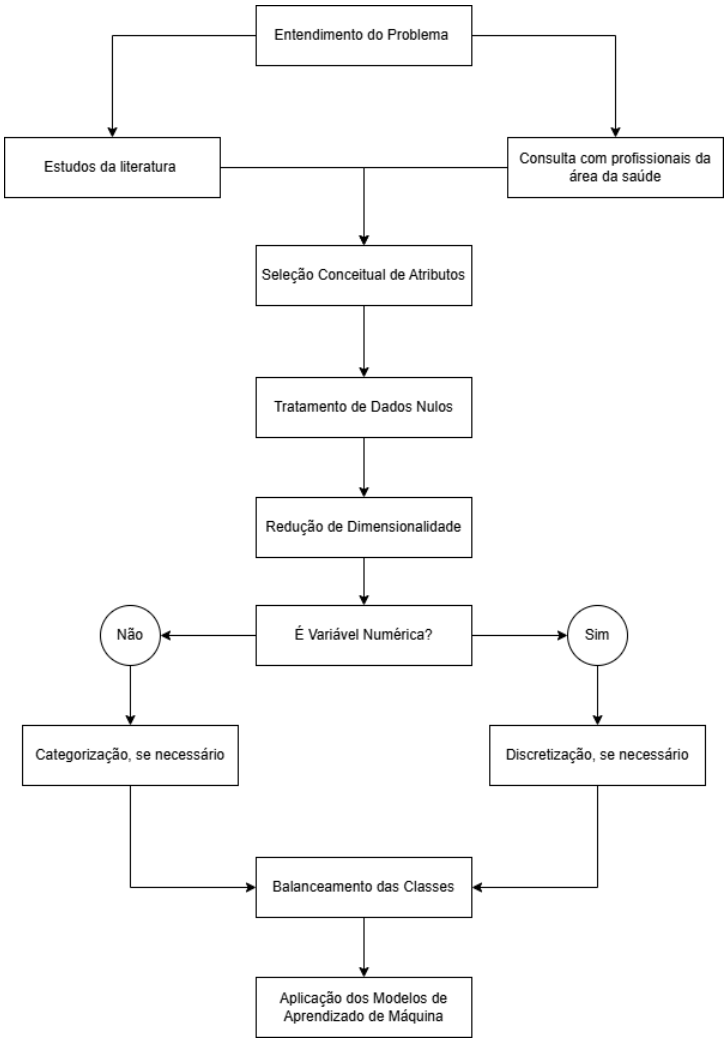

Figura 1: Fluxograma

3.1. Materiais

A Pesquisa Nacional de Saúde (PNS) de 2019 utilizada como fonte de dados neste estudo, é realizada pelo IBGE em parceria com o Ministério da Saúde, é uma importante fonte de dados sobre as condições de saúde da população brasileira. Com o objetivo de fornecer informações detalhadas sobre o perfil de saúde da população, ela coleta dados sobre doenças crônicas, condições de vida, hábitos de saúde, e acesso aos serviços de saúde, entre outros. A PNS possui uma amostra representativa de todos os estados brasileiros, com informações sobre aspectos sociodemográficos, estilos de vida, e fatores de risco, como tabagismo e obesidade.Somado à isso, a PNS possui 293.726 registros e 1.088 atributos, esses atributos estão subdivididos em 26 módulos diferentes de questões. Além de auxiliar em políticas públicas brasileiras, essa base de dados é material de estudos para inúmeros projetos estudantis e científicos.

(imagens aqui sobre a PNS)

3.2 Métodos

Parte 1 - Entendimento do Problema: Nessa etapa, o foco é entender o domínio do problema proposto para o estudo, em outras palavras, significa compreender o contexto, mapear os fatores mais importantes para o problema, as possíveis soluções e o que elas trazem como

consequência para a sociedade por meio dos seus possíveis benefícios. Esse processo envolve a revisão da literatura e a consulta com especialistas da área, nesse caso médicos, buscando uma compreensão clara dos aspectos que influenciam no problema em questão. Para isso, também foi utilizado do método CAPTO [16] que consiste em construir um modelo conceitual com a utilização de conhecimento explicíto, obtido por meio de documentos, artigos e textos e com o uso do conhecimento tácito obtido por meio da comunicação com profissionais especialistas no tema abordado. Com o conhecimento obtido, o próximo passo desse método é construir dimensões, cujo cada dimensão pode vir à possuir muitos aspectos que detalham essas dimensões. Na figura 2 é apresentada o mapa conceitual construído para Hipertensão Arterial e Doenças Cardiovasculares após a aplicação do método CAPTO.

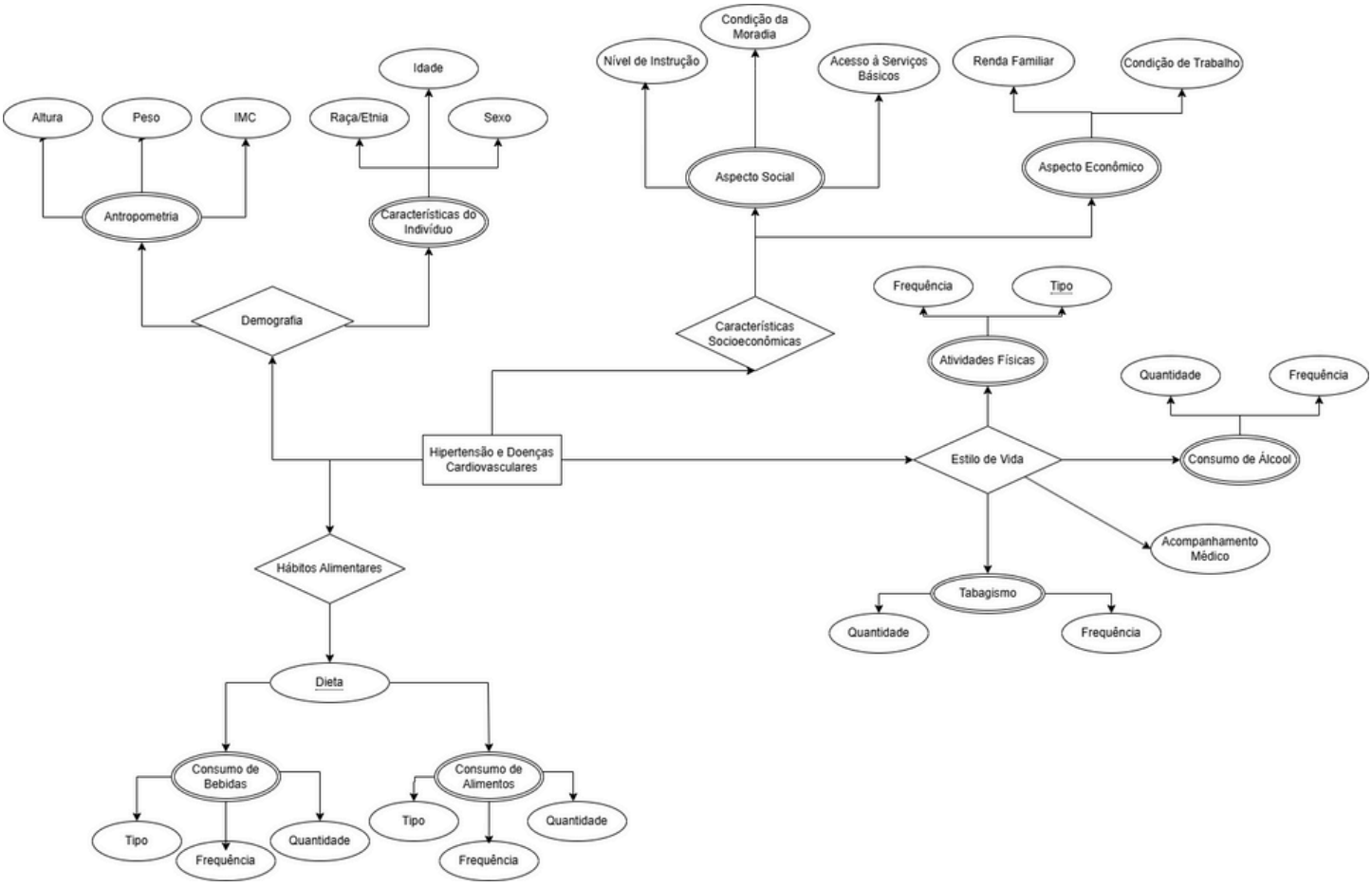

Figura 2: Mapa Conceitual - Método CAPTO

Parte 2 - Seleção Conceitual de Atributos: Após a montagem do mapa conceitual utilizando o método CAPTO, essa etapa consiste em selecionar conceitualmente os atributos escolhidos no mapa dentro da base de dados PNS 2019. Confira na tabela 1 os atributos selecionados:

Atributo	Descrição do Atributo
P00104	Peso em kg do indivíduo.
P00404	Altura em cm do indivíduo.
C006	Sexo do indivíduo.
C008	Idade do indivíduo.
C009	Raça/etnia do indivíduo.
P02601	Consumo de sal do indivíduo.
P034	Nos últimos três meses praticou algum tipo de exercício físico ou esporte?
P035	Quantos dias por semana costumava praticar exercício físico ou esporte?
P050	Atualmente fuma algum produto do tabaco?
P052	No passado, fumou algum produto do tabaco?
P05401	Quanto fuma por dia ou por semana de Cigarros industrializados
P05404	Quanto fuma por dia ou por semana de Cigarros de palha
P05407	Quanto fuma por dia ou por semana de Cigarros de cravo
P05410	Quanto fuma por dia ou por semana de Cachimbos cheios
P05413	Quanto fuma por dia ou por semana de Charutos
P05416	Quanto fuma por dia ou por semana de Narguilé?
P05419	Quanto fuma por dia ou por semana de Outro?
P027	Com que frequência costuma consumir alguma bebida alcoólica?
P02801	Quantos dias por semana costuma consumir alguma bebida alcoólica?
P029	No dia em que bebe, quantas doses de bebida são consumidas?
J037	Nos últimos 12 meses ficou internado por mais de 24h em um hospital?
H001	Quando foi a última vez que consultou um médico?
N001	Em geral, como você avalia sua saúde?
I00102	Tem algum plano de saúde médico particular, de empresa ou órgão público?
N004	Quando sobe uma ladeira, lance de escadas ou caminha no plano, sente dor no peito?
N005	Quando caminha em um lugar plano em velocidade normal, sente desconforto ou dor no peito?
D001	Sabe ler e escrever?
VDD004A	Nível de instrução mais elevado alcançado.
V0026	Tipo de situação censitária.
A01501	Para onde vai o esgoto do banheiro ou do sanitário ou das dejeções?
A016010	Qual o principal destino dado ao lixo?
A01901	Algum morador possui acesso à internet?
A005010	Qual a forma de abastecimento de água deste domícilio.
VDF004	Faixa de rendimento domiciliar per capita.
Q03001	Algum médico já lhe deu o diagnóstico de diabetes?

Tabela 1: Seleção de Atributos da PNS 2019

Além do mais, após a seleção dos atributos na base de dados PNS 2019, alguns deles sofreram o processo de renomeação dos nomes para melhorar a interpretabilidade durante a construção dos códigos e dos modelos como descrito na tabela 2.

Atributo	Nome atribuído
P00104	Peso
P00404	Altura
C006	Sexo
C008	Idade
C009	Raça_etnia
P02601	Consumo_sal
P034	Atividades_fisicas
P035	Freq_atividade_fisica
P027	Frequencia_alcoolismo
P02801	Qtd_alcool_semanal
P029	Qtd_doses_alcoolicas
J037	Ficou_internado
H001	Ultima_consulta
N001	Percepcao_estado_saude
I00102	Tem_plano
N004	Cansa_subida
N005	Cansa_plano
D001	Alfabetizacao
VDD004A	Escolaridade
V0026	Area_moradia
A01501	Esgoto
A016010	Destino_lixo
A01901	Acesso_internet
A005010	Abastecimento_agua
VDF004	Faixa_salarial
Q03001	Tem_diabetes

Tabela 2: Renomeação de Atributos

Os atributos que não foram renomeados são aqueles que posteriormente sofreram um processo de categorização, discretização ou fusão e serão apresentados ao longo do documento nos próximos passos, como é o caso dos atributos pertencentes ao tabagismo (P050, P052, ... , P054019).

Com a pré-seleção de atributos realizada, foi também realizado um filtro nas idades dos participantes dessa pesquisa, sendo incluídas as pessoas com no mínimo 18 anos de idade. Ao final, a base de dados resultou em 45.757 instâncias e 36 atributos.

Parte 3 - Tratamendo de Dados Nulos: Esse passo consistiu em solucionar os dados nulos presentes no conjunto de dados selecionados, para isso, foi tratado os nulos separadamente por cada Aspecto do Mapa Conceitual (Figura 2), exceto pelo aspecto de hábitos alimentares devido a sua não utilização no pré-processamento e no modelo. Em primeiro lugar, houve a remoção de 575 instâncias nos atributos Peso e Altura, subdividos entre: 573 sendo da classe Saudáveis e 2 da classe Hipertensos com DCV, a decisão foi tomada porque não há uma garantia de valores que sejam representativos mesmo com a utilização de técnicas de imputação por KNN devido a ausência de Peso para gerar Alturas artificiais e na ausência de Altura para gerar Pesos artificiais.

No Aspecto de Atividades Físicas, houve a presença de 26.358 registros nulos no atributo Freq_atividade_fisica que foram substituídos pelo valor 0, tal valor se deu ao fato dessas instâncias serem as mesmas que, para o atributo Atividades_fisicas, responderam que não se exercitaram ou praticaram algum esporte nos últimos três meses, o que informa implicitamente que elas possuem uma frequência semanal de 0 dias de prática de atividades físicas.

No que se diz respeito ao Aspecto do Tabagismo, para os atributos P052, P05401, P05404, P05407, P05410, P05413, P05416, P05419 ocorreu 5.923, 39.127, 39.127, 39.127, 39.127, 39.127, 39.127, 39.127 nulos respectivamente. Para as colunas supracitadas, exceto a coluna P052, foi imputado o valor 0 que significa que essas pessoas consomem esses produtos com nenhuma frequência porque essas pessoas são as mesmas que no atributo P050 que perguntam se elas fumam produtos do tabaco elas respondem que nunca fumaram na vida. Já em relação aos nulos que estão presentes no atributo P052 que pergunta aos participantes se eles fumaram no passado, de 5.923 nulos, 5.268 são pessoas que responderam que são fumantes diários, logo essa pergunta não foi aplicada à eles, sendo imputado o valor 1 que significa que fumaram diariamente no passado. Os 655 registros nulos restantes são de pessoas que fumam menos que diariamente atualmente e para estes, o valor 2 foi inserido que tem como significado que fumaram menos que diariamente no passado.

Durante a análise do Aspecto do Alcoolismo, para o atributo Qtd_alcool_semanal foi encontrado 31.409 valores ausentes, deste montante, 25.273 instâncias responderam no atributo Frequencia_alcoolismo que nunca beberam álcool na vida e, como consequência, o valor 0 foi imputado representando uma frequência de 0 dias de consumo de álcool na semana. Os 6.136 nulos restantes são pessoas que responderam que possuem um hábito de consumo de álcool, sendo subdivididos em 5.863 valores pertencentes à classe dos saudáveis e 273 à classe dos hipertensos com DCV. Nesse caso, foi realizada uma análise por classe para verificar a média e o desvio padrão com o intuito de imputar valores mais fidedignos à cada classe, entretanto, a média da classe dos saudáveis sobre frequência semanal de consumo de álcool foi de 0.38 com um desvio padrão de 1.21, já a classe dos hipertensos com doenças cardiovasculares foi obtido uma média de 0.65 com desvio padrão de 1.28. Em suma, foram médias e desvios padrões muito similares e imputar o valor 1 para o consumo semanal foi considerado como impróprio, isso porque essas mesmas instâncias/pessoas responderam no atributo Frequencia_alcoolismo que consomem álcool 1 vez ou menos por mês, o que tornaria o valor de 1 dia por semana equivocado, portanto, foi imputado o valor de 0 para esses casos. Por último, o atributo Qtd_doses_alcoolicas apresentou 25.273 instâncias com valores ausentes que são as mesmas pessoas que afirmaram anteriormente que não consomem e não consumiram álcool em nenhum momento de suas vidas.

Para o Aspecto de Acompanhamento Médico os seguintes atributos apresentaram valores ausentes: Cansa_subida, Cansa_plano, Tem_diabetes com os valores 537, 537 e 3.746 respectivamente. Essas instâncias foram descartadas do conjunto de dados e todas com pertencimento à classe majoritária dos saudáveis.

No Aspecto Social, houve 560 valores nulos para o Atributo Esgoto, subdivididos em 533 pertencendo à classe dos saudáveis e 27 para a classe das pessoas hipertensas com DCV. Foi realizada uma análise minuciosa para essas instâncias para entender os perfis sociais e mais de 60% desses indivíduos recebem até 1 salário mínimo por mês, sendo pessoas com uma instrução baixa, isso significa que, não completaram o ensino médio e residem em área rural sem ligação à rede de abastecimento d'água. Para essas instâncias foi imputado o valor 4 que

representa que o esgoto é descartado em uma fossa comum que melhor representa esse perfil em relação aos valores não nulos.

Por último, o Aspecto Econômico apresentou 13 instâncias com valores nulos no atributo Faixa_salarial e foram descartados da base de dados.

Parte 4 - Remoção de Outliers: Em relação à esta etapa, há como objetivo a remoção dos Outliers que são dados muito discrepantes em relação à distribuição do restante dos dados e podem gerar muitos ruídos pela alta distorção que eles geram.

No Aspecto de Características do Indivíduo, foi realizada uma análise pelo Intervalo Interquartil (IQR) que consiste em encontrar o limite inferior e superior da distribuição dos dados que é calculado de acordo com a figura 3:

$$\text{Limite Inferior} = Q_1 - 1,5 \cdot \text{IQR}$$
$$\text{Limite Superior} = Q_3 + 1,5 \cdot \text{IQR}$$

Onde:

- Q_1: Primeiro quartil (25% dos dados estão abaixo deste valor);
- Q_3: Terceiro quartil (75% dos dados estão abaixo deste valor);
- IQR: Intervalo interquartil, dado por $\text{IQR} = Q_3 - Q_1$.

Figura 3: Cálculo Intervalo interquartil

Nesse estudo, foi aplicado a utilização do boxplot da biblioteca Seaborn do Python, por padrão ela calcula o limite inferior e o limite superior baseando-se na multiplicação de 1.5 pelo IQR. Dessa forma, são considerados Outliers aqueles valores que encontram-se acima do limite superior ou abaixo do limite inferior.

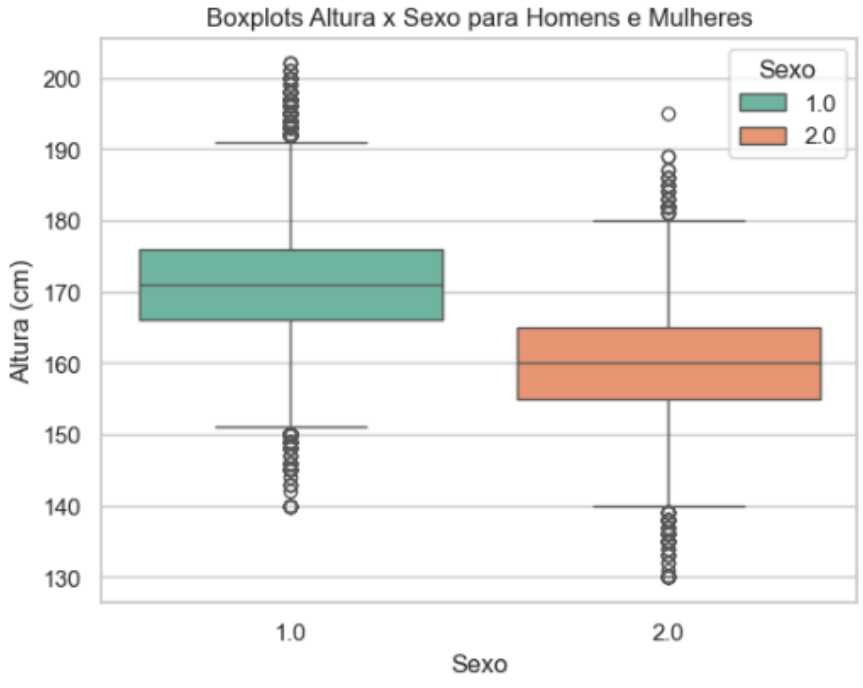

Figura 4: Boxplot Altura x Sexo

Como é observado na figura 4, a utilização do cálculo de Outliers baseado nos limites calculados a partir do IQR encontram, de maneira geral, para os dois sexos Homem (1) e Mulher (2) muitos possíveis outliers mas, visualmente, nenhum valor que distorça muito a distribuição geral dos dados, isso porque a diferença de alturas mantém a variabilidade dos dados originais e portanto foi utilizado do cálculo da multiplicação de 3 pelo IQR ao invés de utilizar o limiar 1.5, isso porque ao aumentar este limiar o processo de cálculo de outliers fica mais conservador e rigoroso, com uma estimação de 0.7% de eliminação dos dados em relação ao total dos dados disponíveis quando trata-se de uma distribuição normal [17].

Em relação ao Peso, o boxplot padrão (figura 5) que utiliza o limiar 1.5 encontrou 4 possíveis outliers abaixo do limite inferior e 847 acima do limite superior. Entretanto, os valores considerados outliers que encontram-se acima do limite superior não foram descartados porque são indivíduos que configuram-se como pessoas com sobrepeso ou obesidade e são importantes para o contexto estudado, logo, foram removidos apenas os outliers inferiores ao limite inferior.

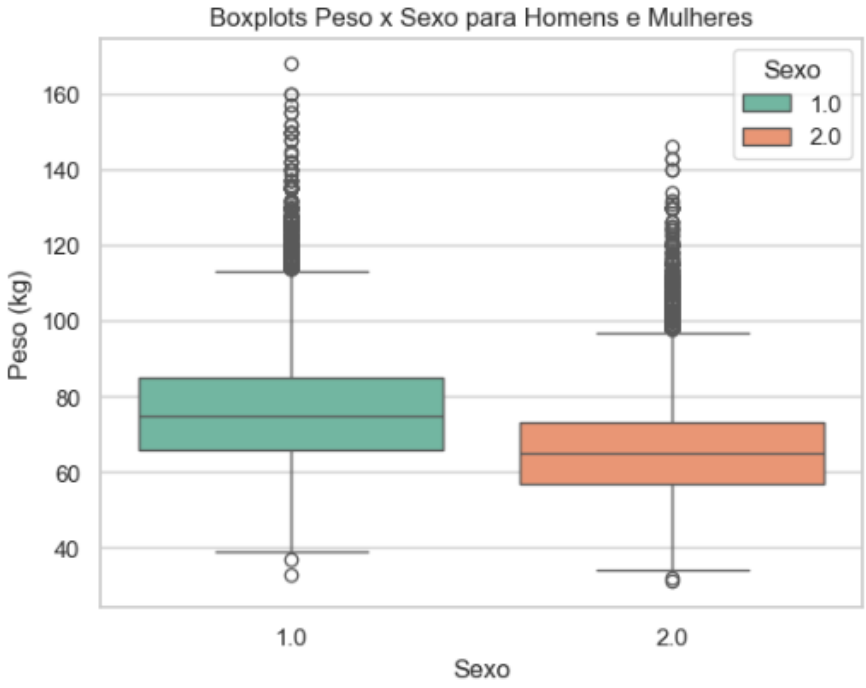

Figura 5: Boxplot Peso x Sexo

Em relação ao consumo de álcool, foram encontradas instâncias que consomem acima de 15 doses de álcool por dia. De acordo com o estudo [18], cada "dose" de álcool geralmente é definida como 14 gramas de álcool puro, o que equivale a uma bebida padrão. Consumir 15 doses significa ingerir 210 gramas de álcool, um valor muito acima dos limites diários recomendados para a saúde. Esse nível de consumo pode rapidamente resultar em intoxicação alcoólica ou danos permanentes a longo prazo. Portanto, essas instâncias foram eliminadas do conjunto de dados.

Parte 4 - Redução de Dimensionalidade: Em relação à esta etapa, ela tem como foco principal reduzir a dimensionalidade de atributos que compartilham da mesma informação para o problema, foi aplicado no aspecto de tabagismo onde os atributos listados a seguir foram fundidos em categorias de tabagismo.

- P050: Atualmente, o(a) Sr(a) fuma algum produto do tabaco?
- P052: E no passado, o(a) Sr(a) fumou algum produto do tabaco?
- P05401: Em média, quanto fuma por dia ou por semana Cigarros industrializados?
- P05404: Em média, quanto fuma por dia ou por semana Cigarros de palha ou enrolados a mão?
- P05407: Em média, quanto fuma por dia ou por semana Cigarros de cravo ou de Bali?
- P05410: Em média, quanto fuma por dia ou por semana Cachimbos (considere cachimbos cheios)?
- P05413: Em média, quanto fuma por dia ou por semana Charutos ou cigarilhas?
- P05416: Em média, quanto fuma por dia ou por semana Narguilé (sessões)?
- P05419: Em média, quanto fuma por dia ou por semana Outro?

Para as categorias formuladas, a tabela 3 apresenta a categoria de tabagismo e sua lógica.

Categoria de Tabagismo	Condição Lógica
Fuma Muito	Se "P050" = 1 (Fuma diariamente atualmente)
Fuma Razoavelmente	Se "P050" = 3 e "P052" = 1 (Não fuma atualmente, mas fumou diariamente no passado)
Fuma Pouco	Se "P050" = 3 e "P052" = 2 (Não fuma atualmente, mas fumou menos que diariamente no passado)
Fuma Pouco	Se "P050" = 2 e ("P05401" = 4 ou "P05404" = 4 ou "P05410" = 4 ou "P05413" = 4 ou "P05416" = 4 ou "P05419" = 4) (Fuma menos que diariamente atualmente e usa produtos de tabaco raramente)
Fuma Razoavelmente	Se "P050" = 2 e ("P05401" = 2 ou "P05404" = 2 ou "P05410" = 2 ou "P05413" = 2 ou "P05416" = 2 ou "P05419" = 2) (Fuma menos que diariamente atualmente e usa produtos de tabaco razoavelmente)
Fuma Muito	Se "P050" = 2 e ("P05401" = 1 ou "P05404" = 1 ou "P05410" = 1 ou "P05413" = 1 ou "P05416" = 1 ou "P05419" = 1) (Fuma menos que diariamente atualmente e usa produtos de tabaco frequentemente)
Não Fuma	Se "P050" = 3 e "P052" = 3 (Não fuma atualmente e nunca fumou no passado)
Não Fuma	Se houver valores ignorados em "P050" ou "P052" (Valores ignorados são tratados como não fumantes)

Tabela 3: Condições lógicas de categorização do tabagismo.

Portanto, 9 atributos do tabaco foram unidos e categorizados para o atributo chamado de Categoria_tabagismo representado na figura a seguir:

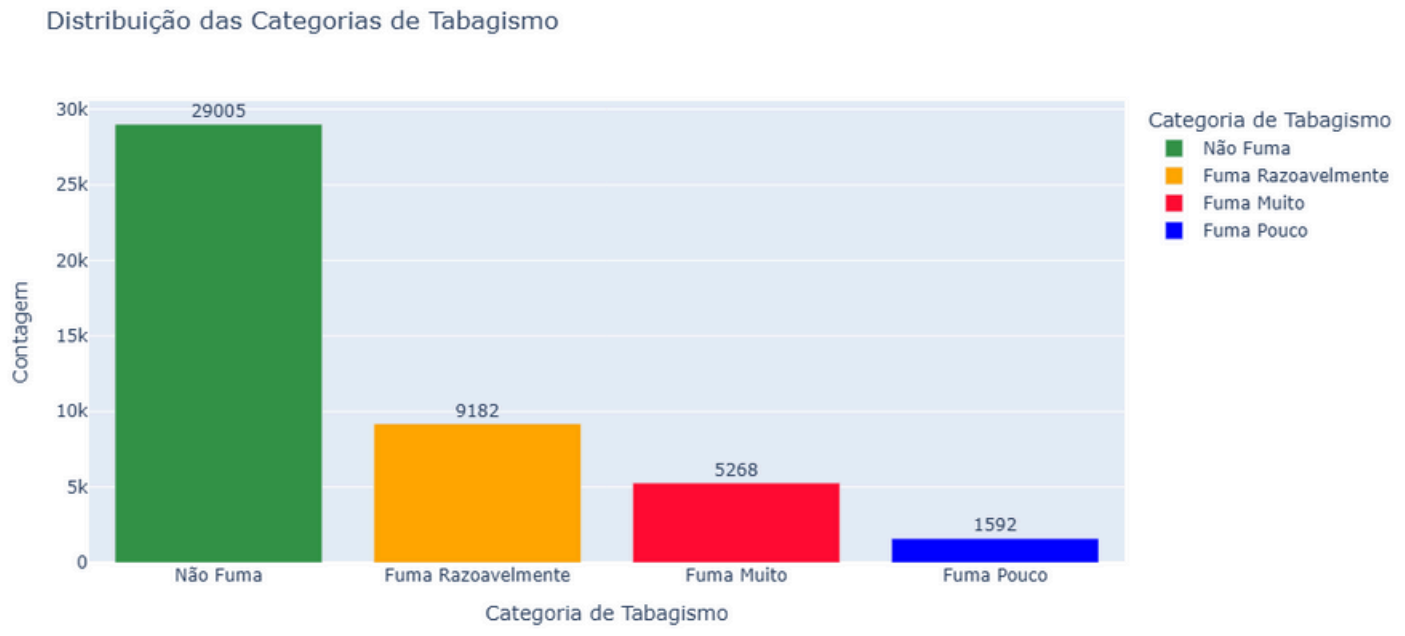

Figura 6: Categorias do tabagismo

Para descrever a característica de alcoolismo dos indivíduos, também foi criada uma codificação com o intuito de reduzir a dimensionalidade das variáveis relacionadas ao álcool, dessa forma, os atributos Qtd_doses_alcoolicas, Frequencia_alcoolismo, Qtd_alcool_semanal foram fundidos e categorizados para apenas um, chamado de categoria_alcoolismo (figura 7).

Categoria de Alcoolismo	Condição Lógica
Não alcoólico	Frequência do alcoolismo = 1
Bebedor social	Quantidade de doses alcoólicas < 1 e Quantidade semanal de álcool < 3 e Frequência de alcoolismo = 2
Bebedor moderado	(Quantidade de doses alcoólicas ≥ 1 e ≤ 2) ou Quantidade semanal de álcool ≤ 3 e Frequência de alcoolismo = 3
Bebedor frequente	(Quantidade de doses alcoólicas > 2) ou Quantidade semanal de álcool ≤ 4 e Frequência de alcoolismo = 4
Bebedor excessivo	Quantidade de doses alcoólicas > 3 ou Quantidade semanal de álcool > 4

Tabela 4: Condições lógicas de categorização do alcoolismo

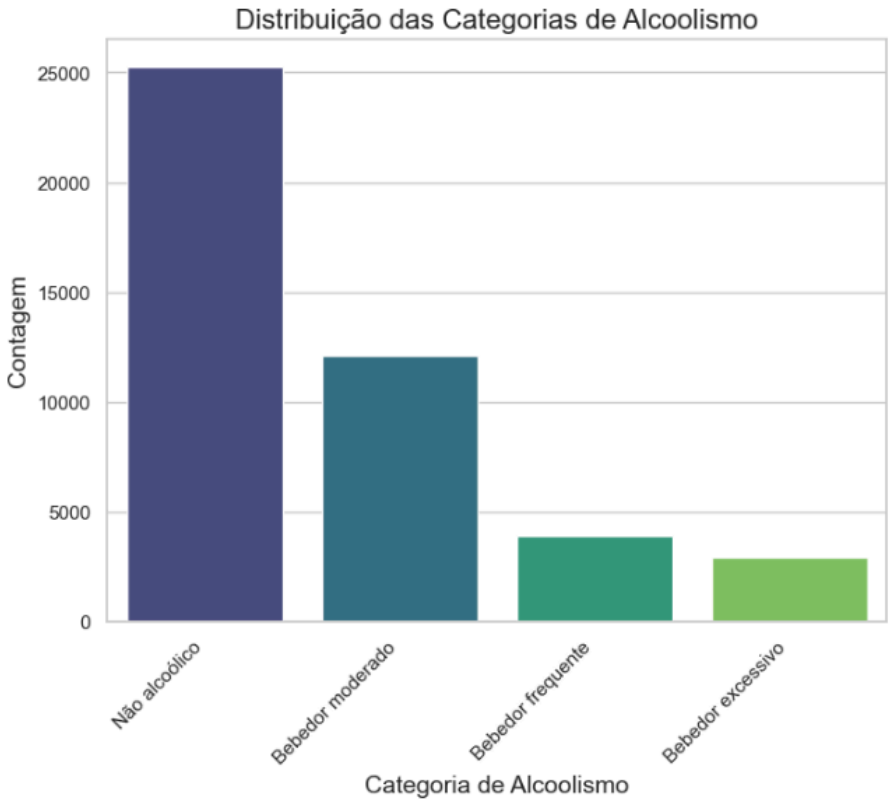

Figura 7: Categorias do alcoolismo

Para mais, houve a categorização do atributo Racas_Etnias entre: Brancos, Pretos e Pardos, sendo que as raças/etnias minoritárias como, amarelos, indigenas e as pessoas que ignoraram essa pergunta foram inseridas na classificação como sendo brancas. Tal medida adotada de agrupação entre três grandes grupos deve-se à predisposição genética das pessoas pardas e pretas à desenvolverem a hipertensão arterial em maior grau em relação às pessoas brancas [19, 20]. As figuras 8 e 9 apresentam a configuração das Raças/Etnias que estavam por padrão na PNS filtrada por idade e pelos atributos selecionados conceitualmente e após a redução entre os três grupos formulados, respectivamente.

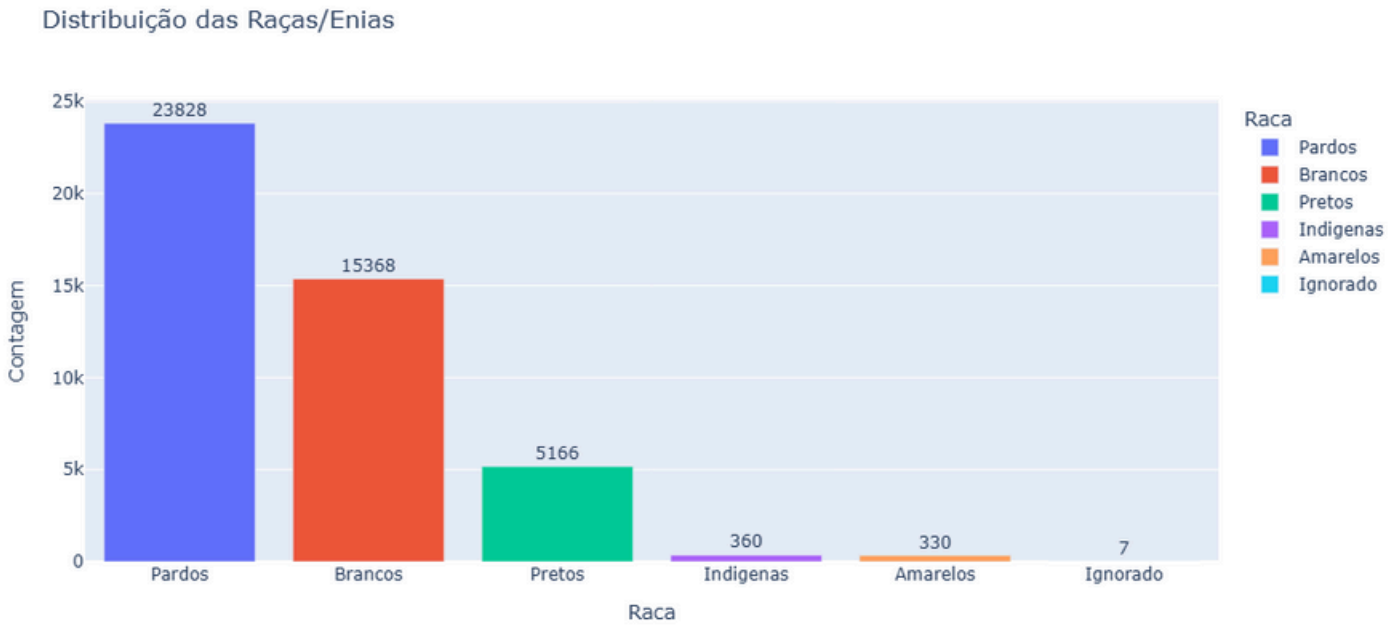

Figura 8: Categorias Raças/Etnias Originais

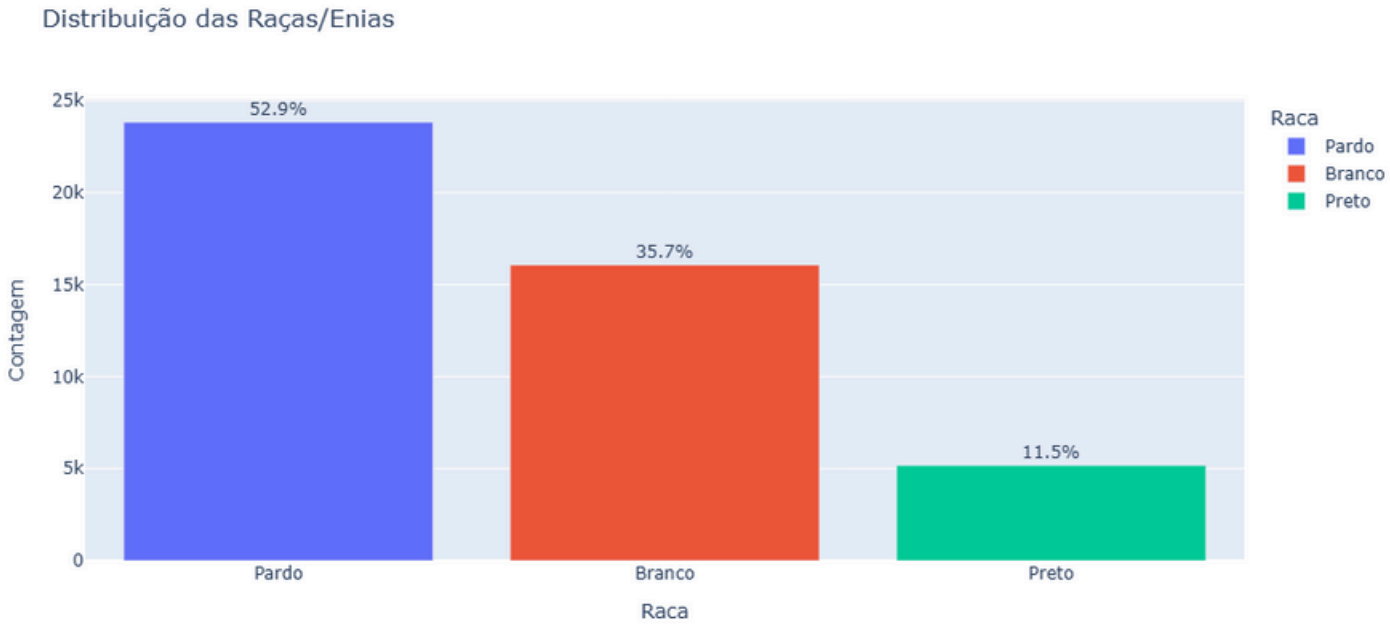

Figura 9: Categorias Raças/Etnias Modificada

Considerando todas as pessoas do conjunto de dados, a soma das porcentagens dos pretos e pardos totalizam 64.4% de indivíduos com uma predisposição genética à desenvolverem problemas de hipertensão.

Por último, foi preciso criar um novo atributo para calcular o IMC de cada indivíduo presente no conjunto de dados, visto que a obesidade e o sobrepeso são fatores de risco para o surgimento da hipertensão arterial e das DCV [21]. O IMC, calcula-se da seguinte forma:

$$IMC = \frac{\text{Peso (kg)}}{\left(\frac{\text{Altura (cm)}}{100}\right)^2}$$

Após o cálculo do IMC, foi feita a categorização desses valores em 4 classes dispostas na tabela 4 de acordo com a OMS [22].

Categoria IMC	Condição Lógica
Baixo Peso	IMC <18.5
Peso Ideal	$18.5 \leq IMC \leq 24.9$
Sobrepeso	$25 \leq IMC \leq 29.9$
Obeso	IMC $\geq$ 30

Tabela 5: Condições lógicas de categorização do IMC

Com a categorização do IMC aplicada, foi criado um novo atributo chamado de Categoria_IMC, disposto na figura 10 que representa um histograma:

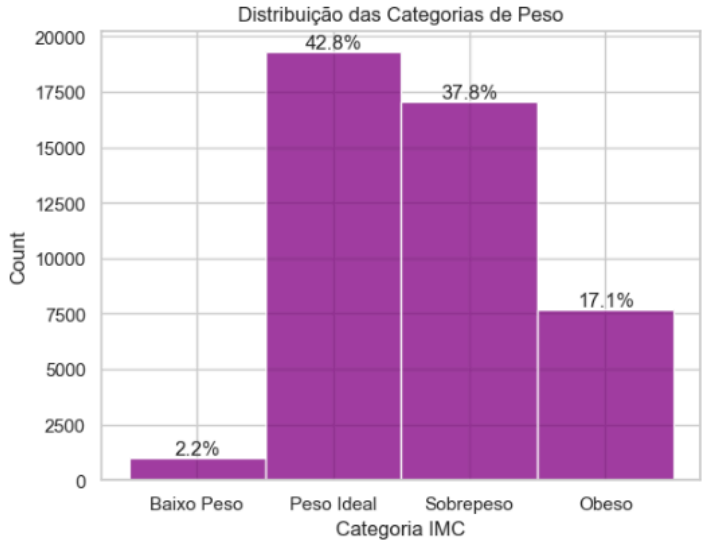

Figura 10: Histograma Categorias IMC

Com a plotagem do histograma, 54,9% das pessoas encontram-se acima do peso ideal de acordo com a OMS, configurando como um fator de risco para o desenvolvimento das doenças crônicas exploradas neste estudo.

Parte 5 - Balanceamento das Classes: Foram identificadas 37.507 instâncias que pertencem à classe dos Saudáveis/sem diagnóstico e 2.961 instâncias que são da classe dos indivíduos que possuem o diagnóstico positivo para as doenças crônicas de hipertensão e doenças cardiovasculares. O primeiro passo foi realizar a divisão do conjunto de dados em treino e teste, sendo 20% do total reservado ao teste e 80% ao conjunto de treinamento.

No que se diz respeito ao conjunto de treinamento, 30.005 instâncias foram postas como sendo da classe dos Saudáveis/sem diagnóstico e 2.369 pertencendo à classe das pessoas com HA e DCV. Diante do desbalanceamento das classes, foi utilizada a técnica de subamostragem aleatória (RandomUnderSampling) que reduziu a classe majoritária e equalizada aleatoriamente

em relação à classe minoritária. Como resultado, foram selecionadas 2.369 instâncias aleatórias das 30.005 instâncias pertencentes à classe dos Saudáveis/sem diagnóstico, totalizando um conjunto de treinamento de 4.738 instâncias, com 25 atributos independentes, isto é, sem a variável target.

Já o conjunto de teste não foi balanceado para manter a representatividade original dos dados e transmitir ao futuro modelo a realidade dos dados. Este conjunto dispõe de 7.502 instâncias da classe Saudáveis e 592 instâncias da classe das pessoas com HA e DCV.

Parte 6 - Aplicação dos Modelos de Aprendizado de Máquina: O último passo da metodologia seguida envolve aplicar os algoritmos de aprendizado de máquina e interpretar os resultados retornados por eles aplicados no contexto do domínio do problema estudado, nesse caso, da classificação de pessoas diagnosticadas com hipertensão e doenças cardiovasculares ou saudáveis/sem diagnóstico. No estudo de O. Loyola González [7] o autor demonstra diferentes vantagens e desvantagens dos modelos estilo caixa preta e os que são consideradas como caixa branca ou, em outras palavras, os interpretáveis. Para o escopo do problema estudado nessa pesquisa, foram escolhidos os modelos: Árvore de Decisão, Floresta Aleatória e o Nayve-Bayes. Em relação aos modelos de caixa-preta adotados, foi utilizado a rede neural classificadora MLP que possui uma menor interpretabilidade porém, por ser mais complexo, há a esperança de um melhor ajuste no comportamento dos dados.

4. Experimentos e Análises dos Resultados

Os experimentos realizados visaram avaliar a performance de diferentes modelos de aprendizado de máquina no contexto proposto. Nesta seção, apresentamos os detalhes da parametrização dos algoritmos e as análises dos resultados obtidos.

4.1. Parametrização dos Algoritmos

Para os modelos de Árvore de Decisão e Floresta Aleatória, foi utilizado o GridSearch e o RandomSearch , ambos da biblioteca Scikit-Learn da linguagem Python, como otimizadores de hiperparâmetros dos modelos, os dois obtiveram uma acurácia muito similar com uma acurácia de, aproximadamente, 86%. Os hiperparâmetros encontrados para esses dois modelos foram:

- Árvore de Decisão:

Hiperparâmetro	Descrição	Valor Definido
`min_samples_split`	Número mínimo de amostras necessário para dividir um nó.	10
`min_samples_leaf`	Número mínimo de amostras que um nó folha deve conter.	4
`max_features`	Proporção de variáveis consideradas em cada divisão.	0.4
`max_depth`	Profundidade máxima da árvore, para evitar overfitting.	10
`criterion`	Métrica para medir a qualidade da divisão.	`entropy`

Tabela 6: Hiperparâmetros Modelo Árvore de Decisão

- Floresta Aleatória:

Hiperparâmetro	Descrição	Valor Definido
max_depth	Profundidade máxima das árvores, limitando seu crescimento para evitar overfitting.	10
max_features	Porcentagem de variáveis usadas para a divisão em cada nó.	0.4
min_samples_split	Número mínimo de amostras necessário para dividir um nó.	10
n_estimators	Número total de árvores na floresta.	150
n_jobs	Número de threads usadas no treinamento (-1 utiliza todos os processadores).	-1

Tabela 7: Hiperparâmetros Modelo Floresta Aleatória

No que concerne à rede neural MLP, estilo caixa-preta, os hiperparâmetros foram:

- Rede Neural MLP Classificadora:

Hiperparâmetro	Descrição	Valor Definido
hidden_layer_sizes	Define a quantidade de neurônios e camadas ocultas na rede. Cada número indica o número de neurônios em uma camada.	(100, 100, 100)
activation	Função de ativação utilizada para transformar a saída de cada neurônio.	relu
alpha	Taxa de regularização L2 para evitar overfitting.	0.0001
learning_rate	Ajuste dinâmico da taxa de aprendizado durante o treinamento.	adaptive
solver	Método de otimização utilizado para ajustar os pesos da rede.	adam
max_iter	Número máximo de iterações para o treinamento.	200

Tabela 8: Hiperparâmetros Modelo MLP

Ademais, para os modelos foi utilizado o método de validação cruzada com k-folds = 10.

4.2. Métricas de Classificação

Quanto as métricas de classificação utilizadas para avaliar o desempenho dos modelos foram adotadas:

I. Precisão (Precision): que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo, indicando a **confiabilidade** dos resultados positivos

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos (VP)}}{\text{Verdadeiros Positivos (VP)} + \text{Falsos Positivos (FP)}}$$

II. Sensibilidade (Recall): avalia a proporção de exemplos positivos corretamente identificados pelo modelo, ou seja, a capacidade do modelo de **detectar** todos os casos positivos.

$$\text{Sensibilidade} = \frac{\text{Verdadeiros Positivos (VP)}}{\text{Verdadeiros Positivos (VP)} + \text{Falsos Negativos (FN)}}$$

III. F1-Score: é a média harmônica entre Precisão e Sensibilidade, fornecendo uma medida balanceada que considera tanto falsos positivos quanto falsos negativos.

$$F1\text{-Score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}$$

Dessa forma, essas métricas levarão em consideração a classe minoritária no momento de avaliação do conjunto de teste que é desbalanceado.

4.3. Desempenho dos Modelos e Discussão dos Resultados

Primeiramente, no que tange à classe das pessoas com hipertensão (HA) e doenças cardiovasculares (DCV) que representam a Classe 0, a tabela 9 traz o resultado de cada algoritmo de aprendizado de máquina nessa classe em específico. Todos os modelos utilizados obtiveram um resultado aproximado entre si, por volta dos 60% de F1-Score.

Algoritmo	Precisão Pessoas HA + DCV	Sensibilidade Pessoas HA + DCV	F1-Score Pessoas HA + DCV
Árvore de Decisão	0.65	0.58	0.61
Floresta Aleatória	0.65	0.58	0.61
Naive Bayes	0.64	0.54	0.58
Rede Neural MLP	0.62	0.57	0.60

Tabela 9: Métricas de classificação dos modelos classe de pessoas HA + DCV

Por outro lado, em relação à classe das pessoas consideradas saudáveis que pertencem à classe 1, isto é, as pessoas que não possuem o diagnóstico positivo das doenças crônicas estudadas, a tabela 10 contém as métricas de classificação e seus valores obtidos.

Algoritmo	Precisão Pessoas Saudáveis	Sensibilidade Pessoas Saudáveis	F1-Score Pessoas Saudáveis
Árvore de Decisão	0.97	0.98	0.97
Floresta Aleatória	0.97	0.98	0.97
Naive Bayes	0.96	0.98	0.97
Rede Neural MLP	0.97	0.97	0.97

Tabela 10: Métricas de classificação dos modelos classe de pessoas saudáveis

Com a observação dos resultados presentes nas tabelas 9 e 10, os modelos entregam um resultado muito similar, sendo a Árvore de Decisão e a Floresta Aleatória apresentando os melhores resultados para a classe de pessoas com HA + DCV com uma precisão de 65% e um sensibilidade de 58%. Isso significa que esses modelos em 65% dos novos casos dessa classe serão acertados na classificação e que 58% deles serão reconhecidos como sendo pessoas que possuem as duas doenças crônicas. Por outro lado, esses mesmos algoritmos trazem 97% de precisão e 97% de sensibilidade, indicando que em 97% dos novos casos que pertencem à classe dos saudáveis serão classificados corretamente e que 97% deles serão detectados. O pior resultado obtido para o contexto médico, que há a priorização de uma sensibilidade maior foi para o algoritmo Naive Bayes em relação à classe das pessoas doentes com 54%.

A seguir, são apresentadas as matrizes de confusão de cada algoritmo:

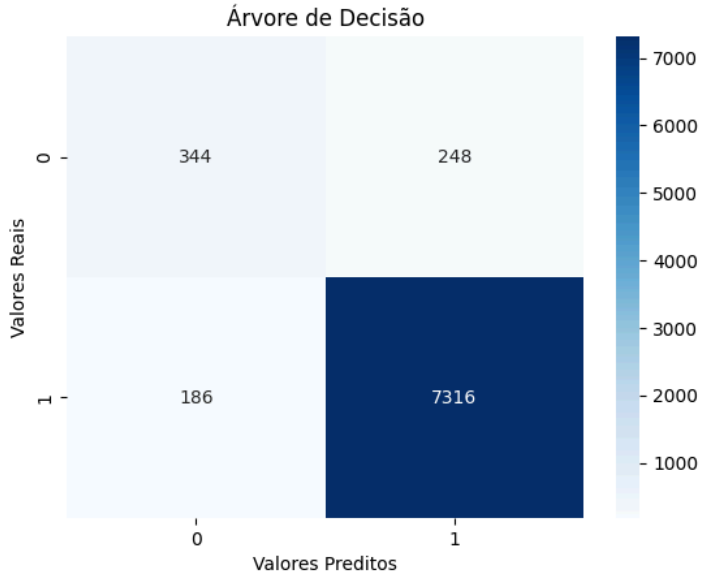

Figura 11: Matriz de Confusão do modelo Árvore de Decisão

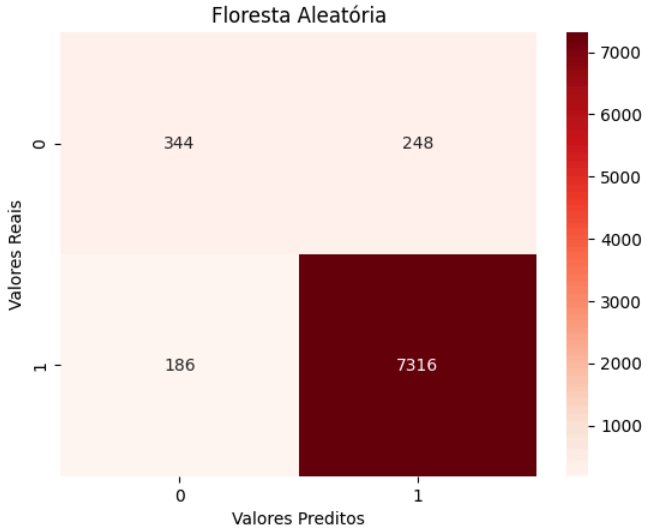

Figura 12: Matriz de Confusão do modelo Floresta Aleatória

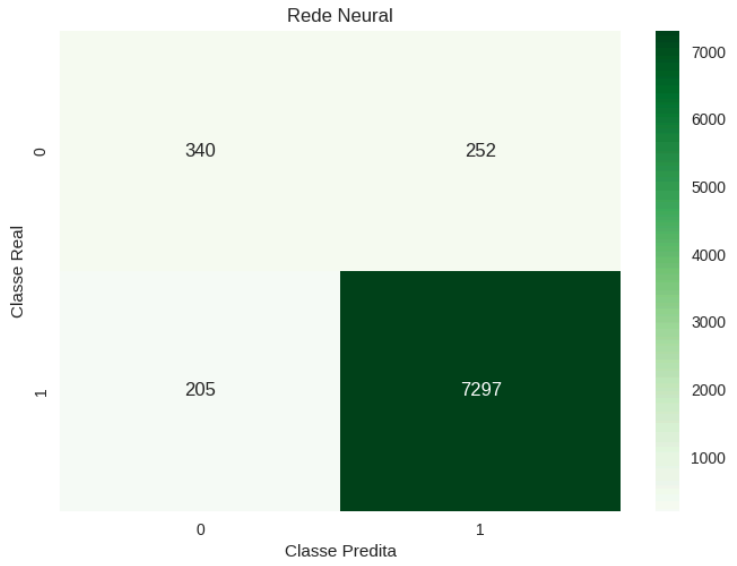

Figura 13: Matriz de Confusão do modelo MLP

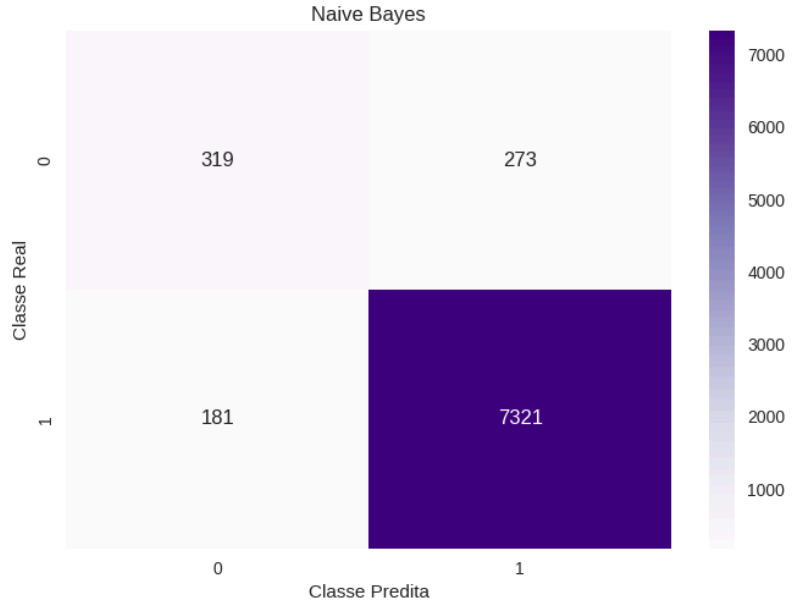

Figura 14: Matriz de Confusão do modelo MLP

O desempenho dos quatro modelos testados são bem similares em relação à classificação errônea das classes, e tal fato pode trazer a ideia de que pela PNS 2019 tratar-se de uma base de dados descritiva sobre saúde e não cliníca, as pessoas que são da classe saudáveis mas foram classificadas como sendo da classe HA + DCV não necessariamente está errado, há a possibilidade dessas pessoas possuirem as doenças crônicas mas não obtiveram o diagnóstico positivo ainda. Por outro lado, as pessoas que possuem o diagnóstico positivo para essas doenças e estão inseridas na classe HA + DCV e que foram classificadas de forma equivocada pertencendo a classe dos saudáveis podem ser pessoas que, em algum momento da vida, obtiveram essas doenças mas que atualmente adotaram um estilo de vida mais saudável como a adoção de prática de atividades físicas, acompanhamento médico recorrente, tratamentos preventivos e portanto possuem um perfil similar com as pessoas que são, de fato, saudáveis.

Portanto, possuem características que são relevantes para o contexto clínico e médico. A priorização da sensibilidade, como discutido, é fundamental em contextos onde o diagnóstico precoce é necessário para a implementação de estratégias de saúde pública eficazes. Isso é especialmente relevante em relação às classes de pessoas com hipertensão (HA) e doenças cardiovasculares (DCV), onde um modelo que minimize o risco de classificação incorreta de pessoas doentes como saudáveis pode ajudar a garantir que essas pessoas recebam o acompanhamento adequado.

Somado à isso, as questões relacionadas à base de dados da PNS 2019, que é uma base de dados descritiva e não clínica, podem justificar a presença de alguns casos em que pessoas classificadas como saudáveis ainda possam ser portadoras das doenças crônicas, mas não diagnosticadas. Por outro lado, aqueles que possuem o diagnóstico positivo, mas são classificados como saudáveis, podem estar em um estágio de melhora ou controle das doenças devido a mudanças em seus hábitos de vida. Isso reforça a necessidade de analisar não apenas as métricas de precisão e sensibilidade, mas também o contexto em que essas classificações ocorrem, permitindo uma interpretação mais rica dos resultados.

5. Conclusões

O estudo atual demonstrou a eficácia dos modelos de aprendizado de máquina na classificação de pessoas com hipertensão e doenças cardiovasculares (HA + DCV), bem como na identificação de pessoas saudáveis. Embora os modelos, incluindo a Árvore de Decisão, a Floresta Aleatória e o Naive Bayes, apresentem um desempenho semelhante em termos de F1-Score, observou-se que a precisão e a sensibilidade para a classe dos saudáveis se destacam, com resultados de 97% em ambos os casos. Isso sugere que os algoritmos foram capazes de identificar corretamente a maioria das pessoas sem doenças crônicas e sem diagnóstico positivo, o que é importante para precaver possíveis avanços dessas doenças crônicas.

Entretanto, é importante salientar as limitações dos modelos em relação à classe de pessoas com HA + DCV, onde a sensibilidade foi inferior, por volta de 60% nos modelos, indicando que uma parcela considerável dessas pessoas não foi identificada corretamente, possivelmente devido à falta de diagnóstico médico formal ou ao efeito de fatores externos, como estilos de vida mais saudáveis. Isso sugere a necessidade de um modelo mais complexo, que leve em conta variáveis adicionais, como o histórico médico detalhado e o comportamento ao longo do tempo.

Além disso, os resultados obtidos nos modelos indicam que a base de dados, sendo descritiva e não clínica, pode apresentar desafios para a precisão das previsões, já que pessoas saudáveis podem, em alguns casos, ainda estar em risco de desenvolver doenças crônicas no futuro, e pessoas com diagnósticos positivos podem ter controlado suas condições com a adoção de hábitos saudáveis. Assim, o estudo contribui para uma melhor compreensão dos limites da modelagem preditiva no contexto de saúde pública e destaca a importância de considerar não apenas os dados clínicos, mas também fatores comportamentais e socioeconômicos.

Para futuras pesquisas, sugere-se a inclusão de mais variáveis, como dados longitudinais de acompanhamento médico e características genéticas, que possam enriquecer os modelos e melhorar sua acurácia, especialmente em relação à classificação da classe de pessoas com hipertensão e doenças cardiovasculares.

6. Referências

[1] World Health Organization (WHO). Global Atlas on Cardiovascular Disease Prevention and Control. Mendis S, Puska P, Norrving B editors. Geneva: World Health Organization; 2011.

[2] Alwan,A., A. Alwaneditor, World Health Organization, 20113168808, English, Book, Switzerland, 9789241564229, Geneva, Global status report on noncommunicable diseases 2010, (176 pp.), World Health Organization, Global status report on noncommunicable diseases 2010., (2011)

[3] Malta, D.C. et al. 2022. Hipertensão arterial e fatores associados: Pesquisa Nacional de Saúde, 2019. *Revista de Saúde Pública*. 56, (dez. 2022), 122. DOI:https://doi.org/10.11606/s1518-8787.2022056004177.

[4] B. Stevens, L. Pezzullo, L. Verdian, J. Tomlinson, A. Georgeand F. Bacal, “The Economic Burden of Heart Conditions in Brazil”, *Arq. Bras. Cardiol.*, vol. 111, no. 1, pp. 29–36, Jul. 2018, doi: 10.5935/abc.20180104.

[5] J. M. de Araújo, R. E. de Alencar Rodrigues, A. N. da Costa Pereira de Arruda Neta, F. E. Leite Lima Ferreira, R. L. F. Cavalcanti de Lima, et al., "The direct and indirect costs of cardiovascular diseases in Brazil," *PLOS ONE*, vol. 17, no. 12, p. e0278891, 2022. doi: 10.1371/journal.pone.0278891.

[6] L. A. AlKaabi, L. S. Ahmed, M. F. Al Attiyah, e M. E. Abdel-Rahman, "Predicting hypertension using machine learning: Findings from Qatar Biobank Study," *PLOS ONE*, vol. 15, no. 10, p. e0240370, 2020. doi: 10.1371/journal.pone.0240370.

[7] C. M. Bhatt, P. Patel, T. Ghetia, e P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023. doi: 10.3390/a16020088.

[8] N. M. de Carvalho, M. P. S. Gomes, e L. E. Zárate, “Mineração de dados no diagnóstico de hipertensão baseado na Pesquisa Nacional em Saúde 2019”, *J Health Inform*, vol. 16, nº Especial, nov. 2024.

[9] A. K. Gárate-Escamila, A. H. El Hassani, e E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, p. 100330, 2020. doi: 10.1016/j.imu.2020.100330.

[10] A. Saboor, M. Usman, S. Ali, A. Samad, A. Ali, M. F. Abrar, e N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2022, p. 1410169, 9 pages, 2022. doi: 10.1155/2022/1410169.

[11] A. D’Souza, "Heart disease prediction using data mining techniques," *International Journal of Research in Engineering and Science (IJRES)*, vol. 3, no. 3, pp. 74-77, Mar. 2015. [Online]. Available: www.ijres.org.

[12] M. S. Gangadhar, K. V. S. Sai, S. H. S. Kumar, K. A. Kumar, M. Kavitha and S. S. Aravinth, "Machine Learning and Deep Learning Techniques on Accurate Risk Prediction of Coronary Heart Disease," *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2023, pp. 227-232, doi: 10.1109/ICCMC56507.2023.10083756.

[13] M. Kavitha, G. Gnaneswar, R Dinesh, Y.R Sai and R. S. Suraj, "Heart disease prediction using hybrid machine learning model", *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 1329-1333, January 2021

[14] H. Ayatollahi, L. Gholamhosseini and M. Salehi, "Predicting coronary artery disease: a comparison between two data mining algorithms", *BMC public health*, vol. 19, no. 1, pp. 1-9, 2019.

[15] S. Kaur, K. Bansal and Y. Kumar, "Machine Learning based Approaches for Accurately Diagnosis and Detection of Hypertension Disease," *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gautam Buddha Nagar, India, 2023, pp. 169-173, doi: 10.1109/UPCON59197.2023.10434428.

[16] L. Zárate, B. Petrocchi, C. D. Maia, C. Felix, e M. Gomes, "CAPTO - A method for understanding problem domains for data science projects: CAPTO - Um método para entendimento de domínio de problema para projetos em ciência de dados," *Concilium*, vol. 23, pp. 922-941, 2023. doi: 10.53660/CLM-1815-23M33.

[17] J. Yang, S. Rahardja, e P. Fränti, "Outlier detection: how to threshold outlier scores?" in *Proc. of the Int. Conf. on Artificial Intelligence, Information Processing and Cloud Computing*, Sanya, China, 2019, pp. 37-42, doi: 10.1145/3371425.3371427.

[18] National Institute on Alcohol Abuse and Alcoholism, "Standard Alcohol Guidelines," *National Institute on Alcohol Abuse and Alcoholism*, 2022. [Online]. Available: https://medicine.howard.edu/sites/medicine.howard.edu/files/2022-08/1.2%20NIAAA%20Standard%20Alcohol%20Guidelines-Senior_0.pdf. [Accessed: Dec. 1, 2024].

[19] Zilbermint, M.; Hannah-Shmouni, F.; Stratakis, C.A. Genetics of Hypertension in African Americans and Others of African Descent. *Int. J. Mol. Sci.* **2019**, *20*, 1081. https://doi.org/10.3390/ijms20051081

[20] C. T. Sousa, A. Ribeiro, S. M. Barreto, L. Giatti, L. Brant, P. Lotufo, D. Chor, A. A. Lopes, S. S. Mengue, A. O. Baldoni, e R. C. Figueiredo, "Diferenças raciais no controle da pressão arterial em usuários de anti-hipertensivos em monoterapia: resultados do estudo ELSA-Brasil," *Arq. Bras. Cardiol.*, vol. 118, no. 3, pp. 614-622, mar. 2022.

[21] T. M. Powell-Wiley, P. Poirier, L. E. Burke, J.-P. Després, P. Gordon-Larsen, C. J. Lavie, S. A. Lear, C. E. Ndumele, I. J. Neeland, P. Sanders, e M.-P. St-Onge, "Obesity and cardiovascular disease: A scientific statement from the American Heart Association," *Circulation*, vol. 143, no. 21, pp. e84-e118, May 2021, doi: https://doi.org/10.1161/CIR.0000000000000973

[22] WHO, *Obesity and Overweight*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. [Accessed: Dec. 2, 2024].

[23] O. Loyola-González, "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View," in *IEEE Access*, vol. 7, pp. 154096-154113, 2019, doi: 10.1109/ACCESS.2019.2949286. keywords: {Machine learning;Mathematical model;Biological system modeling;Gallium nitride;Biological neural networks;Statistical analysis;Computational modeling;Black-box;white-box;explainable artificial intelligence;deep learning}