Mineração de Dados para Caracterizar Indivíduos Hipertensos com Doenças Cardiovasculares no Brasil

Gustavo Costa[1], **Luis Enrique Zárate Gálvez**[1]

[1]Curso Ciência de Dados e Inteligência Artificial
Laboratório de Inteligência Computacional Aplicada - LICAP
Pontificia Universidade Católica de Minas Gerais (PUC-Minas)
CEP — 30140-100 — Belo Horizonte — MG — Brasil

gustavocosta.ds09@gmail.com, zarate@pucminas.br

Abstract. *This study applied data mining to classify healthy individuals and those with hypertension and cardiovascular diseases (HA + CVD) in Brazil, using data from the 2019 National Health Survey (PNS). Algorithms such as Decision Tree, Random Forest, and Naive Bayes were tested. The models performed similarly, with Random Forest achieving 97% accuracy and sensitivity in identifying healthy individuals. However, classifying HA + CVD cases was more challenging, with lower sensitivity, possibly due to the absence of formal diagnoses and lifestyle factors. The results highlight the importance of more detailed and longitudinal data to improve the identification of chronic diseases.*

Resumo. *Este estudo utilizou mineração de dados para classificar indivíduos saudáveis e hipertensos com doenças cardiovasculares (HA + DCV) no Brasil, a partir da PNS 2019. Foram testados algoritmos como Árvore de Decisão, Floresta Aleatória e Naive-Bayes. Os modelos tiveram desempenho semelhante, com a Floresta Aleatória atingindo 97% de precisão e sensibilidade para identificar saudáveis. No entanto, a classificação de HA + DCV foi desafiadora, com menor sensibilidade, possivelmente devido à ausência de diagnósticos formais e fatores como estilo de vida. Os resultados evidenciam a importância de dados mais detalhados e longitudinais para melhorar a identificação de doenças crônicas.*

1. Introdução

As Doenças Cardiovasculares (DCV) configuram-se como a principal causa de morte ao redor do mundo atualmente. 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 [WHO 2011].

A Hipertensão Arterial (HA) é também uma doença crônica, afetando negativamente o sistema cardíaco do indivíduo, sendo um dos principais problemas de saúde pública do 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 [Alwan 2011]. No Brasil, dados da Pesquisa Nacional de Saúde (PNS) de 2019 revelam que 23.9% dos adultos reportaram diagnóstico médico

positivo para HA [Malta et al. 2022], [Sousa and Zarate 2024] 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 [Stevens et al. 2018], [de Araújo et al. 2022].

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 [AlKaabi et al. 2020] utiliza dados de 987 pessoas 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 [Bhatt et al. 2023]. Este trabalho utilizou um conjunto de dados reais contendo 70.000 instâncias para desenvolver modelos preditivos que classificam a ocorrência de doenças cardiovasculares. Os pesquisadores empregaram algoritmos como Floresta Aleatória, Árvore de Decisão, 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.

Esta contribuição tem por objetivo descrever o perfil dos indivíduos que apresentam hipertensão e doenças cardiovasculares. O estudo aplica um processo de descoberta de conhecimento para identificar os principais fatores que caracterizam essa comorbidade na população brasileira.

Os modelos são construídos baseados em Árvore de decisão, Floresta aleatória e *Naive-Bayes* e são construídos considerando duas populações: a) indivíduos saudáveis, e b) indivíduos com presença da comorbidade. Para caracterizar a população brasileira, é considerada a mais recente Pesquisa Nacional de Saúde (PNS) do IBGE [IBGE 2020]. Esse estudo coletou dados sobre a saúde e estilos de vida da população brasileira no ano de 2019. O presente estudo busca revelar padrões para a comorbidade, esperando contribuir com um conhecimento mais contextualizado sobre essa comorbidade no Brasil.

2. Trabalhos Relacionados

A aplicação de técnicas de *machine learning* (ML) na predição de hipertensão arterial (HA) e doenças cardiovasculares (DCV) tem crescido. Estudos internacionais como o de [AlKaabi et al. 2020] usaram modelos como regressão logística para identificar fatores de risco para HA, enquanto outros, como [Bhatt et al. 2023] e [Gárate-Escamila et al. 2020], focaram em DCV, alcançando altas acurácias com algoritmos como XGBoost e Floresta Aleatória, respectivamente.

No cenário brasileiro, que utiliza dados da Pesquisa Nacional de Saúde (PNS), a abordagem é igualmente relevante. Por exemplo, [de Carvalho et al. 2024] também usou a PNS 2019 para diagnosticar hipertensão, obtendo um F1-Score de 75% com Floresta Aleatória e destacando fatores de risco semelhantes aos encontrados neste trabalho. De forma análoga, [Sousa and Zarate 2024] utilizou o mesmo conjunto de dados (PNS 2019) para caracterizar o perfil de indivíduos com Acidente Vascular Cerebral (AVC), reforçando o potencial da base de dados para estudos de saúde pública no Brasil.

Diferentemente desses trabalhos que focam em HA ou DCV (como o AVC) de forma isolada, o presente estudo avança ao investigar a comorbidade de hipertensão e doenças cardiovasculares (HA+DCV) simultaneamente. O objetivo não é apenas predizer, mas caracterizar o perfil desses indivíduos, preenchendo uma lacuna na compreensão dos fatores associados a este grupo de alto risco na população brasileira, com base nos dados da PNS.

3. Metodologia

No âmbito da metodologia aplicada, o estudo proposto seguiu uma série de etapas descritas no Fluxograma contido na Figura 1.

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 e 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 a 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.

3.2. Métodos

3.2.1. Entendimento do Problema

O entendimento do domínio do problema proposto é fundamental para o estudo, pois 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

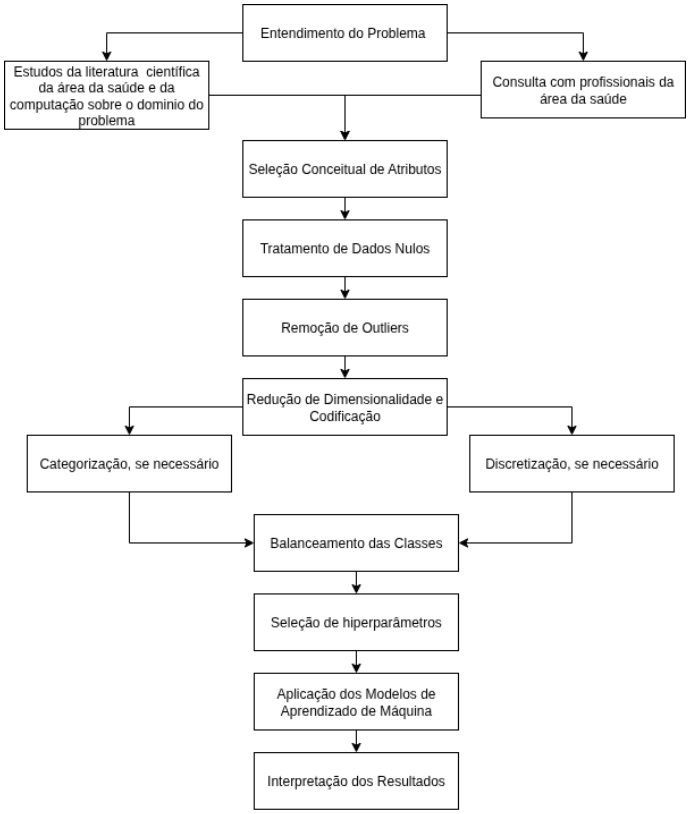

Figura 1. Fluxograma

com especialistas da área, nesse caso médicos, buscando uma compreensão clara e mais profunda dos aspectos que influenciam o problema em questão. Para isso, foi utilizado o método CAPTO [Gonçalves et al. 2024] que permite visualizar o problema e seu contexto com maior quantidade de perspectivas. Na Figura 2 é apresentado o mapa conceitual construído para Hipertensão Arterial e Doenças Cardiovasculares após a aplicação do método CAPTO.

3.2.2. Seleção Conceitual de Atributos

Após a montagem do mapa conceitual utilizando o método CAPTO, a próxima etapa consiste em selecionar conceitualmente os atributos identificados pelo mapa conceitual dentro da base de dados PNS 2019. Após a seleção, alguns deles sofreram o processo de renomeação dos nomes para melhorar a interpretabilidade durante a construção dos códigos e dos modelos, como descrito no quadro 1, os atributos renomeados são detalhados a seguir:

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

Tabela 1. Seleção Conceitual de Atributos da PNS 2019

Atributo	Descrição	Nome Renomeado
P00104	Peso em kg do indivíduo.	Peso
P00404	Altura em cm do indivíduo.	Altura
C006	Sexo do indivíduo.	Sexo
C008	Idade do indivíduo.	Idade
C009	Raça/etnia do indivíduo.	Raça_etnia
P02601	Consumo de sal do indivíduo.	Consumo_sal
P034	Nos últimos três meses praticou exercício físico ou esporte?	Atividades_fisicas
P035	Quantos dias por semana praticava exercício físico?	Freq_atividade_fisica
P050, P052, P05401, P05404, P05407, P05410, P05413, P05416, P05419	Quantidade fumada por dia/semana e histórico de tabagismo.	Categoria_tabagismo
P027	Com que frequência consome bebida alcoólica?	Frequencia_alcoolismo
P02801	Quantos dias por semana consome bebida alcoólica?	Qtd_alcool_semanal
P029	No dia em que bebe, quantas doses são consumidas?	Qtd_doses_alcoolicas
J037	Internação hospitalar nos últimos 12 meses?	Ficou_internado
I1001	Última consulta médica?	Ultima_consulta
I002	Autoavaliação da saúde.	Percepcao_estado_saude
I001	Possui plano de saúde?	Tem_plano
N004	Sente dor no peito ao subir escadas/ladeiras?	Cansa_subida
N005	Sente dor no peito ao caminhar normalmente?	Cansa_plano
D001	Sabe ler e escrever?	Alfabetizacao
VDD004A	Nível de instrução mais alto alcançado.	Escolaridade
V0026	Tipo de situação censitária.	Area_moradia
A01501	Forma de esgoto do domicílio.	Esgoto
A01601	Destino do lixo.	Destino_lixo
A01901	Possui acesso à internet?	Acesso_internet
A00501	Abastecimento de água do domicílio.	Abastecimento_agua
VDF004	Renda domiciliar per capita.	Faixa_salarial
Q03001	Diagnóstico de diabetes por médico?	Tem_diabetes

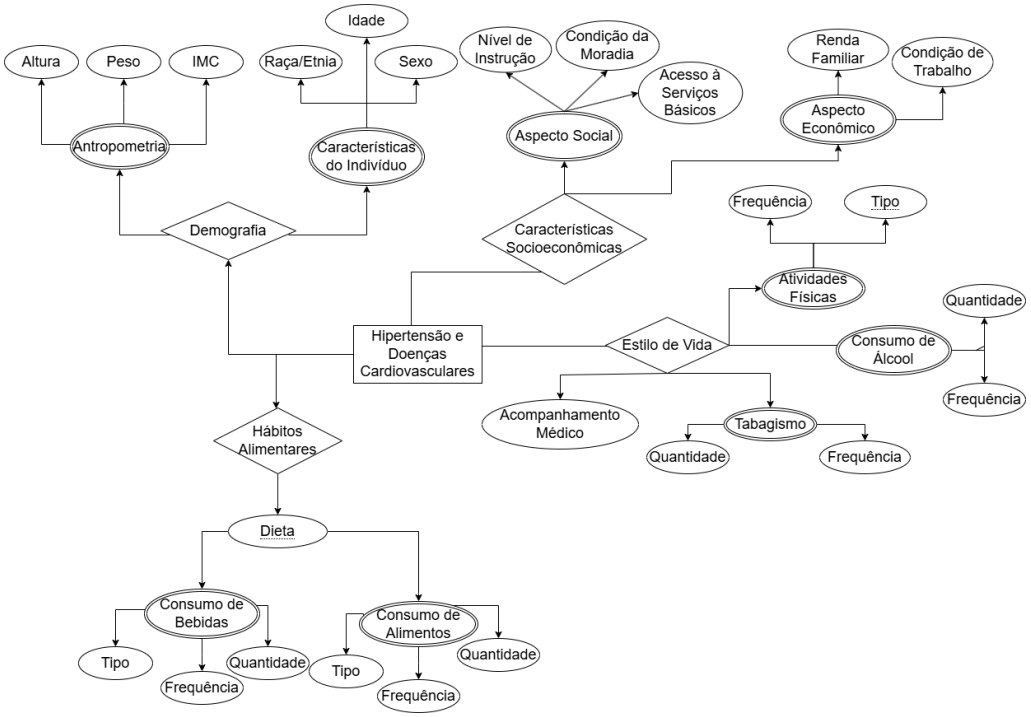

Figura 2. Mapa Conceitual - Método CAPTO

de idade. Ao final, a base de dados resultou em 45.757 instâncias e 36 atributos.

3.2.3. Tratamento de dados ausentes

Os dados ausentes foram tratados de maneira específica para cada Aspecto do modelo conceitual (Figura 2), com exceção de "Hábitos Alimentares". Primeiramente, no Aspecto Antropométrico, 575 registros (573 saudáveis e 2 hipertensos com DCV) foram removidos dos atributos "Peso"e "Altura", devido à inviabilidade de imputação via técnicas como KNN, já que a ausência de uma variável impossibilitava a estimativa confiável da outra. No Aspecto de Atividades Físicas, 26.358 valores nulos em "Freq_atividade_fisica"foram substituídos por 0, pois correspondiam a indivíduos que não praticaram exercícios ou esportes nos últimos três meses.

Para o Aspecto do Tabagismo, os atributos relacionados ao consumo atual (ex: P05401) tiveram 39.127 nulos imputados com 0, já que os indivíduos associados a esses registros declararam nunca ter fumado. No atributo P052 (histórico de tabagismo), 5.268 nulos foram preenchidos com 1 (fumantes diários no passado) e 655 com 2 (fumantes ocasionais no passado), conforme respostas vinculadas a outros indicadores. No Aspecto do Alcoolismo, 25.273 nulos em "Qtd_alcool_semanal e Qtd_doses_alcoolicas"receberam 0, correspondendo a indivíduos que nunca consumiram álcool. Os 6.136 nulos restantes, após análise por classe (média próxima de 0 dias/semana), também foram imputados com 0, considerando o consumo declarado de menor ou igual a 1 vez por mês.

No Aspecto de Acompanhamento Médico, 537 registros ausentes em Cansa_subida e "Cansa_plano", além de 3.746 em "Tem_diabetes", foram removidos por pertencerem à classe majoritária (saudáveis). Para o Aspecto Social, 560 nulos em "Es-

goto"foram preenchidos com 4 (fossa comum), baseado no perfil socioeconômico (60% com renda menor ou igual a 1 salário mínimo, residência rural). Por fim, no Aspecto Econômico, 13 instâncias com valores ausentes em "Faixa_salarial foram excluídas".

3.2.4. Remoção de Outliers

Em relação a 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 seguinte fórmula:

$$LimiteInferior = Q_1 - 1.5 \times IQR$$

$$LimiteSuperior = Q_3 + 1.5 \times IQR$$

$$IQR = Q_3 - Q_1$$

Para este estudo, foi utilizado o boxplot da biblioteca Seaborn do Python. Por padrão, este método 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 se encontram acima do limite superior ou abaixo do limite inferior.

Dentro do aspecto Antropométrico, a utilização do cálculo de *outliers*, para o atributo altura, baseado nos limites calculados a partir do IQR, encontra, de maneira geral, para ambos os sexos muitos possíveis *outliers*, mas nenhum valor que distorça muito a distribuição geral dos dados. Isso porque a diferença de alturas mantém a variabilidade dos dados originais. Portanto, foi utilizado o cálculo da multiplicação de 3 pelo IQR em vez de utilizar o limiar 1.5. Isso porque ao aumentar este limiar o processo de cálculo de *outliers* fica mais conservador e menos 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 [Yang et al. 2019].

Em relação ao Peso, o boxplot padrão 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.

Em relação ao consumo de álcool, foram encontradas instâncias que consomem acima de 15 doses de álcool por dia. De acordo com o estudo [National Institute on Alcohol Abuse and Alcoholism 2022], 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.

3.2.5. Redução de Dimensionalidade

Esta etapa tem como objetivo reduzir a dimensionalidade de atributos que compartilham da mesma informação para o problema. Por exemplo, foi aplicado no aspecto de tabagismo, onde os atributos relacionados ao tipo de produto de tabaco consumido e a frequência de consumo foram agrupados em categorias que relacionam essas duas características, construindo os seguintes grupos: Fuma Muito, Fuma Razoavelmente, Fuma Pouco e Não Fuma. Portanto, 9 atributos relacionados a tabaco foram unidos e categorizados para o atributo chamado de *Categoria_tabagismo* mostrado na Tabela 2.

Tabela 2. Categorias de Tabagismo e Condições Lógicas

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) (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) (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) (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).

Para descrever a característica de alcoolismo dos indivíduos, também foi criada uma codificação de fusão 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 atributo, chamado de *categoria_alcoolismo* (ver Tabela 3).

Tabela 3. Categorias de Alcoolismo e Condições Lógicas

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 do alcoolismo = 2.
Bebedor moderado	(Quantidade de doses alcoólicas $\geq$ 1 e $\leq$ 2) ou Quantidade semanal de álcool $\leq$ 3 e Frequência do alcoolismo = 3.
Bebedor frequente	(Quantidade de doses alcoólicas $>$ 2) ou Quantidade semanal de álcool $\leq$ 4 e Frequência do alcoolismo = 4.
Bebedor excessivo	Quantidade de doses alcoólicas $>$ 3 ou Quantidade semanal de álcool $>$ 4.

Houve também a categorização do atributo “Racas_Etnia”s entre: Brancos, Pretos e Pardos, sendo que as raças/etnias minoritárias como, amarelos, indígenas 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 a desenvolverem a hipertensão arterial em maior grau em relação às pessoas brancas [Zilbermint et al. 2019], [Sousa et al. 2022].

Por último, foi necessário 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 [Powell-Wiley et al. 2021]. Após o cálculo do IMC, foi feita a categorização desses valores em 4 classes dispostas em: Baixo Peso, Peso Ideal, Sobrepeso e Obeso de acordo com a OMS [WHO 2021], ver Tabela 4.

Tabela 4. Classificação do IMC conforme diretrizes da OMS

Categoria	Intervalo de IMC (kg/m²)
Baixo Peso	$< 18{,}5$
Peso Ideal	18,5 – 24,9
Sobrepeso	25,0 – 29,9
Obeso	$\geq 30{,}0$

3.2.6. 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 para doença) e 2.369 pertencendo à classe das pessoas com HA e DCV. Diante do desbalanceamento das classes, foi utilizada a técnica de subamostragem aleatória (Random_Under_Sampling) que reduziu a classe majoritária e equalizou aleatoriamente em relação à classe minoritária. Como resultado, foram selecionadas 2.369 instâncias aleatórias das 30.005 instâncias pertencentes à classe dos Saudáveis, totalizando um conjunto de treinamento de 4.738 instâncias, com 25 atributos independentes, acrescentado do atributo *classe* = $\{$*Saudável, Hipertensão+Doenças_Cardiovasculares*$\}$.

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. A base de dados encontra-se no link: https://github.com/licapLaboratory/DataBase-PNS-Hipertensao-Cardio

3.2.7. Aplicação dos Modelos de Aprendizado de Máquina

A distinção entre modelos caixa-preta e caixa-branca (interpretáveis) é discutida em [Loyola-González 2019]. Com base nisso, para o escopo deste estudo foram escolhidos os modelos de Árvore de Decisão, Floresta Aleatória e Naive Bayes por suas características complementares: a Árvore de Decisão foi selecionada pela alta interpretabilidade; a Floresta Aleatória pelo seu robusto desempenho preditivo; e o Naive Bayes como um *baseline*

computacionalmente eficiente para dados categóricos.

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 e métricas de avaliação

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. Ambos os modelos obtiveram uma acurácia muito similar, com uma acurácia de aproximadamente 86%. Os hiperparâmetros encontrados para a Árvore de Decisão foram *max_depth* = 10, *min_samples_split* = 10, *min_samples_leaf* = 4, *max_features* = 0.4 e o critério de divisão escolhido foi a entropia. Já a Floresta Aleatória utilizou dos mesmos hiperparâmetros citados e seus valores acrescidos de *min_samples_split* = 10, *n_estimators* = 150 e *n_jobs* = -1, permitindo o uso de todos os núcleos do processador durante o treinamento e acelerando a execução.

Ademais, para os modelos foi utilizado o método de validação cruzada com *k-folds* = 10. Como métricas de avaliação, foram utilizadas a Precisão, definida como a razão entre os verdadeiros positivos e a soma de verdadeiros positivos e falsos positivos, a Sensibilidade, calculada como a razão entre os verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos, e o F1-Score, que é a média harmônica entre Precisão e Sensibilidade.

4.2. Desempenho dos Modelos e Discussão dos Resultados

Primeiramente, no que tange à classe das pessoas com hipertensão (HA) e doenças cardiovasculares (DCV), a Tabela 5 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. Esse resultado indica que cerca de 4 em 10 casos podem não ser corretamente classificados, limitando a efetividade dos modelos para a triagem preventiva e sugere desafios importantes na identificação automatizada desse grupo, especialmente considerando a necessidade de intervenções precoces em saúde pública.

Tabela 5. Desempenho dos modelos na classificação de pessoas com HA + DCV

Algoritmo	Precisão	Sensibilidade	F1-Score
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

Por outro lado, em relação à classe das pessoas consideradas saudáveis, isto é, as pessoas que não possuem o diagnóstico positivo das doenças crônicas estudadas, a Tabela 6 contém as métricas de classificação e seus valores obtidos. Os três modelos apresentaram desempenho significativamente superior, com precisão e sensibilidade próximas de 97% e 98%, evidenciando alta capacidade de identificação de casos sem doenças crônicas.

Tabela 6. Desempenho dos modelos na classificação de pessoas saudáveis

Algoritmo	Precisão	Sensibilidade	F1-Score
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

Observando os resultados das Tabelas 5 e 6, observa-se que os modelos (baseado em Árvore de decisão, Floresta aleatória e Naive-Bayes) apresentam desempenho semelhante nas classes HA+DCV e saudáveis, porém com maior assertividade para saudáveis (97% de precisão e sensibilidade), sugerindo possível viés para essa classe. A similaridade entre as instâncias, devido à natureza descritiva (não clínica) da base PNS 2019, pode explicar sobreposições: indivíduos classificados erroneamente como saudáveis podem ter doenças não diagnosticadas, enquanto os classificados como HA+DCV podem estar em controle clínico.

A Árvore de Decisão e a Floresta Aleatória destacaram-se levemente na classe HA+DCV (65% precisão, 58% sensibilidade), enquanto o *Naive-Bayes* teve a menor sensibilidade (54%), crítica em contextos que exigem diagnóstico precoce. Esses resultados reforçam a necessidade de priorizar métricas como sensibilidade em saúde pública.

Na análise de importância das variáveis, a Faixa_Etaria_Numerica foi a mais relevante, corroborando a relação entre idade e risco de doenças crônicas. Percepcao_estado_saude, Cansa_subida, Cansa_plano e Tem_diabetes também se destacaram, indicando que sintomas subjetivos e comorbidades associadas são indicativos precoces. A variável Ultima_consulta reforça a importância do acompanhamento médico regular para prevenção. A interpretação contextual dos dados é essencial, especialmente em estudos populacionais que buscam traçar perfis de saúde baseados em fatores socioeconômicos e comportamentais.

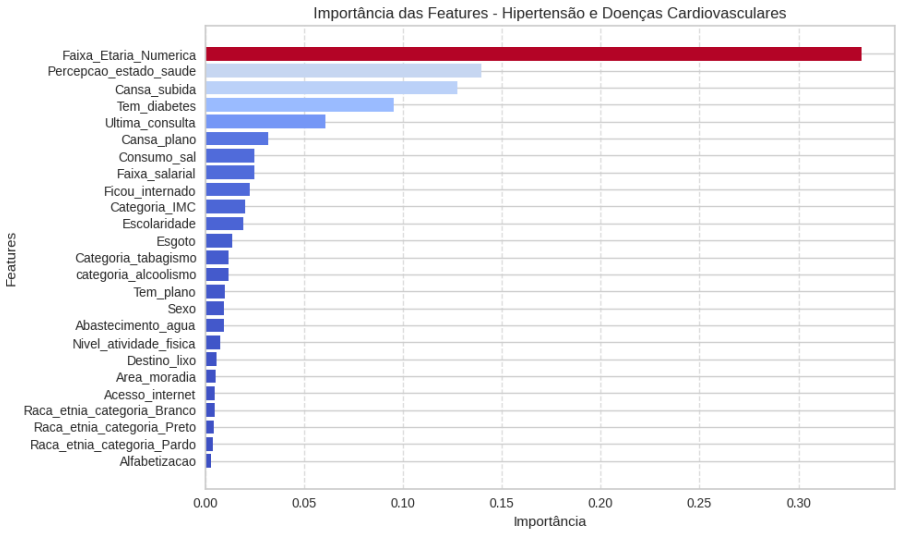

Figura 3. Feature Importance dos Atributos Utilizados

5. Conclusões

Os modelos de aprendizado de máquina avaliados neste estudo demonstraram eficácia na classificação de indivíduos com características saudáveis, com precisão e sensibilidade de

97%. No entanto, para a classe de indivíduos com hipertensão e doenças cardiovasculares (HA + DCV), os modelos apresentaram sensibilidade reduzida, em torno de 60%, o que sugere desafios significativos na detecção desse grupo. Essa limitação pode estar associada à ausência de diagnósticos formais ou à influência de fatores comportamentais, como a adoção de hábitos saudáveis que diminuem os sintomas, destacando a necessidade de modelos mais robustos que integrem variáveis temporais (histórico médico longitudinal) e contextuais (mudanças no estilo de vida).

A análise de importância das variáveis revelou que atributos como faixa etária, percepção subjetiva de saúde, cansaço, desconforto torácico ao esforço e diagnóstico prévio de diabetes foram determinantes para a classificação, corroborando evidências científicas sobre fatores de risco cardiovascular. Estes achados, por sua vez, oferecem incentivos para ações práticas de saúde pública: a relevância da idade reforça a necessidade de rastreamento ativo em populações mais velhas, enquanto a importância de sintomas autorrelatados sugere o potencial de questionários simples como ferramentas de triagem inicial de baixo custo. A natureza descritiva da base de dados, contudo, limitou a capacidade de identificar indivíduos assintomáticos, o que reforça a importância de complementar dados populacionais com informações clínicas detalhadas e acompanhamento temporal para capturar a complexidade das doenças crônicas.

Para avançar na precisão dos modelos, especialmente na classificação de HA + DCV, futuros estudos devem priorizar a incorporação de dados longitudinais como exames médicos seriados, monitoramento de hábitos e marcadores genéticos, além de explorar técnicas como o aprendizado profundo para modelar interações não lineares entre variáveis. A validação em bases de dados com maior diversidade socioeconômica e detalhamento diagnóstico também é crucial para garantir a generalização dos resultados. Por fim, o estudo ressalta que a modelagem preditiva em saúde pública exige uma abordagem integrada, combinando dados quantitativos, contextos clínicos e fatores comportamentais, a fim de enfrentar os desafios impostos pelas doenças crônicas de forma mais eficaz.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e da Fundação de Ampara à Pesquisa do Estado de Minas Gerais (FAPEMIG), Processo No xxxx/2025.

Referências

AlKaabi, L., Ahmed, L., Al Attiyah, M., and Abdel-Rahman, M. (2020). Predicting hypertension using machine learning: Findings from qatar biobank study. *PLOS ONE*, 15(10):e0240370.

Alwan, A. (2011). *Global status report on noncommunicable diseases 2010*. World Health Organization, Geneva. 176 pp.

Bhatt, C., Patel, P., Ghetia, T., and Mazzeo, P. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2):88.

de Araújo, J., de Alencar Rodrigues, R., da Costa Pereira de Arruda Neta, A., et al. (2022). The direct and indirect costs of cardiovascular diseases in brazil. *PLOS ONE*, 17(12):e0278891.

de Carvalho, N., Gomes, M., and Zárate, L. (2024). Mineração de dados no diagnóstico de hipertensão baseado na pesquisa nacional em saúde 2019. *J Health Inform*, 16(Especial).

Gárate-Escamila, A., El Hassani, A., and Andrès, E. (2020). Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, 19:100330.

Gonçalves, L., Franca, D., and Zarate, L. (2024). Relevância do entendimento do domínio de problema na construção de modelos computacionais de aprendizado. In *Anais do XVIII Brazilian e-Science Workshop*, pages 135–142, Porto Alegre, RS, Brasil. SBC.

IBGE (2020). Pesquisa nacional de saúde 2019 - instituto brasileiro de geografia e estatística. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=25921&t=resultados`. Acesso em: 2024-07-15.

Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Malta, D. et al. (2022). Hipertensão arterial e fatores associados: Pesquisa nacional de saúde, 2019. *Revista de Saúde Pública*, 56:122.

National Institute on Alcohol Abuse and Alcoholism (2022). Standard alcohol guidelines.

Powell-Wiley, T., Poirier, P., Burke, L., et al. (2021). Obesity and cardiovascular disease: A scientific statement from the american heart association. *Circulation*, 143(21):e84–e118.

Sousa, C., Ribeiro, A., Barreto, S., et al. (2022). Diferenças raciais no controle da pressão arterial em usuários de anti-hipertensivos em monoterapia: resultados do estudo elsa-brasil. *Arq. Bras. Cardiol.*, 118(3):614–622.

Sousa, M. and Zarate, L. (2024). A epidemia silenciosa: Explorando os determinantes comportamentais e socioeconômicos da deficiência renal crônica no brasil. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 318–327, Porto Alegre, RS, Brasil. SBC.

Stevens, B., Pezzullo, L., Verdian, L., Tomlinson, J., George, A., and Bacal, F. (2018). The economic burden of heart conditions in brazil. *Arq. Bras. Cardiol.*, 111(1):29–36.

WHO (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization, Geneva.

WHO (2021). Obesity and overweight.

Yang, J., Rahardja, S., and Fränti, P. (2019). Outlier detection: how to threshold outlier scores? In *Proc. of the Int. Conf. on Artificial Intelligence, Information Processing and Cloud Computing*, pages 37–42.

Zilbermint, M., Hannah-Shmouni, F., and Stratakis, C. (2019). Genetics of hypertension in african americans and others of african descent. *Int. J. Mol. Sci.*, 20(5):1081.