Caracterização de Perfis Clínicos: Aplicação de Algoritmos de Aprendizado de Máquina na Identificação de Hipertensos com Doenças Cardiovasculares

Gustavo Costa
G.Costa*
gustavocosta.ds09@gmail.com
PUC Minas
Belo Horizonte, Minas Gerais

ABSTRACT

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.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

KEYWORDS

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

ACM Reference Format:
Gustavo Costa and G.Costa. 2018. Caracterização de Perfis Clínicos: Aplicação de Algoritmos de Aprendizado de Máquina na Identificação de Hipertensos com Doenças Cardiovasculares. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

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 MATERIAIS E MÉTODOS

2.1 Descrição da Base de Dados

A base de dados utilizada neste presente estudo foi a Pesquisa Nacional de Saúde (PNS) 2019, nela há informações importantes sobre hábitos de vida, aspectos sociodemográficos, por exemplo, a frequência de álcool consumida por um indivíduo, acesso à saneamento básico, doenças crônicas, entre outros. Na base disponibilizada pelo Ministério da Saúde, inicialmente, há uma representatividade de todos os estados brasileiros na amostra coletada. Para mais, a PNS possui 293.726 registros e 1.088 atributos, dentre estes atributos, há uma subdivisão em 26 módulos diferentes de questões. Essa base de dados é fonte de informação para a produção de diversos estudos e pesquisas científicas e contribui para a implementação de políticas públicas de saúde brasileiras.

Ainda, foram utilizadas as instâncias que representam apenas as pessoas com, no mínimo, 18 anos de idade. Além disto, a PNS 2019 foi distribuída em duas classes:

- **Pessoas Sem Diagnóstico:** indivíduos que responderam não para o diagnóstico das comorbidades hipertensão arterial e doenças cardiovasculares e que responderam diferente de sim para as outras doenças crônicas presentes no questionário.
- **Pessoas Hipertensas com Doenças Cardiovasculares:** indivíduos que responderam sim para o diagnóstico de hipertensão arterial e responderam sim para o diagnóstico de doenças cardiovasculares, essa classe pode ser entendida também com a sigla HA + DCV.

Para as três classes, a quantidade de registros são, respectivamente, 42.742, 20.804 e 3.015. Após os tratamentos descritos, a base de dados converteu-se em 66.561 instâncias com 1.088 atributos. O estudo [] propõe o método CAPTO que consiste em realizar uma seleção conceitual de atributos baseando-se em conhecimento explicíto (fontes literárias) e conhecimento tácito (especialistas do assunto do problema) para realizar uma seleção conceitual de features, contribuindo para a redução da dimensionalidade da base de dados e mantendo informações cruciais para a construção de bons modelos. À face do exposto, este método foi aplicado na base de dados resultando na pré-seleção dos atributos listados a seguir:

- **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 domicílio?
- **VDF004:** Faixa de rendimento domiciliar per capita.
- **Q03001:** Algum médico já lhe deu o diagnóstico de diabetes?
- **Condicao:** Variável alvo contendo os rótulos das classes.

De acordo com a lista supracitada, há 36 atributos selecionados com a aplicação do método CAPTO.

2.2 Descrição dos Atributos

Para a descrição dos atributos, primeiramente, foi realizada uma renomeação de alguns dos 36 atributos pré-selecionados para melhorar a interpretabilidade. É importante ressaltar que nem todos foram renomeados, isso porque alguns foram pré-selecionados com o intuito de sofrerem uma fusão, mas também serão descritos com seus nomes originais da PNS. Confira na tabela 1 a renomeação desses atributos.

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

Table 1: Atributos que sofreram renomeação

Em relação à descrição de cada atributo, confira a listagem individual à seguir:

- **Sexo**: Este atributo indica o gênero da pessoa entrevistada, sendo codificado como um valor categórico. Os valores possíveis são:
  - 1: Homem
  - 2: Mulher
- **Idade**: A idade da pessoa é um valor numérico nominal, representando a quantidade de anos de vida. A idade mínima registrada é de 18 anos e a máxima é de 100 anos.

- **Altura**: Refere-se à altura do indivíduo em centímetros. É um valor numérico nominal, com registros variando de 128 cm até 200 cm.
- **Peso**: Indica o peso corporal do indivíduo em quilogramas, também um valor numérico nominal. O menor valor registrado é 25 kg, enquanto o maior é 170 kg.
- **Raça_Etnia**: Esse atributo representa a raça ou etnia com a qual o entrevistado se identifica, categorizado da seguinte forma:
  - 1: Branco
  - 2: Preto
  - 3: Amarelo
  - 4: Pardo
  - 5: Indígena
  - 9: Ignorado
- **Consumo_Sal**: Esse atributo avalia o nível de consumo de sal do indivíduo, em uma escala ordinal:
  - 1: Muito Alto
  - 2: Alto
  - 3: Adequado
  - 4: Baixo
  - 5: Muito Baixo
- **Atividades_fisicas**: Informa se o indivíduo praticou alguma atividade física nos últimos três meses, com respostas categóricas:
  - 1: Sim
  - 2: Não
- **Freq_atividade_fisica**: Diz respeito a quantos dias o indivíduo se exercita na semana, é um atributo numérico ordinal e tem os seguintes valores:
  - 0: zero dias na semana.
  - 1: um dia na semana.
  - 2: dois dias na semana.
  - 3: três dias na semana.
  - 4: quatro dias na semana.
  - 5: cinco dias na semana.
  - 6: seis dias na semana.
  - 7: sete dias na semana.
- **Freq_alcoolismo**: Atributo cujo indivíduo informa quantas vezes ele consome álcool no mês.
  - 1: não bebo nunca.
  - 2: menos de uma vez por mês.
  - 3: uma vez ou mais por mês.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Qtd_alcool_semanal**: Este atributo informa quantos dias na semana o indivíduo consome bebida alcoólica. É categórico e possui uma ordem.
  - 1 a 7: dias na semana.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Qtd_doses_alcoolicas**: Este atributo pergunta para a pessoa que bebe quantas doses na ocasião ela consome. É categórica e ordinal.
  - 1 a 98: doses consumidas na ocasião.
  - 99: ignorado.
  - Nulo: não foi aplicada a questão.
- **Ficou_internado**: Este atributo pergunta para o indivíduo se ele foi internado por 24 horas ou mais nos últimos 12 meses. É uma variável categórica sem ordem.
  - 1: Sim.
  - 2: Não.
  - Nulo: não foi aplicada a questão.
- **Ultima_consulta**: Este atributo questiona o entrevistado quanto tempo faz desde a última consulta médica realizada na data do questionário, configura-se como categórico e possui uma ordem temporal.
  - 1: até 15 dias.
  - 2: mais de 15 dias até 1 mês.
  - 3: mais de 1 mês até 6 meses.
  - 4: mais de 6 meses até 1 ano.
  - 5: há mais de 1 ano.
  - 9: ignorado
  - Nulo: não foi aplicada a questão.
- **Percepcao_estado_saude**: Este atributo interroga o indivíduo sobre a percepção própria do estado de saúde. É um atributo categórico e ordinal que mede a qualidade da saúde.
  - 1: muito boa.
  - 2: boa.
  - 3: regular.
  - 4: ruim.
  - 5: muito ruim.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Tem_plano**: É um atributo categórico nominal, o objetivo é questionar o entrevistado se ele possui ou não algum tipo de plano de saúde.
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Cansa_subida**: É um atributo categórico nominal,a finalidade é saber se o indivíduo se cansa ao subir ladeiras, escadas e se sente dor ou algum desconforto no peito.
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Cansa_plano**: É um atributo categórico nominal,a finalidade é saber se o indivíduo se cansa ao caminhar

em lugares planos em velocidade normal e se sente dor ou algum desconforto no peito.
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Alfabetizacao**: Este atributo avalia se a pessoa possui alfabetização básica (sabe ler e escrever), com respostas categóricas e sem ordem:
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Escolaridade**: Indica o nível de instrução do entrevistado, sendo um valor ordinal e categórico disposto nas seguintes categorias:
  - 1: sem instrução.
  - 2: ensino fundamental incompleto.
  - 3: ensino fundamental completo.
  - 4: ensino médio incompleto.
  - 5: ensino médio completo.
  - 6: ensino superior incompleto.
  - 7: ensino superior completo.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Area_moradia**: Atributo que refere-se à localidade da residência do entrevistado, é categórico e nominal, disposto nas seguintes opções:
  - 1: urbano.
  - 2: rural.
- **Esgoto**: Atributo de escopo social que pergunta ao entrevistado qual é o destino dado ao esgoto da residência, é categórico não ordinal com as respostas:
  - 1: rede geral de esgoto ou pluvial.
  - 2: fossa séptica ligada à rede.
  - 3: fossa séptica não ligada à rede.
  - 4: fossa rudimentar.
  - 5: vala.
  - 6: rio, lago, córrego ou mar.
  - 7: outra.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Destino_lixo**: Atributo que informa qual é o destino dado ao lixo produzido na residência, também é um atributo categórico e não ordinal com as seguintes opções de resposta:
  - 1: coleta diretamente por serviço de limpeza.
  - 2: coletado em caçamba de serviço de limpeza.
  - 3: queimado (na propriedade).
  - 4: enterrado (na propriedade).
  - 5: jogado em terreno baldio ou logradouro.
  - 6: outro.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Acesso_internet**: Atributo que questiona se algum morador da residência possui acesso à internet, é categórico e sem ordem:
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não aplicável.
- **Abastecimento_agua**: Atributo que informa qual é a forma de abastecimento da residência, também é categórico e não possui uma ordem:
  - 1: rede geral de distribuição.
  - 2: poço profundo ou artesiano.
  - 3: poço raso, freático ou cacimba.
  - 4: fonte ou nascente.
  - 5: água da chuva armazenada.
  - 6: outra.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Faixa_salarial**: Atributo que informa qual é a forma de abastecimento da residência, também é categórico e não possui uma ordem:
  - 1: rede geral de distribuição.
  - 2: poço profundo ou artesiano.
  - 3: poço raso, freático ou cacimba.
  - 4: fonte ou nascente.
  - 5: água da chuva armazenada.
  - 6: outra.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **Tem_diabetes**: Atributo que pergunta ao entrevistado se ele já recebeu o diagnóstico positivo para diabetes:
  - 1: sim.
  - 2: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **P050**: Pergunta ao indivíduo entrevistado se ele fuma atualmente e qual a frequência, sendo categorizado ordinalmente nas seguintes opções de resposta:
  - 1: sim, diariamente.
  - 2: sim, menos que diariamente.
  - 3: não fumo atualmente.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.
- **P052**: Pergunta se o indivíduo já fumou anteriormente, sendo categorizado ordinalmente em:
  - 1: sim, diariamente.
  - 2: sim, menos que diariamente.
  - 3: não.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.

- **P05401, P05404, P05407, P05410, P05410, P05413, P05416, P05419**: Este atributo coleta dados sobre a frequência de uso de produtos específicos de tabaco (cigarros, narguilé, entre outros), com as seguintes opções ordinais:
  - 1: um ou mais por dia.
  - 2: um ou mais por semana.
  - 3: menos que uma vez por semana.
  - 4: menos que uma vez por mês.
  - 5: não fuma esse produto.
  - 9: ignorado.
  - Nulo: não foi aplicada a questão.

2.3 Etapas da Metodologia Aplicada

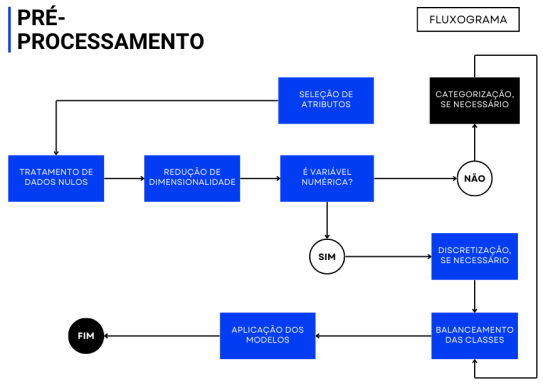

Figure 1: Fluxograma Pré-processamento

Para o pré-processamento dos dados, a figura 1 ilustra o fluxograma das etapas seguidas para melhorar a qualidade dos dados até a aplicação dos modelos.

2.3.1. Seleção de Atributos: Na primeira etapa, foram selecionados os atributos da PNS 2019 supracitados neste estudo de acordo com o método CAPTO [16] para a base de dados pré-selecionada, ao todo são 36 atributos incluindo a variável target para o problema de classificação.

2.3.2. Tratando 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 domínio do problema, exceto pelo aspecto de hábitos alimentares devido sua não utilização no pré- processamento e no modelo. Em primeiro lugar, houve a remoção de 575 instâncias nos atributos Peso e Altura, subdivididos 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 à 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_plan 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.

2.3.3. 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 seguinte fórmula:

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

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

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 IQR = $Q_3 - Q_1$.

Valores menores que o Limite Inferior ou maiores que o Limite Superior são considerados *outliers*.

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 1.5 pelo IQR. Dessa forma, são considerados Outliers aqueles valores que encontram-se acima do limite superior ou abaixo do limite inferior.

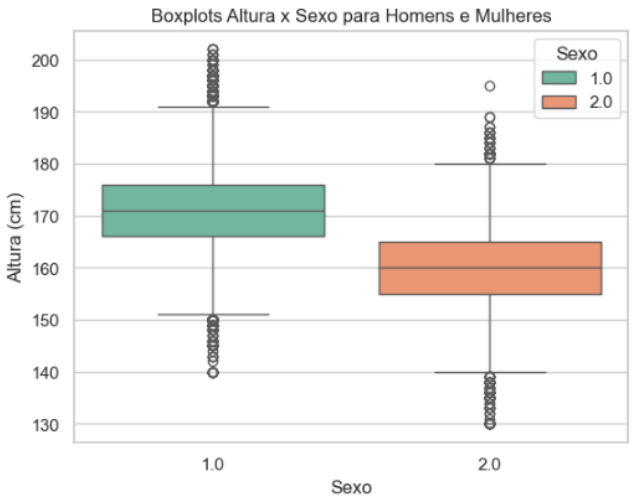

Figure 2: Boxplot Altura x Sexo

Como é observado na figura 2, a utilização do cálculo de Outliers baseado nos limites calculados a partir do IQR encontram, de maneira geral, para os dois sexos Homem (1) 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 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 3) 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.

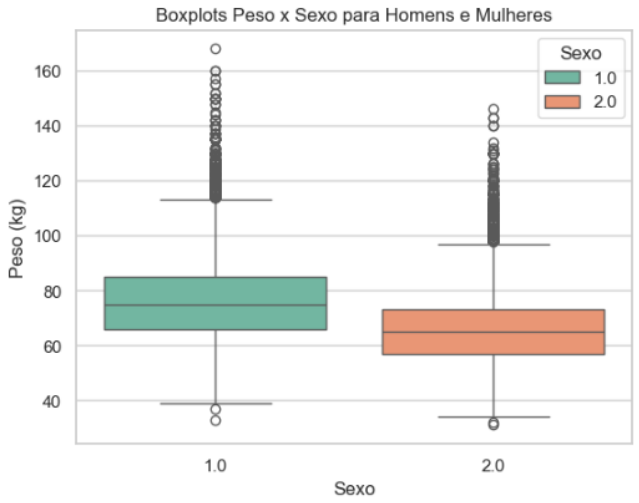

Figure 3: Boxplot Peso x Sexo

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.

2.3.4. Redução de Dimensionalidade e Codificação: 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 o 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 figura 4 apresenta a categoria de tabagismo e sua lógica.

Table 2: Categorias de Tabagismo e suas 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).

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

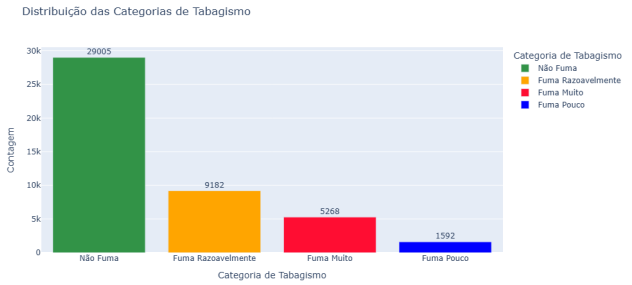

Figure 4: Categorias do tabagismo

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 (tabela 3), dessa forma, os atributos Qtd_doses_alcoolicas, Frequencia_alcoolismo, Qtd_alcool_semanal foram fundidos e categorizados para apenas um, chamado de categoria_alcoolismo (figura 5)

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 ≥ 1 e ≤ 2) ou Quantidade semanal de álcool ≤ 3 e Frequência do alcoolismo = 3
Bebedor frequente	(Quantidade de doses alcoólicas > 2) ou Quantidade semanal de álcool ≤ 4 e Frequência do alcoolismo = 4
Bebedor excessivo	Quantidade de doses alcoólicas > 3 ou Quantidade semanal de álcool > 4

Table 3: Categorias de Alcoolismo e suas Condições Lógicas.

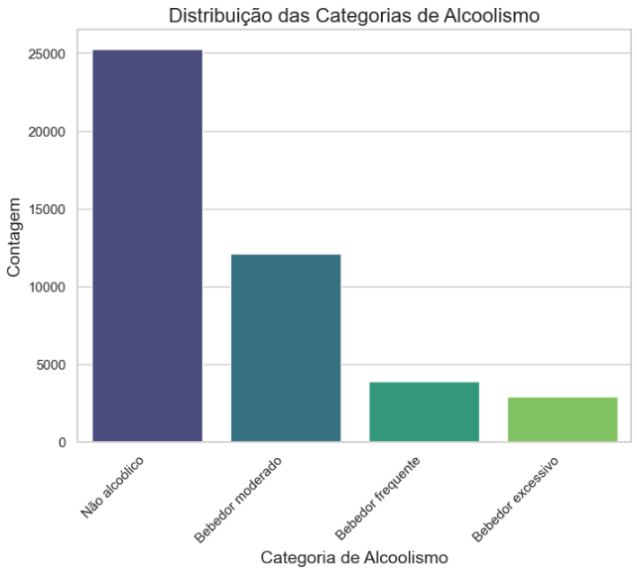

Figure 5: Categorias do alcoolismo

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 6 e 7 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.

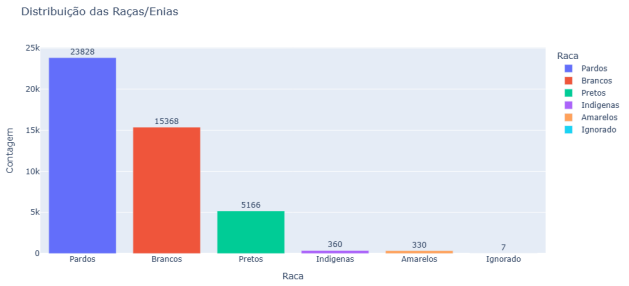

Figure 6: Categorias Raças/Etnias Originais

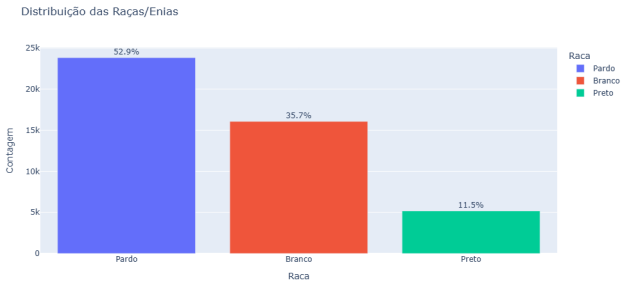

Figure 7: Categorias Raças/Etnias Formulada

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 ≤ 24.9
Sobrepeso	$25 \leq$ IMC ≤ 29.9
Obeso	IMC ≥ 30

Table 4: Categorias do IMC e suas condições lógicas.

Com a categorização do IMC aplicada, foi criado um novo atributo chamado de Categoria_IMC, disposto na figura 8 que representa um histograma:

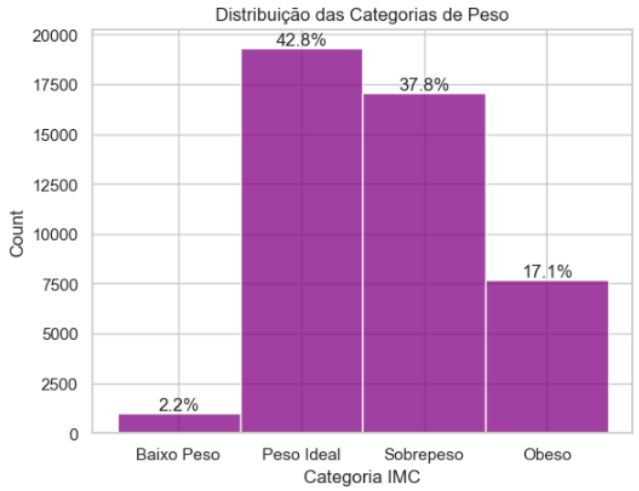

Figure 8: Categorias Raças/Etnias Formulada

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.

2.3.5. 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.

2.3.6. 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.

2.3.7. 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:

- **Min_samples_split**: Número mínimo de amostras necessárias para dividir um nó. Neste caso, definido como 10. Valores maiores impedem a criação de nós muito pequenos, o que reduz o overfitting, mas pode limitar a complexidade do modelo.
- **Min_samples_leaf**: Número mínimo de amostras que um nó folha deve conter. Configurado como 4, garante que os nós finais não sejam baseados em uma quantidade muito pequena de dados, melhorando a generalização.
- **Max_features**: Proporção das características totais consideradas ao procurar a melhor divisão. Um valor de 0.4 significa que 40% das variáveis disponíveis serão

avaliadas, o que introduz uma leve aleatoriedade e reduz o overfitting.
- **Max_depth**: Profundidade máxima da árvore. Com limite de 10, evita que a árvore cresça excessivamente e se ajuste demais aos dados de treino (overfitting), mas ainda mantém uma boa capacidade de aprendizado.
- **Criterion**: Métrica usada para avaliar a qualidade de uma divisão. Configurada como `entropy`, que é baseada na teoria da informação. Essa métrica mede o grau de desordem (ou incerteza) nos dados e busca minimizá-la em cada divisão.

Já para o modelo caixa-preta Classificador Neural MLP, os hiperparâmetros utilizados foram:

- **hidden_layer_sizes**: Define a arquitetura da rede neural, especificando o número de neurônios em cada camada oculta. O valor `(100, 100, 100)` indica que a rede possui três camadas ocultas, cada uma com 100 neurônios, permitindo a modelagem de relações complexas nos dados.
- **activation**: Função de ativação para os neurônios. Configurada como `relu` (Rectified Linear Unit), que introduz não-linearidade no modelo e ajuda a evitar o problema de gradientes desaparecendo.
- **alpha**: Parâmetro de regularização L2. Definido como 0.0001, evita overfitting ao penalizar pesos muito grandes na rede.
- **learning_rate**: Determina como a taxa de aprendizado varia durante o treinamento. Configurada como `adaptive`, ajusta automaticamente a taxa de aprendizado, diminuindo-a quando a performance no conjunto de validação para de melhorar.
- **solver**: Algoritmo usado para otimizar os pesos da rede. Configurado como `adam` (Adaptive Moment Estimation), que é eficiente e bem adequado para grandes bases de dados e redes com muitas variáveis.
- **max_iter**: Número máximo de iterações durante o treinamento. Definido como 200, garante que o modelo tenha tempo suficiente para convergir antes de interromper.

2.4 Métricas de avaliação de qualidade

Os modelos de aprendizado de máquina foram avaliados por meio das seguintes métricas de avaliação:

- Precision: que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo, indicando a *confiabilidade* dos resultados positivo.

$$\text{Precision} = \frac{TP}{TP + FP}$$

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

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

3 RESULTADOS E DISCUSSÕES

Primeiramente, no que tange à classe das pessoas com hipertensão (HA) e doenças cardiovasculares (DCV) que representam a Classe 0, a figura 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

Figure 9: Métricas de classificação dos modelos classe 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 figura 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

Figure 10: Métricas de classificação dos modelos classe saudáveis

Com a observação dos resultados presentes nas figuras 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 uma 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, onde 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:

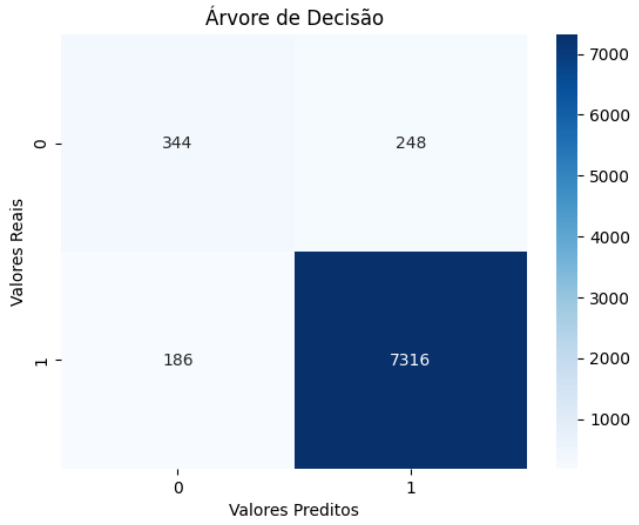

Figure 11: Matriz de confusão Árvore de Decisão

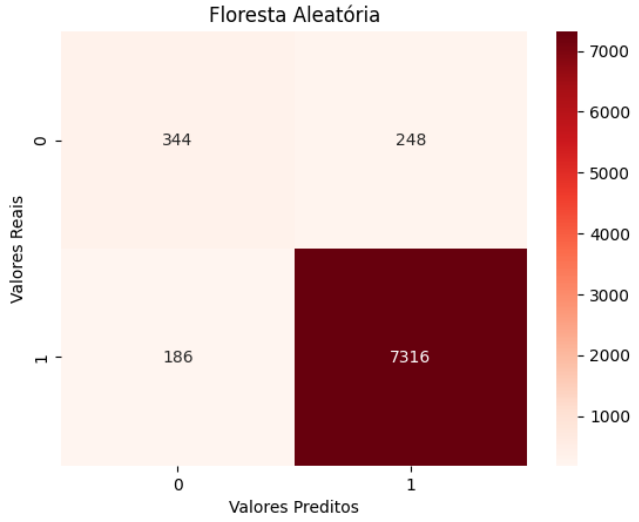

Figure 12: Matriz de confusão Floresta Aleatória

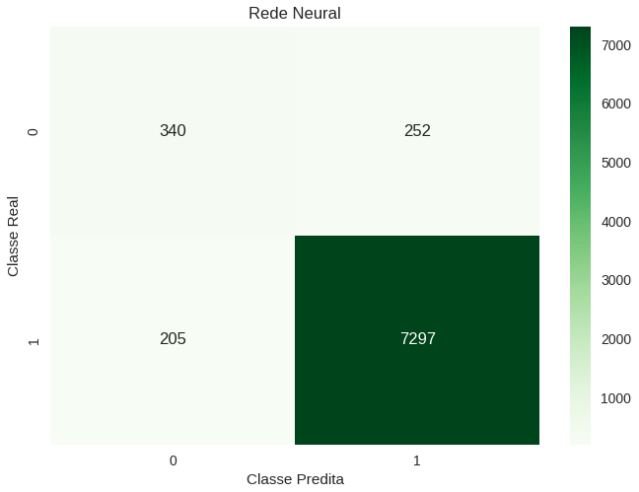

Figure 13: Matriz de confusão Rede Neural Classificadora MLP

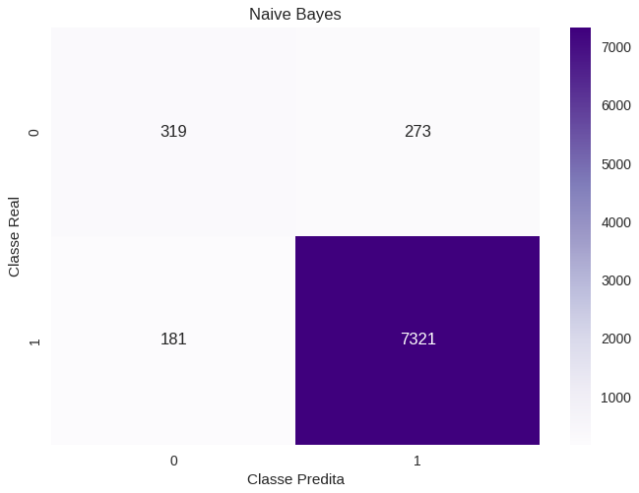

Figure 14: Matriz de confusão Naive Bayes

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.

4 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.

5 UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL

Foi utilizada a ajuda do chat GPT da OpenAI para ajustar as tabelas e figuras na sintaxe LaTeX. Somado à isso, houve a utilização do Gemini do Google para a criação das fórmulas das métricas de classificação e das codificações realizadas na etapa de pré-processamento em formato LaTeX também, como a fórmula do IMC em LaTeX.

6 CÓDIGO DESENVOLVIDO

Notebooks GitHub: Gustavo Costa

7 REFERÊNCIAS BIBLIOGRÁFICAS

[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), 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/202208/1.2%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

[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/factsheets/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