UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Rivaldo Correia Santos Junior

PREDIÇÃO DE DOENÇA CARDÍACA E IDENTIFICAÇÃO DE FATORES DE RISCO COM TÉCNICAS ESTATÍSTICAS E DE MACHINE LEARNING

São Cristóvão - SE

2025

Rivaldo Correia Santos Junior

PREDIÇÃO DE DOENÇA CARDÍACA E IDENTIFICAÇÃO DE FATORES DE RISCO COM TÉCNICAS ESTATÍSTICAS E DE MACHINE LEARNING

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Cleber Martins Xavier

São Cristóvão - SE

2025

Rivaldo Correia Santos Junior

PREDIÇÃO DE DOENÇA CARDÍACA E IDENTIFICAÇÃO DE FATORES DE RISCO COM TÉCNICAS ESTATÍSTICAS E DE MACHINE LEARNING

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Aprovado em 11/09/2025.

Prof. Dr. Cleber Martins Xavier
Orientador

Prof. Dr. Allan Robert da Silva
Membro da Banca

Prof. Dr. Ulisses Vieira Guimarães
Membro da Banca

São Cristóvão - SE
2025

Dedico este trabalho à minha família, que foi a base e a força que me sustentaram em cada passo dessa jornada acadêmica. Desde o momento da minha entrada na UFS, vocês estiveram ao meu lado, com amor, paciência e apoio incondicional, acreditando em mim quando, por vezes, eu mesmo duvidava de minha capacidade. Cada esforço de vocês, cada gesto de carinho, foi um alicerce essencial que me impulsionou a seguir em frente, mesmo nos momentos mais desafiadores. A todos que, com sacrifícios e dedicação, me proporcionaram as oportunidades que hoje me permitem alcançar mais um sonho. Não são apenas as palavras que expresso aqui que podem traduzir o quanto sou grato a vocês, mas toda a minha vida e todas as minhas conquistas. Este trabalho é, antes de tudo, um reflexo do amor e da força de cada um de vocês. Este trabalho é dedicado a vocês, que sempre foram meu porto seguro, minha inspiração e minha razão para persistir.

Agradecimentos

A realização deste trabalho não teria sido possível sem o apoio, carinho e contribuição de muitas pessoas, que de alguma forma, fizeram parte dessa caminhada.

Agradeço, primeiramente, à minha família, que sempre me ofereceu amor, apoio incondicional e a motivação necessária para superar os desafios. A presença de vocês foi fundamental para minha jornada, e é fato: sem o apoio de cada um, eu não teria chegado até aqui.

Aos meus amigos, que, com seu carinho, compreensão e dedicação, me deram forças e apoio em todos os momentos, celebrando comigo as pequenas vitórias ao longo do caminho e sendo um alicerce constante em todos os momentos.

Aos professores, colegas e amigos da universidade, que contribuíram para a minha formação acadêmica com ensinamentos, trocas de experiências, aprendizado coletivo e muitos momentos de diversão. Vocês foram peças-chave nesta trajetória.

A cada um de vocês, meu sincero agradecimento. Este trabalho é resultado de muitas mãos e corações que me ajudaram a torná-lo realidade.

“Tudo o que somos é resultado do que pensamos.”

(Buda)

Resumo

As doenças cardiovasculares (DCVs) configuram uma das principais causas de mortalidade mundial, representando um desafio silencioso para a saúde pública, inclusive no Brasil. Diante desse cenário, este trabalho tem como objetivo desenvolver um modelo preditivo capaz de prever a presença de doença cardíaca e identificar os principais fatores de risco associados, contribuindo para estratégias de prevenção e gestão de riscos. Para isso, a metodologia consistiu em explorar dados e ajustar modelos de classificação, utilizando técnicas estatísticas tradicionais e de aprendizado de máquina (Regressão Logística, Random Forest e K-Nearest Neighbors - KNN). Esses métodos foram avaliados por métricas como F1-score e AUC-ROC. Para a identificação dos fatores de risco, utilizou-se o V de Cramer para medir a associação entre as variáveis preditoras e a variável alvo. Os resultados evidenciaram a capacidade preditiva da regressão logística com seleção de variáveis, que apresentou o melhor equilíbrio entre as classes, com F1-score de 0,23 para a classe positiva e AUC-ROC de 0,7859. Os modelos Random Forest e KNN, embora eficientes na classe majoritária, demonstraram desempenho inferior na detecção de casos positivos. A análise dos fatores de risco, por sua vez, reforçou a associação da doença com idade, hipertensão, diabetes, colesterol alto, AVC e problema de sono. Em conclusão, o estudo demonstra a viabilidade de aplicar métodos de aprendizado de máquina na predição de doença cardíaca, apontando a importância da escolha de variáveis para aprimorar a capacidade preditiva dos modelos em cenários de desequilíbrio entre classes.

Palavras-chave: DOENÇA CARDÍACA. FATORES DE RISCO. APRENDIZADO DE MÁQUINA.

Abstract

Cardiovascular diseases (CVDs) are one of the leading causes of global mortality, representing a silent challenge to public health, including in Brazil. Given this scenario, this study's objective is to develop a predictive model capable of forecasting the presence of heart disease and identifying the main associated risk factors, thereby contributing to prevention and risk management strategies. To achieve this, the methodology involved exploring data and fitting classification models using traditional statistical techniques and machine learning (Logistic Regression, Random Forest, and K-Nearest Neighbors - KNN). These methods were evaluated using metrics such as F1-score and AUC-ROC. For the identification of risk factors, Cramer's V was employed to measure the association between predictor variables and the target variable. The results highlighted the predictive capability of logistic regression with variable selection, which showed the best balance between classes, with an F1-score of 0.23 for the positive class and an AUC-ROC of 0.7859. The Random Forest and KNN models, while efficient in the majority class, demonstrated inferior performance in detecting positive cases. The analysis of risk factors, in turn, reinforced the association of the disease with age, hypertension, diabetes, high cholesterol, stroke, and sleep problems. In conclusion, the study demonstrates the feasibility of applying machine learning methods in heart disease prediction, highlighting the importance of variable selection to improve the predictive capacity of models in scenarios of class imbalance.

Keywords: HEART DISEASE. RISK FACTORS. MACHINE LEARNING.

Resumen

Las enfermedades cardiovasculares (ECV) se configuran como una de las principales causas de mortalidad a nivel mundial, representando un desafío silencioso para la salud pública, incluso en Brasil. Ante este escenario, este trabajo tiene como objetivo desarrollar un modelo predictivo capaz de predecir la presencia de enfermedad cardíaca e identificar los principales factores de riesgo asociados, contribuyendo a estrategias de prevención y gestión de riesgos. Para ello, la metodología consistió en explorar datos y ajustar modelos de clasificación, utilizando técnicas estadísticas tradicionales y de aprendizaje automático (Regresión Logística, Random Forest y K-Nearest Neighbors - KNN). Estos métodos fueron evaluados mediante métricas como F1-score y AUC-ROC. Para la identificación de los factores de riesgo, se utilizó la V de Cramer para medir la asociación entre las variables predictoras y la variable objetivo. Los resultados evidenciaron la capacidad predictiva de la regresión logística con selección de variables, que presentó el mejor equilibrio entre las clases, con un F1-score de 0,23 para la clase positiva y un AUC-ROC de 0,7859. Los modelos Random Forest y KNN, aunque eficientes en la clase mayoritaria, mostraron un rendimiento inferior en la detección de casos positivos. El análisis de los factores de riesgo, a su vez, reforzó la asociación de la enfermedad con la edad, hipertensión, diabetes, colesterol alto, accidente cerebrovascular y problemas de sueño. En conclusión, el estudio demuestra la viabilidad de aplicar métodos de aprendizaje automático en la predicción de enfermedades cardíacas, señalando la importancia de la selección de variables para mejorar la capacidad predictiva de los modelos en escenarios de desequilibrio entre clases.

Palabras clave: ENFERMEDAD CARDÍACA. FACTORES DE RIESGO. APRENDIZAJE AUTOMÁTICO.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DCVs	Doenças Cardiovasculares
KNN	K-Nearest Neighbors (Algoritmo dos K Vizinhos Mais Próximos)
ML	Machine Learning (Aprendizado de Máquina)
PNS	Pesquisa Nacional de Saúde
RF	Random Forest (Floresta Aleatória)
RL	Regressão Logística
SMOTE	Synthetic Minority Over-sampling Technique (Técnica de Sobreamostragem de Minorias Sintéticas)
SVM	Support Vector Machine (Máquina de Vetores de Suporte)

Sumário

1 INTRODUÇÃO

Vivemos em um cenário global onde o avanço da ciência médica convive paradoxalmente com uma das maiores crises silenciosas da humanidade: as doenças cardiovasculares (DCVs) (World Heart Federation, 2023). Apesar de não se tornarem manchetes diárias, continuam, discretamente, provocando mortes em uma escala que superam guerras, pandemias e desastres naturais somados. Segundo a World Heart Federation (2023), em 2021, mais de 20,5 milhões de pessoas morreram em decorrência de DCVs, o equivalente a aproximadamente uma em cada três mortes registradas no planeta.

De acordo com Floresti (2024), aproximadamente 400 mil brasileiros morreram em 2022 por doenças cardiovasculares, o que representa cerca de 45 mortes por hora, reforçando que, apesar de serem amplamente conhecidas, as DCVs permanecem subestimadas na percepção coletiva e na priorização das políticas públicas.

As doenças cardíacas compreendem um conjunto de condições que afetam diretamente o funcionamento do coração, diferenciando-se das demais doenças cardiovasculares que incluem patologias vasculares (World Health Organization, 2021). Entre as principais doenças cardíacas estão a doença arterial coronariana (DAC), caracterizada pelo acúmulo de placas ateroscleróticas nas artérias coronárias, que pode resultar em angina e infarto do miocárdio (XAVIER et al., 2013). A insuficiência cardíaca é definida como a incapacidade do coração de bombear sangue adequadamente para atender às necessidades do organismo, ocasionando sintomas como fadiga, dispneia e edema (PONIKOWSKI et al., 2016). As arritmias referem-se a distúrbios no ritmo cardíaco, incluindo a fibrilação atrial, que aumentam o risco de eventos tromboembólicos (JANUARY et al., 2019). Por fim, as cardiomiopatias representam um grupo heterogêneo de doenças do músculo cardíaco que podem ser classificadas em dilatadas, hipertróficas ou restritivas, provocando alterações estruturais e funcionais do miocárdio (ELLIOTT et al., 2014).

A complexidade do desenvolvimento das DCVs transcende causas isoladas. Elas são fruto de uma teia intricada de fatores: genéticos, metabólicos, comportamentais e ambientais. Condições como hipertensão, diabetes, dislipidemia, sedentarismo e tabagismo atuam como vetores silenciosos, muitas vezes imperceptíveis até que se manifestem na forma de infartos, insuficiências cardíacas ou acidentes vasculares fatais (MANSUR; FAVARATO, 2012). A medicina preventiva, por mais eficiente que seja, enfrenta o desafio de antever, com precisão, quem está realmente em risco e quando esse risco pode se materializar.

A compreensão do perfil epidemiológico das doenças cardiovasculares no país é fortemente apoiada por pesquisas nacionais de grande abrangência, como a Pesquisa Nacional de Saúde (PNS). Trata-se de um levantamento epidemiológico realizado periodicamente pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em parceria com o Ministério da Saúde, que coleta

dados sobre as condições de saúde da população brasileira, seus determinantes e o acesso aos serviços de saúde (BRASIL, 2014). A PNS é fundamental para o planejamento e avaliação de políticas públicas na área da saúde, fornecendo informações detalhadas sobre a prevalência de doenças crônicas, fatores de risco comportamentais e uso de serviços de saúde (IBGE, 2019). A partir de seus dados, é possível analisar a prevalência de fatores de risco para doenças cardíacas, como hipertensão arterial, diabetes e obesidade, além de identificar desigualdades regionais e sociais que influenciam os resultados de saúde, permitindo o direcionamento de estratégias de prevenção e controle (MALTA et al., 2022).

É nesse contexto que a estatística moderna e as técnicas de aprendizado de máquina se consolidam como ferramentas estratégicas no enfrentamento desse problema. Modelos estatísticos tradicionais, como a Regressão Logística, oferecem estruturas interpretáveis e robustas, fundamentais para a compreensão clínica. Entretanto, a crescente utilização de métodos de *Machine Learning* (ML), como *Random Forest* e *K-Nearest Neighbors* (KNN), tem se mostrado eficaz na detecção de padrões complexos e na melhoria do desempenho preditivo em cenários de alta dimensionalidade (KRITTANAWONG et al., 2017).

Diante desse panorama, este trabalho tem como objetivo construir um modelo capaz de prever a presença de doença cardíaca e identificar os principais fatores de risco associados, utilizando técnicas estatísticas e de aprendizado de máquina. Dessa forma, busca-se contribuir para o aprimoramento de estratégias de prevenção e gestão de riscos em saúde cardiovascular, por meio de evidências quantitativas robustas.

2 OBJETIVOS

2.1 Geral

Construir um modelo capaz de prever a presença de doença cardíaca e identificar os principais fatores de risco associados, utilizando técnicas estatísticas e de aprendizado de máquina.

2.2 Específicos

- Explorar estatisticamente o comportamento das variáveis envolvidas, com ênfase em características clínicas e comportamentais;
- Ajustar modelos preditivos de classificação, como Regressão Logística, Random Forest e K-Nearest Neighbors (KNN);
- Comparar o desempenho dos modelos utilizando métricas quantitativas como acurácia, precisão, recall, F1-score e AUC-ROC;
- Analisar a robustez e a aplicabilidade dos modelos no apoio à tomada de decisão clínica e em estratégias de prevenção a doenças cardíacas.

3 JUSTIFICATIVA

As doenças cardíacas continuam sendo uma das principais causas de morte no Brasil e no mundo. Apesar dos avanços no diagnóstico e tratamento, a prevenção ainda é a melhor estratégia, e isso depende diretamente da capacidade de prever e identificar fatores de risco precocemente. Com o volume crescente de dados em saúde, técnicas estatísticas e de aprendizado de máquina oferecem novas possibilidades para fortalecer a medicina preventiva, contribuindo para políticas públicas mais eficientes e redução da mortalidade.

Embora existam diversos estudos voltados para a predição de doenças cardiovasculares, muitos estão centrados em modelos tradicionais, limitados em termos de precisão ou capacidade de lidar com variáveis complexas. Este trabalho se justifica por buscar comparar e aplicar diferentes algoritmos (Regressão Logística, K-Nearest Neighbors e Random Forest) sobre um banco de dados real, com foco não apenas na predição, mas na interpretação dos fatores de risco mais influentes, oferecendo suporte à decisão clínica com base em evidências quantitativas robustas.

4 REVISÃO LITERÁRIA

4.1 Doenças Cardíacas no Brasil e no Mundo

As doenças cardíacas englobam um grupo de enfermidades que afetam diretamente a estrutura e a função do coração, sendo uma das principais causas de mortalidade no mundo (World Heart Federation, 2023). Em 2021, estima-se que mais de 20 milhões de pessoas tenham falecido em decorrência de doenças cardiovasculares, majoritariamente por causas cardíacas diretas, como a doença arterial coronariana e a insuficiência cardíaca (World Heart Federation, 2023).

No Brasil, o cenário é igualmente preocupante. Conforme levantamento divulgado pela revista Pesquisa FAPESP, cerca de 400 mil brasileiros morreram em 2022 devido a doenças cardíacas e circulatórias, o que equivale a aproximadamente 45 óbitos por hora (FLORESTI, 2024). Além dos impactos clínicos, essas doenças acarretam consequências econômicas e sociais, afetando a qualidade de vida da população e pressionando o sistema de saúde com hospitalizações frequentes, intervenções complexas e tratamentos prolongados (FLORESTI, 2024).

Além da mortalidade, as doenças cardíacas geram alta morbidade, especialmente em grupos vulneráveis. Bichara et al. (2024), ao analisarem a relação entre o Índice de Vulnerabilidade Social (IVS) e a mortalidade por doenças cardíacas no Brasil, identificaram desigualdades regionais acentuadas, com maior incidência em áreas com piores indicadores sociais. O estudo demonstra que o acesso desigual a informações, diagnósticos e tratamentos adequados contribui para a disparidade nos padrões de mortalidade, para além de fatores genéticos ou comportamentais.

Nesse contexto, a antecipação da ocorrência dessas doenças por meio da identificação precoce de fatores de risco é essencial para a prevenção e gestão em saúde pública (LIU et al., 2025). Essa predição possibilita o direcionamento de políticas eficazes e intervenções precoces, minimizando desfechos adversos e reduzindo os custos ao sistema de saúde (LIU et al., 2025).

Os fatores de risco para doenças cardíacas são amplamente conhecidos e incluem hipertensão arterial, diabetes mellitus, dislipidemia, tabagismo, obesidade e sedentarismo, os quais contribuem para o desenvolvimento de condições como a doença arterial coronariana e a insuficiência cardíaca, impactando significativamente a morbimortalidade cardiovascular no Brasil e globalmente (OLIVEIRA et al., 2022).

De acordo com dados da Pesquisa Nacional de Saúde (PNS), analisados por Malta et al. (2022), a prevalência de hipertensão arterial autorreferida entre adultos brasileiros foi de 23,9% em 2019, com maior incidência entre mulheres, idosos e indivíduos com menor escolaridade. O estudo também relacionou a hipertensão com excesso de peso, obesidade e baixa prática de atividade física, reforçando o papel dos fatores metabólicos e comportamentais no contexto

urbano brasileiro.

A hipertensão arterial, comumente associada ao envelhecimento, e o sedentarismo, combinados com o sobrepeso e a obesidade, são fatores que elevam substancialmente o risco de eventos cardiovasculares graves (CARLUCCI et al., 2013).

A desigualdade no acesso à saúde também exerce influência crucial. Conforme a Organização Pan-Americana da Saúde (Organização Pan-Americana da Saúde, 2021), a hipertensão não controlada é responsável por mais da metade das doenças cardíacas nas Américas, principalmente em países de baixa e média renda, onde os sistemas de acompanhamento clínico são fragilizados.

Adicionalmente, fatores psicossociais, como estresse crônico e depressão, têm sido associados ao agravamento das doenças cardíacas, especialmente em mulheres hipertensas (NASCIMENTO; GOMES; SARDINHA, 2011).

4.2 Técnicas Estatísticas e de Machine Learning Aplicadas à Prevenção de Doenças Cardíacas

4.2.1 Definição de Machine Learning

A área de Machine Learning (ML), ou aprendizado de máquina, é um subcampo da inteligência artificial que se dedica ao desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e melhorar seu desempenho em determinadas tarefas sem necessidade de programação explícita (MITCHELL, 2006).

Segundo Mitchell (2006), "Dizemos que um programa de computador aprende com a experiência E em relação a uma tarefa T e a uma medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E."

Essa definição formaliza o aprendizado de máquina como um processo em que um algoritmo analisa dados (experiência E), ajusta seus parâmetros de acordo com os padrões identificados e melhora sua capacidade de realizar uma determinada tarefa (T), sendo avaliado por uma métrica de desempenho específica (P).

Por exemplo, na detecção de doenças cardíacas, um modelo pode ser treinado utilizando um conjunto de diagnósticos históricos (experiência E), com o objetivo de classificar novos pacientes com doença cardíaca ou não (tarefa T). O desempenho do modelo pode ser avaliado por métricas como acurácia, precisão e recall (medida de desempenho P), indicando sua eficácia nos prognósticos.

4.2.2 Diferença entre Aprendizado Supervisionado e Não Supervisionado

O aprendizado de máquina pode ser classificado de duas formas: supervisionado e não supervisionado (JAMES et al., 2013).

O aprendizado supervisionado ocorre quando o modelo é treinado com um conjunto de dados rotulado, ou seja, cada amostra possui uma entrada associada a uma saída conhecida. Assim, o algoritmo aprende a mapear padrões dos dados de entrada para a saída correta, minimizando um erro predefinido (GOODFELLOW; BENGIO; COURVILLE, 2016). Para ilustrar, considere que um modelo é treinado com milhares de fichas de pacientes, onde cada ficha já contém um diagnóstico final, como "paciente com doença cardíaca"ou "sem doença cardíaca". O modelo aprende a correlacionar os sintomas e exames com esses diagnósticos já confirmados e, a partir daí, consegue prever o diagnóstico para novos pacientes.

Por outro lado, o aprendizado não supervisionado trabalha com dados não rotulados, ou seja, sem uma categorização prévia. Em vez de aprender a associar entradas e saídas, o modelo busca identificar padrões ocultos nos dados, agrupando observações semelhantes e detectando anomalias (MURPHY, 2012). Por exemplo, um modelo pode receber dados de pacientes sem nenhum diagnóstico. Sua tarefa é simplesmente encontrar semelhanças, o que pode levar à descoberta de grupos de pacientes com perfis de risco em comum, como ser mais idoso e ter colesterol alto, sem que ninguém tenha fornecido um rótulo prévio.

Além dessas abordagens, existe também o aprendizado semi-supervisionado, que combina aspectos dos dois métodos anteriores. Esse tipo de técnica é útil quando apenas uma parte do conjunto de dados é rotulada, sendo o restante utilizado para refinar a aprendizagem do modelo (CHAPELLE; SCHOLKOPF; ZIEN, 2006). Um exemplo prático seria um banco de dados com milhares de pacientes, mas com o diagnóstico de apenas uma pequena porcentagem deles. O modelo utiliza o que aprendeu com essa pequena amostra para tentar classificar os demais pacientes, refinando seu próprio conhecimento no processo.

A escolha entre aprendizado supervisionado e não supervisionado depende da disponibilidade de rótulos nos dados e do objetivo da análise. No presente trabalho, optou-se por utilizar aprendizado supervisionado, visto que o conjunto de dados empregado possui histórico dos paciente com diagnóstico com e sem doença cardíaca, permitindo a construção de modelos preditivos capazes de classificar novos diagnósticos com base nos padrões aprendidos.

4.3 Algoritmos Utilizados na Detecção de Doenças Cardíacas

O avanço das técnicas de aprendizado de máquina tem proporcionado ferramentas poderosas para estudos sobre doenças cardíacas. Dentre os diversos algoritmos disponíveis, a Regressão Logística, o Random Forest e o K-Nearest Neighbors (KNN) são amplamente utilizados devido à sua robustez, interpretabilidade e bom desempenho em dados clínicos (SHIMIZU et al., 2024).

A regressão logística é um modelo estatístico que estima a probabilidade de ocorrência de um evento binário, como a presença ou ausência de doença cardíaca, a partir de variáveis preditoras contínuas ou categóricas. Sua aplicabilidade na área da saúde é amplamente reconhecida,

permitindo a identificação de fatores de risco e a estratificação de pacientes (MALTA et al., 2022). Por exemplo, em estudos cardiovasculares, a regressão logística tem sido utilizada para prever eventos como infarto do miocárdio com base em parâmetros clínicos como idade, pressão arterial e histórico familiar (SHIMIZU et al., 2024).

O algoritmo Random Forest consiste em um conjunto (ensemble) de árvores de decisão construídas a partir de subconjuntos aleatórios dos dados e variáveis preditoras, o que contribui para a redução do overfitting[1] e melhora a capacidade preditiva. Este método é especialmente eficaz na análise de dados complexos e de alta dimensionalidade, comuns em estudos biomédicos (NECIOSUP-BOLAñOS; CIEZA-MOSTACERO, 2024). Na área de doenças cardíacas, Random Forest tem sido aplicado com sucesso na predição de risco cardiovascular e na identificação de variáveis preponderantes para o desenvolvimento da doença, evidenciando alta acurácia e sensibilidade (SHIMIZU et al., 2024).

O K-Nearest Neighbors (KNN) é um método baseado em instâncias que classifica um novo exemplo com base na maioria das classes dos seus k vizinhos mais próximos no espaço de características. Apesar de sua simplicidade, o KNN pode ser bastante eficiente em tarefas de classificação clínica, desde que os dados estejam bem pré-processados e normalizados (BRIJITH, 2023). Em estudos sobre doenças cardíacas, o KNN tem sido utilizado para detectar padrões em dados eletrocardiográficos e sinais vitais, auxiliando na classificação de pacientes em grupos de risco (SHIMIZU et al., 2024).

Neste estudo, os três algoritmos serão aplicados e comparados com o intuito de identificar qual deles apresenta o melhor desempenho preditivo para os dados coletados na Pesquisa Nacional de Saúde (PNS), que reúne informações detalhadas sobre o estado de saúde da população brasileira. Essa escolha visa garantir que o modelo selecionado seja robusto e adequado à realidade epidemiológica nacional, permitindo uma análise precisa e relevante para a detecção precoce de doenças cardíacas.

A escolha adequada entre esses algoritmos depende das características do conjunto de dados, da necessidade de interpretabilidade e da complexidade do problema clínico. Em muitos casos, a comparação entre modelos ou a combinação de múltiplos algoritmos (ensemble) pode proporcionar melhores resultados (SHIMIZU et al., 2024).

4.4 Técnicas de Processamento de Dados

O pré-processamento de dados é uma etapa fundamental no desenvolvimento de modelos de aprendizado de máquina, que visa transformar os dados brutos em um formato adequado para a modelagem e otimizar o desempenho do modelo. Para isso, técnicas como normalização, padronização, remoção de outliers e balanceamento das classes são essenciais para reduzir vieses,

[1] Overfitting refere-se a um problema em que o modelo de aprendizado de máquina se ajusta excessivamente aos dados de treinamento, capturando ruídos e detalhes que não são representativos para dados novos e não vistos. Isso faz com que o modelo tenha um desempenho ruim em dados reais.

melhorar o desempenho dos modelos preditivos e garantir maior robustez e acurácia (BRIJITH, 2023).

O desbalanceamento das classes é um desafio frequente em conjuntos de dados clínicos, especialmente na predição de doenças cardíacas, onde a quantidade de informações de uma classe é muito superior à outra. Para mitigar esse problema, técnicas como a Sobreamostragem de Minorias Sintéticas (SMOTE) são amplamente utilizadas para aumentar a representação da classe minoritária (ZHANG; LIU; WANG, 2023). Por outro lado, a Subamostragem Aleatória (Random UnderSampling - RUS), que reduz a quantidade da classe majoritária, ajuda a prevenir o viés dos algoritmos em favor da classe predominante, melhorando a detecção de eventos minoritários em doenças cardiovasculares (HASANAH; SOLEH; SADIK, 2024).

Além disso, técnicas como a seleção de atributos ajudam a eliminar variáveis irrelevantes ou redundantes, reduzindo a complexidade computacional e melhorando a interpretabilidade do modelo (JAMES et al., 2013).

4.5 Técnicas de Normalização e Padronização

A normalização e a padronização de dados são técnicas essenciais no contexto de Machine Learning, pois preparam os dados de maneira adequada antes de serem processados pelos modelos. Essas abordagens são fundamentais para otimizar o desempenho dos modelos e assegurar que variáveis com escalas ou unidades diferentes sejam tratadas de forma uniforme e justa (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.5.1 Normalização

A normalização de dados é o processo de ajustar os valores das características para um intervalo específico, frequentemente entre 0 e 1. Essa técnica é particularmente útil quando as características possuem escalas distintas e é necessário que todas elas estejam dentro de um mesmo intervalo. A fórmula utilizada para realizar a normalização é:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{4.1}$$

Onde:

- x_{norm} representa o valor normalizado.
- x é o valor original da característica.
- x_{min} é o valor mínimo da característica.
- x_{max} é o valor máximo da característica.

A realização desse processo assegura que as variáveis com diferentes escalas possam ser tratadas de maneira equilibrada e contribuam de forma adequada para o desempenho dos modelos de aprendizado de máquina (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.5.2 Padronização

A técnica usado nesse estudo será a da padronização. Em aprendizado de máquina, a padronização é uma técnica essencial de pré-processamento que ajusta os dados para que tenham média zero ($\mu = 0$) e desvio padrão unitário ($\sigma = 1$). Essa transformação é fundamental para algoritmos que dependem de medidas de distância, como KNN e SVM, pois garante que todas as variáveis possuam a mesma escala, evitando vieses decorrentes de grandezas numéricas distintas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A fórmula utilizada para a padronização dos dados é:

$$z = \frac{x - \mu}{\sigma} \tag{4.2}$$

Onde:

- x representa o valor original da observação;
- μ corresponde à média das observações;
- σ é o desvio padrão das observações;
- z representa o valor padronizado.

A padronização é frequentemente utilizada quando os dados seguem uma distribuição normal, tornando-se uma alternativa mais apropriada em relação à normalização em cenários onde a variabilidade das características é relevante para a análise (HE; GARCIA, 2009). Sua aplicação contribui para a melhoria da qualidade dos dados, possibilitando análises mais precisas, geração de relatórios confiáveis e embasando tomadas de decisão mais assertivas.

4.6 Técnicas de Balanceamento de Dados

O balanceamento de dados é uma preocupação crítica em tarefas de aprendizado de máquina. Esse desbalanceamento pode levar a modelos enviesados, apresentando um alto desempenho na predição da classe majoritária, mas baixo desempenho na identificação da classe minoritária (HE; GARCIA, 2009).

Quando os dados desbalanceados não são tratados adequadamente, o modelo projetado pode apresentar dificuldade na identificação dos exemplos da classe minoritária, uma vez que ele se ajusta para minimizar o erro geral, favorecendo a classe majoritária (HE; GARCIA, 2009).

Para mitigar esse problema, diversas técnicas de balanceamento de dados são utilizadas, e na literatura, destacam-se o uso do Synthetic Minority Over-sampling Technique (SMOTE) e o Undersampling.

4.6.1 Synthetic Minority Over-sampling Technique (SMOTE)

O SMOTE é uma técnica de oversampling que gera exemplos sintéticos da classe minoritária para equilibrar o conjunto de dados. Diferente da replicação simples de amostras existentes, ele cria novos pontos de dados interpolando entre exemplos próximos da classe minoritária (CHAWLA et al., 2002). O uso dessa técnica é recomendado quando o desequilíbrio entre as classes é significativo, como é o caso do conjunto de dados usado nesse estudo.

4.6.2 Undersampling

A técnica de undersampling consiste na redução da quantidade de amostras da classe majoritária, equilibrando a distribuição das classes (HE; GARCIA, 2009). Diferentes métodos podem ser aplicados para selecionar quais dados serão retirados, como a exclusão aleatória de amostras e técnicas que buscam remover amostras sem comprometer significamente a informação da classe.

De acordo com a literatura, o uso dessa técnica é recomendado quando o desequilíbrio não é extremo e a redução do conjunto de dados é aceitável. A partir dessa informação, apesar de apresentada aqui, a mesma não será utilizada dada as características da amostra de dados usada nesse estudo.

4.7 Métricas de Avaliação

A análise do desempenho de modelos de classificação em Machine Learning é uma fase essencial para verificar a eficácia do modelo na previsão das classes das amostras. Diversas métricas de avaliação estão disponíveis para quantificar como um modelo de classificação está se comportando, permitindo uma análise mais precisa de sua performance.

Algumas das métricas mais utilizadas para avaliação de modelos de classificação, como a Acurácia, a Precisão, o Recall e o F1-Score. Essas métricas são fundamentais para fornecer uma visão abrangente sobre o desempenho do modelo, especialmente em cenários onde as classes podem ser desbalanceadas.

4.8 Métricas de Avaliação

Para avaliar o desempenho dos modelos preditivos, foram utilizadas métricas de classificação que permitem mensurar a capacidade dos algoritmos de identificar corretamente as classes em um conjunto de dados. Em problemas de classificação, as principais métricas para um conjunto de dados desbalanceado são precisão, recall, F1-score e acurácia (JAMES et al., 2013).

4.8.1 Precisão

A precisão é uma métrica que avalia a confiança de um modelo quando ele prevê que um exemplo pertence a uma determinada classe. Ela é calculada como a razão entre o número de exemplos corretamente classificados como positivos e o número total de exemplos classificados como positivos. Em outras palavras, a precisão indica a proporção de acertos entre os casos que o modelo previu como positivos (JAMES et al., 2013).

A fórmula para calcular a precisão é dada por:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{4.3}$$

Onde:

- TP (True Positives) representa os exemplos que foram corretamente classificados como positivos.
- FP (False Positives) representa os exemplos que foram classificados como positivos, mas na realidade pertencem à classe negativa.

A precisão é uma métrica fundamental, especialmente quando é importante minimizar os falsos positivos, como em um cenário de triagem médica, onde a classificação incorreta de um indivíduo saudável como doente pode gerar ansiedade e custos desnecessários (JAMES et al., 2013).

4.8.2 Recall

O recall, também conhecido como sensibilidade, é uma métrica de avaliação que mede a capacidade de um modelo em identificar corretamente os exemplos positivos em um conjunto de dados (JAMES et al., 2013). Ele é calculado pela razão entre o número de verdadeiros positivos (TP) e o número total de exemplos que realmente pertencem à classe positiva, ou seja, a soma dos verdadeiros positivos (TP) e os falsos negativos (FN). A fórmula para o recall é:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.4}$$

Onde:

- TP (True Positives) são os exemplos corretamente classificados como positivos pelo modelo.
- FN (False Negatives) são os exemplos positivos que o modelo classificou incorretamente como negativos.

O recall é uma métrica crucial quando o custo associado a um falso negativo (não identificar um caso positivo) é elevado, como na detecção de uma doença grave.

4.8.3 F1-Score

O F1-Score é a média harmônica entre a precisão e o recall, oferecendo um equilíbrio entre essas duas métricas. Ele é particularmente importante em situações onde as classes estão desbalanceadas, pois oferece uma visão mais equilibrada da performance do modelo, levando em consideração tanto os falsos positivos quanto os falsos negativos (JAMES et al., 2013). A fórmula para calcular o F1-Score é dada por:

$$F_1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{4.5}$$

Onde:

- Precisão (%) é a proporção de acertos entre os exemplos classificados como positivos.
- Recall (%) é a proporção de acertos entre todos os exemplos que realmente são positivos.

4.8.4 Acurácia

A acurácia é uma métrica simples e amplamente utilizada que mede a proporção de previsões corretas feitas pelo modelo (JAMES et al., 2013). Ela é calculada pela razão entre o número total de classificações corretas (verdadeiros positivos e verdadeiros negativos) e o número total de exemplos no conjunto de dados. A fórmula para calcular a acurácia é dada por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.6}$$

Onde:

- TP (True Positives) são os exemplos corretamente classificados como positivos.
- TN (True Negatives) são os exemplos corretamente classificados como negativos.
- FP (False Positives) são os exemplos classificados como positivos, mas que na verdade são negativos.
- FN (False Negatives) são os exemplos classificados como negativos, mas que na verdade são positivos.

Embora a acurácia seja fácil de entender e interpretar, ela pode ser enganosa em conjuntos de dados desbalanceados, onde uma classe é muito mais comum do que a outra. Nesses cenários, um modelo pode atingir uma alta acurácia simplesmente classificando a classe majoritária

corretamente, sem capturar adequadamente os exemplos da classe minoritária. Portanto, é necessário ter cautela ao utilizar a acurácia como única métrica de avaliação (JAMES et al., 2013).

4.8.5 Curva ROC e AUC

A Curva ROC (Receiver Operating Characteristic) é uma ferramenta gráfica amplamente utilizada para avaliar o desempenho de modelos de classificação binária, mostrando a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos em diferentes limiares de decisão (FAWCETT, 2006). A análise da curva ROC permite visualizar a capacidade do modelo em distinguir entre as classes positiva e negativa.

A métrica AUC (Area Under the Curve), ou Área Sob a Curva, representa a área total sob a curva ROC. O valor da AUC varia entre 0 e 1, onde 1 indica um classificador perfeito e 0,5 corresponde a um classificador aleatório, sem poder discriminatório (HANLEY; MCNEIL, 1982). Quanto maior a AUC, melhor é o desempenho do modelo em separar corretamente as classes.

A curva ROC e a AUC são especialmente úteis em cenários com classes desbalanceadas, pois avaliam o desempenho do modelo independentemente do limiar de classificação, oferecendo uma visão mais robusta da qualidade preditiva (FAWCETT, 2006).

4.9 Trabalhos Relacionados

Diante da complexidade e da natureza multifatorial das doenças cardiovasculares, a estatística e o aprendizado de máquina se consolidaram como ferramentas essenciais para a análise de dados epidemiológicos e para a predição de risco. Uma das abordagens mais comuns nesse cenário é a regressão logística, frequentemente utilizada para estimar a probabilidade de eventos cardíacos adversos a partir de variáveis como idade, pressão arterial e histórico clínico (SHIMIZU et al., 2024). Além disso, o avanço tecnológico permitiu a ampla exploração de técnicas de aprendizado de máquina, como Random Forest, K-Nearest Neighbors (KNN) e XGBoost, em estudos comparativos com modelos tradicionais. Por exemplo, Shimizu et al. (SHIMIZU et al., 2024) avaliaram esses algoritmos com base em métricas como a Área Sob a Curva (AUC) e acurácia, observando que o Random Forest apresentou melhor performance interna (AUC = 0,871; acurácia = 0,794) e externa (AUC = 0,786; acurácia = 0,710), superando modelos tradicionais. Esse estudo também evidenciou que modelos de ensemble ofereceram um incremento de 3,6% no AUC em comparação com escalas de risco convencionais, reforçando a eficácia dessas técnicas para a predição em saúde cardiovascular.

Esses achados são consistentes com revisões de maior abrangência, como a conduzida por Neciosup-Bolaños e Cieza-Mostacero (2024), que analisaram 32 estudos e reforçaram que o Random Forest frequentemente atinge acurácia entre 88% e 95%, mostrando-se especialmente

eficaz em cenários de alta dimensionalidade. Em paralelo, modelos híbridos de deep learning, como arquiteturas CNN-LSTM, vêm demonstrando superioridade em termos de sensibilidade, especificidade e AUC. Shishehbore e Awan (2024), por exemplo, destacaram a integração desses modelos com dados provenientes de dispositivos vestíveis e registros eletrônicos, ampliando as possibilidades de monitoramento contínuo e predição em tempo real.

No contexto da saúde pública, o uso de machine learning também tem se consolidado como ferramenta relevante para análise de variáveis socioeconômicas, demográficas e de acesso aos serviços de saúde. Um exemplo é o trabalho de Bergamini et al. (2020), que utilizou oito algoritmos distintos para estimar taxas de mortalidade por Doença Isquêmica do Coração (IHD) nos municípios do Paraná. O modelo Support Vector Machine (SVM) apresentou o melhor desempenho (RMSE $\approx$ 0,79), além de forte correlação com os dados observados, contribuindo para identificar regiões de maior risco e desigualdade no cuidado cardiovascular.

Resultados semelhantes foram reportados por Delpino et al. (2025), que aplicaram algoritmos como Random Forest, XGBoost e regressão logística para prever mortalidade por todas as causas, incluindo as cardíacas. O Random Forest novamente obteve destaque (AUC = 0,92), e a análise de interpretabilidade via SHAP revelou variáveis como idade, IMC, uso de medicamentos e nível de atividade física como principais preditores. Esse estudo reforça a importância da integração de modelos preditivos em sistemas de saúde para apoiar políticas preventivas mais eficazes.

A literatura também tem discutido aspectos metodológicos fundamentais. Brijith (2023) ressaltam a importância de técnicas de pré-processamento, como normalização, padronização e balanceamento de classes (e.g., SMOTE), para garantir robustez e confiabilidade dos modelos. De forma complementar, estudos clássicos (FAWCETT, 2006; HANLEY; MCNEIL, 1982) reforçam a necessidade de métricas adequadas, como a curva ROC e a AUC, para avaliação em cenários de classes desbalanceadas.

Por fim, um desafio recorrente apontado na literatura é a interpretabilidade dos modelos, aspecto essencial para adoção clínica. Shimizu et al. (2024) demonstraram o uso de técnicas como LIME e valores SHAP para explicar decisões individuais dos modelos, facilitando a compreensão pelos profissionais de saúde e ampliando a aplicabilidade prática dos algoritmos.

Com base nesses trabalhos, observa-se que a combinação de estatística clássica e aprendizado de máquina tem se consolidado como abordagem promissora para a predição de doenças cardíacas. Nesse contexto, o presente estudo propõe comparar o desempenho dos algoritmos Regressão Logística, Random Forest e KNN utilizando dados da PNS, contribuindo para identificar o modelo mais eficaz no cenário brasileiro.

5 METODOLOGIA

5.1 Tipo de Pesquisa

Este estudo caracteriza-se como uma pesquisa aplicada, pois visa desenvolver um modelo baseado em Machine Learning e métodos estatísticos para auxiliar na detecção de doenças cardíacas, integrando variáveis clínicas e psicossociais, de modo a fornecer uma solução prática para um problema de grande relevância em saúde pública. Além disso, trata-se de uma pesquisa explicativa, pois busca compreender a relação entre essas variáveis e a ocorrência das doenças cardíacas, analisando como o balanceamento de dados, a normalização e a padronização afetam o desempenho dos modelos (GIL, 2019).

5.2 Conjunto de Dados

Os dados utilizados neste estudo provêm da Pesquisa Nacional de Saúde (PNS)[1] de 2019, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em parceria com o Ministério da Saúde. A PNS é um levantamento domiciliar que coleta informações detalhadas sobre condições de saúde, hábitos de vida, doenças crônicas, fatores de risco e características sociodemográficas da população brasileira. Para este trabalho, foram selecionadas variáveis relevantes para a análise da presença de doenças cardíacas e fatores associados, incluindo características sociodemográficas, condições clínicas, hábitos de vida e padrões alimentares.

5.2.1 Variáveis Selecionadas

As variáveis selecionadas para a análise inicial foram agrupadas em quatro grandes categorias: *sociodemográficas*, *condições clínicas*, *hábitos de vida* e *alimentação*.

5.2.1.0.1 Variáveis sociodemográficas

- **C006 - Sexo:** sexo do participante (masculino/feminino);
- **C008 - Idade:** idade do morador na data de referência;
- **C009 - Cor ou raça:** categorizada em branca, preta, parda, amarela, indígena ou ignorado.

5.2.1.0.2 Variáveis sobre dor torácica

- **N004, N005, N008:** presença de dor ou desconforto no peito ao realizar esforços físicos e localização da dor.

[1] <https://www.pns.icict.fiocruz.br/>

5.2.1.0.3 Variáveis de sono e antropometria

- **N010 - Problemas de sono:** frequência de dificuldades para dormir ou manter o sono, transformada em variável dicotômica para análise;
- **P00104 - Peso (kg)** e **P00404 - Altura (cm):** utilizados para cálculo do índice de massa corporal (IMC).

5.2.1.0.4 Hábitos de vida

- **P027 - Consumo de álcool:** frequência de ingestão de bebidas alcoólicas, transformada em variável dicotômica;
- **P034 - Atividade física:** prática de exercícios físicos ou esportes nos últimos três meses;
- **P050 e P052 - Tabagismo atual e passado:** uso presente ou histórico de produtos do tabaco.

5.2.1.0.5 Condições clínicas

Q00201 - Hipertensão arterial, **Q03001 - Diabetes**, **Q060 - Colesterol alto**, **Q06306 - Doença cardíaca**, **Q068 - AVC**, **Q074 - Asma**, **Q092 - Depressão**, **Q120 - Câncer:** presença de diagnóstico médico de cada condição, codificada como variável dicotômica (sim/não).

5.2.1.0.6 Variáveis de alimentação

Foram consideradas perguntas sobre a frequência de consumo de diferentes grupos alimentares (**P00601 a P00623**). A partir dessas informações, foram construídos escores de consumo alimentar, categorizando os participantes em níveis *baixo*, *moderado* ou *alto* para consumo de alimentos saudáveis e não saudáveis. Essa agregação permitiu analisar a relação entre padrões alimentares e presença de doenças cardíacas de forma mais sintética e informativa.

5.3 Pré-Processamento de Dados

5.3.1 Tratamento de Valores Ausentes

Para garantir a consistência das análises, os valores ausentes, originalmente representados como "NA", foram padronizados como valores nulos e posteriormente removidos do conjunto de dados. Dessa forma, a base de dados resultante passou de 293.726 registros iniciais para 71.060 observações válidas, todas contendo informações completas nas 40 variáveis selecionadas inicialmente. Esse procedimento assegurou que as análises subsequentes fossem realizadas sobre um conjunto consistente de dados, sem interferência de casos com informações ausentes.

5.3.2 Seleção e Transformação de Variáveis

Após o tratamento de valores ausentes e a definição da amostra final, procedeu-se à seleção e transformação das variáveis de interesse para a análise. Inicialmente, foi calculado o Índice de Massa Corporal (IMC) a partir das variáveis de peso e altura dos participantes, utilizando a fórmula padrão:

$$IMC = \frac{\text{peso (kg)}}{\text{altura}^2 (m^2)} \tag{5.1}$$

Em seguida, foram criadas variáveis compostas para avaliar os padrões de alimentação dos participantes. Para tanto, os alimentos relatados na pesquisa foram categorizados em dois grupos: alimentos saudáveis (como arroz, feijão, carnes, ovos, verduras, frutas e leite) e alimentos não saudáveis (como refrigerantes, sucos industrializados, doces, salgadinhos, embutidos e alimentos ultraprocessados). Cada item foi convertido em uma variável binária, onde 1 indicava consumo e 0 ausência de consumo. A partir dessas variáveis, foram calculados escores agregados para a alimentação saudável e não saudável, somando-se os valores de consumo de cada grupo. Posteriormente, para cada escore, os participantes foram classificados em quartis[2] com base na distribuição dos dados, o que permitiu categorizar os perfis de consumo em "Baixo", "Moderado", "Alto" e "Muito Alto", conforme detalhado na Tabela 1.

Tabela 1 – Classificação dos Escores de Alimentação por Quartis

Categoria	**Escore de Alimentação Saudável**	**Escore de Alimentação Não Saudável**
Baixo	Até 5	Até 1
Moderado	De 6 a 7	Igual a 2
Alto	Igual a 8	Igual a 3
Muito Alto	Acima de 8	Acima de 3

Fonte: Elaborado pelos autores.

Posteriormente, realizou-se a limpeza do conjunto de dados, removendo as colunas originais de alimentos individuais, mantendo apenas os escores agregados e as demais variáveis relevantes para análise. As variáveis foram então renomeadas de forma mais descritiva, incluindo: Sexo, Idade, Raça/Cor, Problemas de Sono, Consumo de Álcool, Atividade Física, Tabagismo Atual e Passado, Hipertensão, Diabetes, Colesterol Alto, AVC, Asma, Depressão, Câncer, Alimentação Saudável, Alimentação Não Saudável, IMC e Doença Cardíaca.

Para garantir consistência e facilitar a análise estatística, as variáveis categóricas foram recodificadas em formato numérico. Por exemplo, Sexo foi transformado em 0 para feminino e 1 para masculino; Problemas de Sono, Consumo de Álcool, Atividade Física, Tabagismo,

[2] Quartil é uma medida estatística que divide um conjunto de dados ordenados em quatro partes iguais. O primeiro quartil (Q1) representa 25% dos dados, o segundo quartil (Q2) representa 50% dos dados (equivalente à mediana), e o terceiro quartil (Q3) representa 75% dos dados.

condições clínicas e diagnóstico de Doença Cardíaca também foram binarizados, atribuindo 1 à presença do evento ou condição e 0 à sua ausência. O IMC foi categorizado segundo classificação internacional padrão (baixo peso, peso normal, sobrepeso e graus de obesidade).

Dessa forma, o conjunto de dados final ficou estruturado e padronizado, permitindo a realização de análises descritivas e inferenciais de forma coerente e consistente.

5.3.2.1 Análise de Associação entre Variáveis

Para a identificação dos principais fatores de risco associados à Doença Cardíaca, utilizou-se o Coeficiente de V de Cramer. Essa métrica, baseada no teste de qui-quadrado, mede a força da associação entre duas variáveis nominais, como as variáveis preditoras e a variável alvo. O valor do V de Cramer varia de 0 (ausência total de associação) a 1 (associação perfeita) e, por ser normalizado, permite a comparação entre diferentes variáveis, independentemente do número de categorias (HAIR et al., 2009). Sua fórmula é expressa por:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \tag{5.2}$$

Essa análise foi crucial para selecionar as variáveis mais relevantes, orientando as etapas subsequentes de modelagem preditiva.

5.3.3 Divisão, Padronização e Codificação dos Dados

Após a limpeza, seleção e transformação das variáveis, os dados foram divididos em conjuntos de treinamento e teste. Essa divisão tem como objetivo permitir a construção de modelos de predição e a validação de seu desempenho em dados não utilizados durante o treinamento. Optou-se por uma proporção de 70% para treinamento e 30% para teste, garantindo que ambos os conjuntos fossem representativos da amostra total.

Para a variável idade, que é contínua, aplicou-se padronização (z-score), transformando os dados de modo que apresentassem média zero e desvio padrão unitário. Essa etapa é importante para algoritmos sensíveis à escala das variáveis, como o KNN, garantindo que a idade não domine o cálculo de distâncias.

A codificação OneHotEncoder foi aplicada à variável raça/cor, de forma a transformá-la em variáveis dummy, permitindo que modelos que exigem entrada numérica interpretem corretamente as categorias. Essa transformação foi utilizada apenas nos modelos KNN e regressão logística, que não conseguem lidar diretamente com variáveis categóricas representadas por números inteiros.

5.4 Seleção do Melhor Modelo

Para a escolha do melhor modelo preditivo de Doença Cardíaca, foram testados três algoritmos: Regressão Logística, K-Nearest Neighbors (KNN) e Random Forest. A avaliação do desempenho de cada modelo considerou métricas amplamente utilizadas em problemas de classificação binária: precision, recall, f1-score e AUC-ROC.

Cada modelo foi treinado e testado utilizando os dados previamente processados e o desempenho dos modelos foi comparado de acordo com as métricas mencionadas, permitindo a identificação do algoritmo com melhor capacidade preditiva para o conjunto de dados estudado.

5.5 Ferramentos Utilizadas

O desenvolvimento e a experimentação dos modelos foram realizados utilizando um conjunto de ferramentas amplamente reconhecidas para análise de dados e aprendizado de máquina.

O projeto foi implementado na linguagem **Python**, escolhida devido à sua versatilidade e ampla adoção na área de ciência de dados. As principais bibliotecas utilizadas incluem:

- **Pandas** - manipulação e análise de dados tabulares;
- **NumPy** - operações matemáticas e manipulação de arrays;
- **Scikit-learn** - implementação de algoritmos de Machine Learning, padronização de dados, codificação OneHotEncoder e métricas de avaliação;
- **Imbalanced-learn** - aplicação de técnicas de balanceamento de dados, como SMOTE;
- **Matplotlib e Seaborn** - visualização gráfica de dados e resultados.

Para a execução do código e a realização das análises foi utilizado o **Google Colab**, que permite execução em nuvem, facilitando o processamento de grandes volumes de dados. O uso dessa ferramenta possibilitou a realização de testes eficientes e uma análise detalhada dos modelos desenvolvidos.

6 RESULTADOS

6.1 Análise Exploratória de Dados

Com a base devidamente preparada, deu-se início à etapa de Análise Exploratória com o objetivo de compreender melhor as características da amostra, identificar possíveis distorções e já antever os desafios que impactarão a modelagem.

De início, foi possível identificar um desbalanceamento nos dados em relação à variável-alvo (**Doença Cardíaca**). Como mostra a Tabela 2, 94,17% dos indivíduos da amostra declararam **não ter** diagnóstico de doença cardíaca, enquanto 5,83% declararam **ter** o diagnóstico. Esse tipo de desproporção é importante pois pode impactar a performance de modelos preditivos, exigindo o uso de estratégias de balanceamento ou métricas específicas de avaliação durante a modelagem.

Tabela 2 – Distribuição dos casos de doença cardíaca nos dados da Pesquisa Nacional de Saúde (PNS) de 2019 para a amostra analisada.

Diagnóstico de Doença Cardíaca	Frequência Absoluta	Frequência Relativa (%)
Não (0)	66.920	94,17
Sim (1)	4.140	5,83
Total	71.060	100,00

Fonte: Elaborado pelos autores.

Na sequência da análise, observou-se a composição da amostra em relação ao **sexo** dos participantes. Como apresenta a Tabela 3, 56,18% dos entrevistados se identificaram como mulheres e 43,82% como homens, totalizando os 71.060 registros considerados após o processamento dos dados.

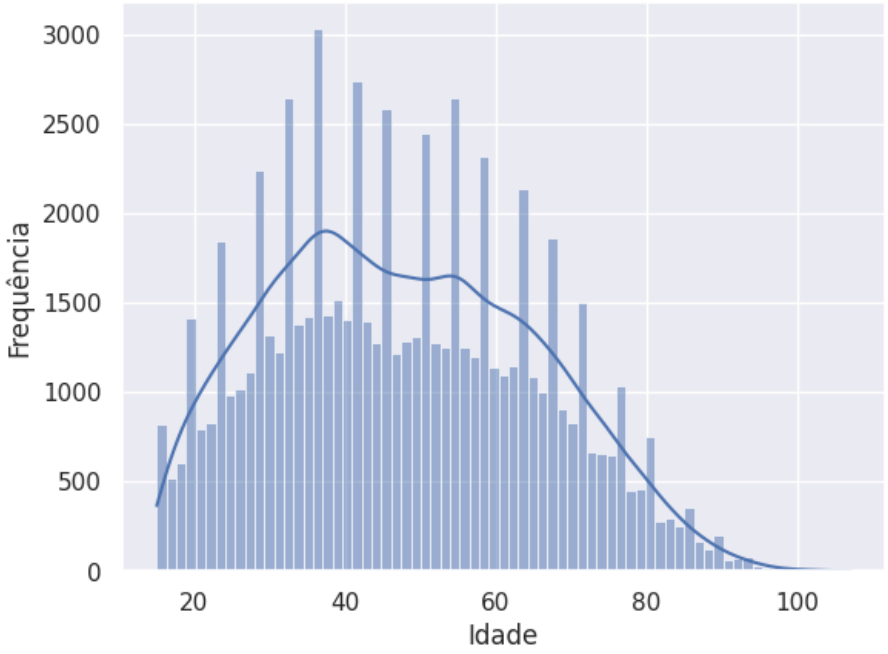
Tabela 3 – Distribuição dos participantes por sexo na amostra analisada.

Sexo	Frequência Absoluta	Frequência Relativa (%)
Mulher (2)	39.924	56,18
Homem (1)	31.136	43,82
Total	71.060	100,00

Fonte: Elaborado pelos autores.

A Figura 1 apresenta a distribuição da variável **Idade**, com uma curva de densidade kernel (KDE) que ajuda a visualizar a forma da distribuição.

Figura 1 – Distribuição etária dos participantes da PNS 2019



Fonte: Elaborado pelo autor.

Observa-se que a idade dos participantes varia amplamente, com uma concentração de indivíduos em faixas etárias intermediárias entre 40 e 60 anos. A curva KDE indica uma distribuição possivelmente assimétrica, com um pico central e uma cauda que se estende para idades mais avançadas, o que reflete a presença de idosos na amostra. Essa variabilidade é relevante, pois a idade é um fator de risco conhecido para doenças cardíacas, e a distribuição observada pode influenciar os resultados da modelagem preditiva.

Dando sequência à caracterização da amostra, a Tabela 4 apresenta a distribuição geral dos participantes em relação às variáveis analisadas, incluindo informações sobre etnia, hábitos de saúde, condições médicas e padrões alimentares.

A variável Raça/Cor evidencia que a maior proporção dos participantes se declarou parda (49,37%), seguida por brancos (38,18%). Os grupos de pretos, amarelos, indígenas e casos ignorados representam proporções menores (10,97%, 0,77%, 0,70% e 0,01%, respectivamente). Essa distribuição reforça a diversidade étnico-racial da amostra, um fator relevante para análises de saúde que podem envolver determinantes sociais e biológicos associados à raça/cor.

Tabela 4 – Panorama detalhado dos participantes da amostra segundo variáveis de saúde e comportamentais.

Variável	Categoria	Frequência Absoluta	Frequência Relativa (%)
Raça/Cor	Parda (4)	35.083	49,37
	Branca (1)	27.132	38,18
	Preta (2)	7.798	10,97
	Amarela (3)	546	0,77
	Indígena (5)	494	0,70
	Ignorado (9)	7	0,01
	Total	71.060	100,00
Problemas de Sono	Não (0)	45.123	63,50
	Sim (1)	25.937	36,50
	Total	71.060	100,00
Consumo de Álcool	Não (0)	53.960	75,94
	Sim (1)	17.100	24,06
	Total	71.060	100,00
Atividade Física	Não (0)	41.040	57,75
	Sim (1)	30.020	42,25
	Total	71.060	100,00
Tabagismo Atual	Não (0)	71.060	100,00
	Total	71.060	100,00
Tabagismo Passado	Não (0)	49.070	69,05
	Sim (1)	21.990	30,95
	Total	71.060	100,00
Hipertensão	Não (0)	50.466	71,02
	Sim (1)	20.594	28,98
	Total	71.060	100,00
Diabetes	Não (0)	64.479	90,74
	Sim (1)	6.581	9,26
	Total	71.060	100,00
Colesterol Alto	Não (0)	59.097	83,16
	Sim (1)	11.963	16,84
	Total	71.060	100,00
AVC	Não (0)	69.402	97,67
	Sim (1)	1.658	2,33
	Total	71.060	100,00
Asma	Não (0)	67.535	95,04
	Sim (1)	3.525	4,96
	Total	71.060	100,00
Depressão	Não (0)	64.275	90,45
	Sim (1)	6.785	9,55
	Total	71.060	100,00
Câncer	Não (0)	69.004	97,11
	Sim (1)	2.056	2,89
	Total	71.060	100,00
Alimentação Saudável	Moderado	23.807	33,50
	Baixo	23.110	32,52
	Muito Alto	14.146	19,91
	Alto	9.997	14,07
	Total	71.060	100,00
Alimentação Não Saudável	Baixo	28.499	40,11
	Muito Alto	15.962	22,46
	Moderado	14.997	21,10
	Alto	11.602	16,33
	Total	71.060	100,00

Fonte: Elaborado pelos autores.

Em relação à variável **Problemas de Sono**, observa-se que 36,50% dos participantes relataram dificuldades relacionadas ao sono, enquanto 63,50% afirmaram não apresentar tais problemas. Este dado indica que mais de um terço da amostra convive com distúrbios do sono, condição frequentemente associada a maior risco de doenças cardiovasculares, metabólicas e transtornos mentais.

A análise do **Consumo de Álcool** mostra que a maior parte da amostra (75,94%) declarou não consumir bebidas alcoólicas, enquanto 24,06% informaram possuir esse hábito. Já a variável **Atividade Física** revelou que 57,75% dos participantes não praticam atividade física regularmente, ao passo que 42,25% afirmaram manter algum tipo de exercício nos últimos três meses. Tais dados indicam que hábitos relacionados ao estilo de vida, potencialmente impactantes na saúde cardiovascular, estão presentes de forma significativa na população analisada.

No que tange ao **Tabagismo Atual**, observa-se que nenhum participante relatou uso de tabaco, tornando essa variável ineficaz para análise comparativa. Entretanto, a variável **Tabagismo no Passado** evidencia que 30,95% dos indivíduos já tiveram contato com produtos derivados do tabaco, enquanto 69,05% nunca tiveram esse hábito.

As condições de saúde pré-existentes foram avaliadas a partir das variáveis **Hipertensão**, **Diabetes**, **Colesterol Alto**, **AVC**, **Asma**, **Depressão** e **Câncer**. Os resultados indicam que 28,98% dos participantes possuem diagnóstico de hipertensão, 9,26% de diabetes, 16,84% de colesterol alto, 2,33% de histórico de AVC, 4,96% de asma, 9,55% de depressão e 2,89% de câncer. Esses dados refletem o perfil de saúde da amostra e apontam para a relevância de tais condições na investigação da doença cardíaca.

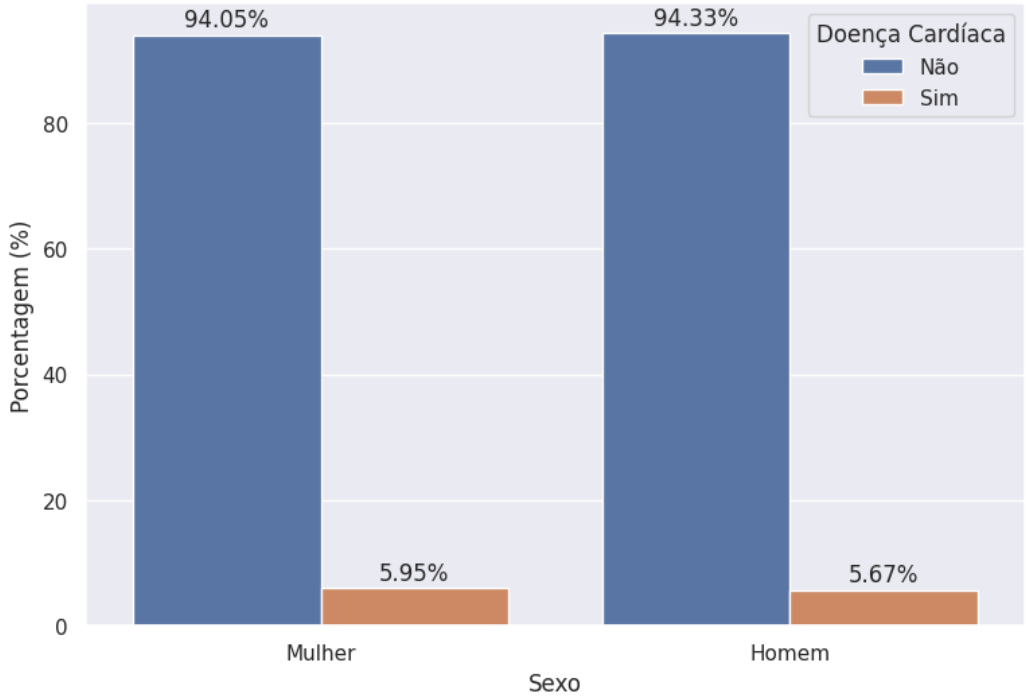
Por fim, a variável **Alimentação Saudável** mostra que 33,50% dos participantes apresentam hábitos alimentares moderados, 32,52% baixos, 19,91% muito altos e 14,07% altos. A distribuição sugere heterogeneidade nos padrões alimentares, aspecto que pode influenciar fatores de risco associados à saúde cardiovascular. Além da alimentação saudável, a variável **Alimentação Não Saudável** foi analisada, evidenciando que 40,11% dos participantes apresentam hábitos alimentares classificados como baixos, 22,46% muito altos, 21,10% moderados e 16,33% altos, conforme apresentado na Tabela 4.

A partir dessa visão geral, torna-se possível avançar para uma etapa mais aprofundada da investigação, em que cada variável é analisada em relação à variável-alvo (*Doença Cardíaca*). Esse procedimento busca identificar padrões e possíveis associações relevantes, permitindo compreender de que forma diferentes fatores se relacionam com a presença ou ausência da condição.

A análise inicial buscou examinar a distribuição dos casos de doença cardíaca de acordo com a variável *Sexo*, tendo como referência a variável-alvo (*Doença Cardíaca*). Essa etapa permite verificar possíveis diferenças na ocorrência da condição entre homens e mulheres, fornecendo uma primeira aproximação sobre o papel das características sociodemográficas no

perfil dos participantes. A Figura 2 ilustra esse comportamento.

Figura 2 – Prevalência de Doença Cardíaca em homens e mulheres na amostra.



Fonte: Elaborado pelo autor.

Observa-se que, tanto entre homens quanto entre mulheres, a maioria dos indivíduos declarou não possuir diagnóstico médico para doença cardíaca.

Foi realizada a tabulação cruzada entre a variável Doença Cardíaca e a variável Raça/Cor. A Tabela 5 apresenta a distribuição dos indivíduos com e sem diagnóstico de doença cardíaca, de acordo com a classificação de raça/cor.

Tabela 5 – Casos de Doença Cardíaca por Raça/Cor na amostra estudada.

Raça/Cor	**Não tem Doença Cardíaca**	**Tem Doença Cardíaca**	**Total**
Branca (1)	25.290 (93,2%)	1.842 (6,8%)	27.132
Preta (2)	7.372 (94,5%)	426 (5,5%)	7.798
Amarela (3)	516 (94,5%)	30 (5,5%)	546
Parda (4)	33.274 (94,9%)	1.809 (5,1%)	35.083
Indígenas (5)	461 (93,3%)	33 (6,7%)	494
Ignorado (9)	7 (100,0%)	0 (0,0%)	7
Total	66.920	4.140	71.060

Fonte: Elaborado pelos autores.

Observa-se que a maior parte da amostra é composta por indivíduos autodeclarados pardos (33.274 sem e 1.809 com doença cardíaca) e brancos (25.290 sem e 1.842 com doença

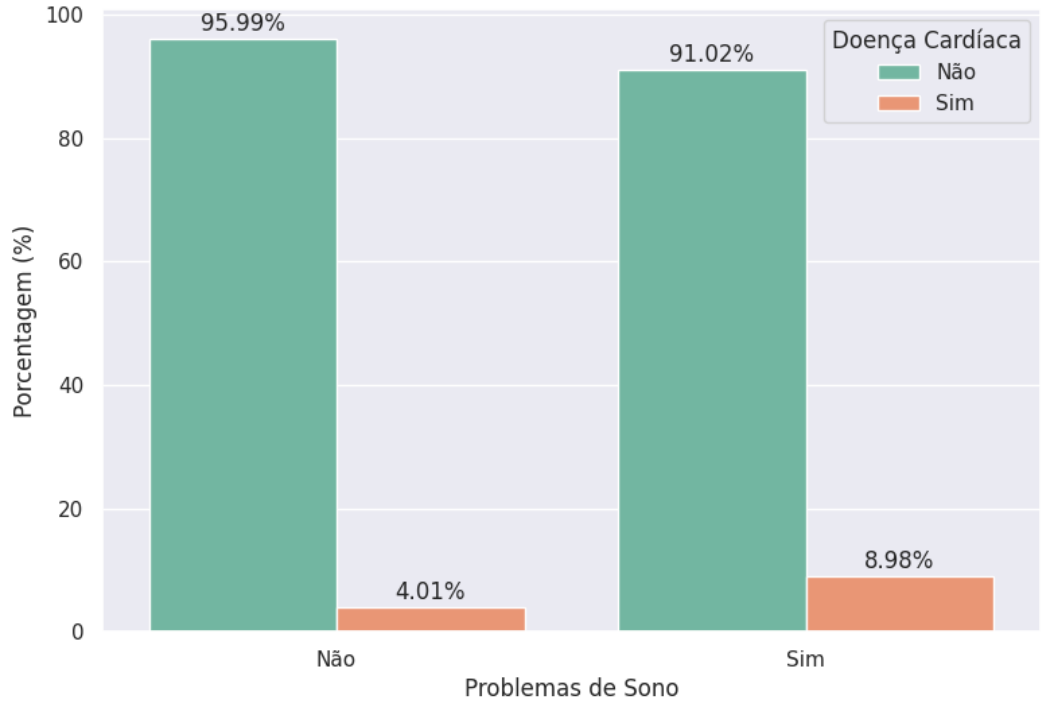
cardíaca). Entre os indivíduos pardos e brancos, embora os números absolutos de pessoas com diagnóstico sejam expressivos, em termos relativos, o percentual de ocorrência de doença cardíaca é semelhante.

Já entre os indivíduos autodeclarados pretos (426 casos em 7.372 indivíduos) percebe-se uma proporção discretamente mais elevada de casos em relação ao observado entre brancos e pardos, o que pode sugerir diferenças estruturais relacionadas a determinantes sociais e de saúde.

Grupos de menor representatividade, como amarelos (30 casos) e indígenas (33 casos), apresentaram números absolutos bastante reduzidos, o que limita inferências estatísticas robustas. Por fim, a categoria "Ignorado" não apresentou casos de doença cardíaca, mas representa uma parcela irrelevante da amostra.

Dando continuidade à investigação dos fatores possivelmente associados à presença de doença cardíaca, analisou-se a relação entre a variável **Problemas de Sono** e o diagnóstico de **Doença Cardíaca**. A Figura 3 apresenta a distribuição dos participantes com e sem a condição, estratificada conforme a presença ou ausência de distúrbios do sono.

Figura 3 – Prevalência de Doença Cardíaca em participantes com/sem problemas de sono.



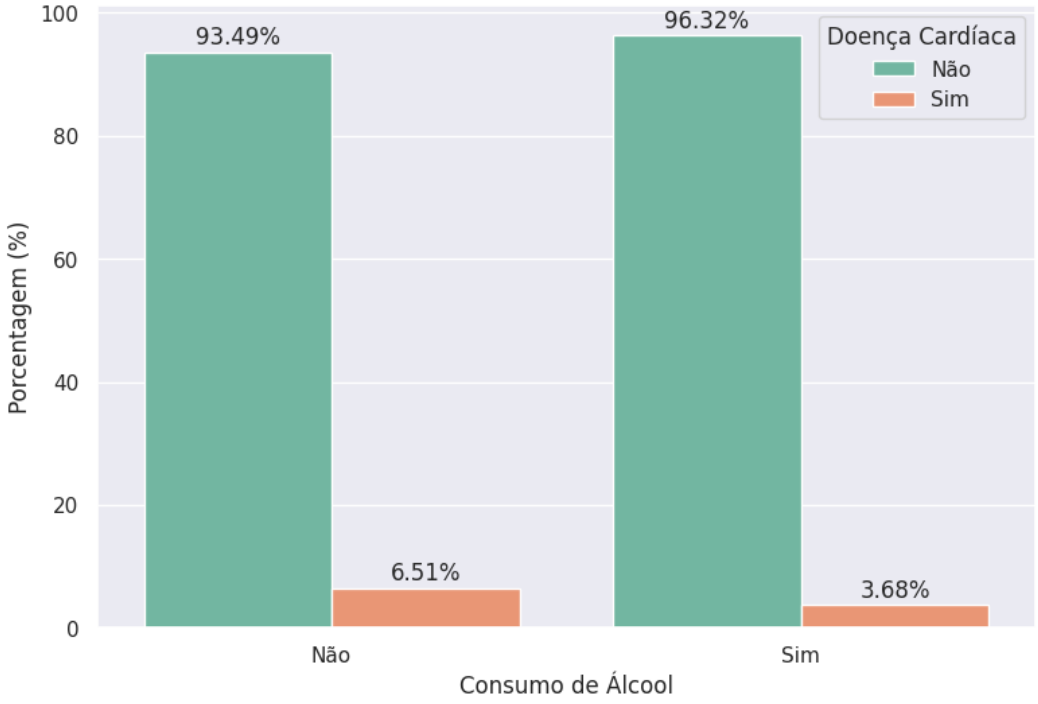
Fonte: Elaborado pelo autor.

Nota-se que a maioria dos indivíduos que relatam não apresentar problemas de sono também não possuem diagnóstico médico para doença cardíaca. Embora o grupo que relata problemas de sono seja numericamente menor na amostra total, ele concentra uma proporção significativamente maior de diagnósticos positivos para a condição cardíaca, com 8,98% dos

indivíduos desse grupo apresentando a doença, em contraste com 4,01% do grupo que não relata problemas de sono. Essa diferença sugere uma possível associação entre a qualidade do sono e o risco cardiovascular, o que está em consonância com evidências já descritas na literatura médica. Assim, a variável Problemas de Sono será considerada nas etapas subsequentes de modelagem, dada sua potencial relevância no contexto analisado.

A Figura 4 mostra a relação entre o consumo de álcool e o diagnóstico de **Doença Cardíaca**.

Figura 4 – Distribuição da doença cardíaca na amostra segundo consumo de álcool entre os participantes



Fonte: Elaborado pelo autor.

Embora a maior parte dos participantes que consomem álcool não tenha recebido esse diagnóstico, observa-se que a proporção de casos positivos é menor entre os que consomem do que entre os que não consomem. Dos 17.100 participantes que relataram consumir bebidas alcoólicas, 629 (3,68%) têm diagnóstico de doença cardíaca. Já entre os 53.960 que afirmaram não consumir, 3.511 (6,51%) relataram a condição. Esses dados chamam atenção para uma inversão em relação ao que se poderia esperar intuitivamente, sendo necessário considerar que o não consumo pode estar relacionado à presença prévia da doença ou a mudanças no estilo de vida após o diagnóstico.

A Tabela 6 consolida os dados referentes a todas as variáveis clínicas e comportamentais analisadas, fornecendo uma visão integrada e detalhada do perfil da amostra.

Tabela 6 – Casos de doença cardíaca estratificados por variáveis clínicas e comportamentais dos participantes.

Variável	**Não tem Doença Cardíaca**	**Tem Doença Cardíaca**	**Total**
Atividade Física (Não)	38.245 (93,18%)	2.795 (6,82%)	41.040
Atividade Física (Sim)	28.675 (95,52%)	1.345 (4,48%)	30.020
Tabagismo (Sim)	20.073 (91,28%)	1.917 (8,72%)	21.990
Tabagismo (Não)	46.847 (95,47%)	2.223 (4,53%)	49.070
Hipertensão (Sim)	17.904 (86,94%)	2.690 (13,06%)	20.594
Hipertensão (Não)	49.016 (97,12%)	1.450 (2,88%)	50.466
Diabetes (Sim)	5.613 (85,29%)	968 (14,71%)	6.581
Diabetes (Não)	61.307 (95,08%)	3.172 (4,92%)	64.479
Colesterol Alto (Sim)	10.371 (86,69%)	1.592 (13,31%)	11.963
Colesterol Alto (Não)	56.549 (95,70%)	2.548 (4,30%)	59.097
AVC (Sim)	1.198 (72,26%)	460 (27,74%)	1.658
AVC (Não)	65.722 (94,69%)	3.680 (5,31%)	69.402
Asma (Sim)	3.178 (90,16%)	347 (9,84%)	3.525
Asma (Não)	63.742 (94,38%)	3.793 (5,62%)	67.535
Depressão (Sim)	5.943 (87,59%)	842 (12,41%)	6.785
Depressão (Não)	60.977 (95,02%)	3.298 (4,98%)	64.275
Câncer (Sim)	1.771 (86,14%)	285 (13,86%)	2.056
Câncer (Não)	65.149 (94,41%)	3.855 (5,59%)	69.004
Total	66.920	4.140	71.060

Fonte: Elaborado pelos autores.

Entre os indivíduos não praticantes de atividade física, 6,81% relataram diagnóstico de doença cardíaca, enquanto entre aqueles que praticam esse percentual foi de 4,48%, indicando uma maior ocorrência da condição entre os que não praticam. O histórico de tabagismo também mostra diferenciação: 8,72% dos participantes que já fumaram possuem diagnóstico, contra 4,53% entre os que nunca fumaram, sugerindo uma possível associação entre tabagismo prévio e risco cardiovascular.

A presença de hipertensão evidencia ainda maior impacto, com 13,06% dos hipertensos apresentando doença cardíaca, em contraste com 2,87% dos não hipertensos. Situação semelhante é observada em participantes com diabetes, nos quais 14,71% relataram diagnóstico de doença cardíaca, comparado a 5,17% entre os não diabéticos. De forma análoga, o colesterol elevado está associado a maior ocorrência da condição (13,31% versus 4,31%), reforçando a relevância dessas variáveis para a modelagem preditiva.

Eventos vasculares prévios também se destacam: 27,74% dos indivíduos com histórico de AVC apresentaram doença cardíaca, valor significativamente superior aos 5,60% observados entre aqueles sem AVC. Entre os asmáticos, 9,84% relataram diagnóstico de doença cardíaca, contra 5,95% entre não asmáticos, sugerindo que a asma pode estar associada a maior risco cardiovascular. A presença de depressão também evidencia diferenças proporcionais importantes

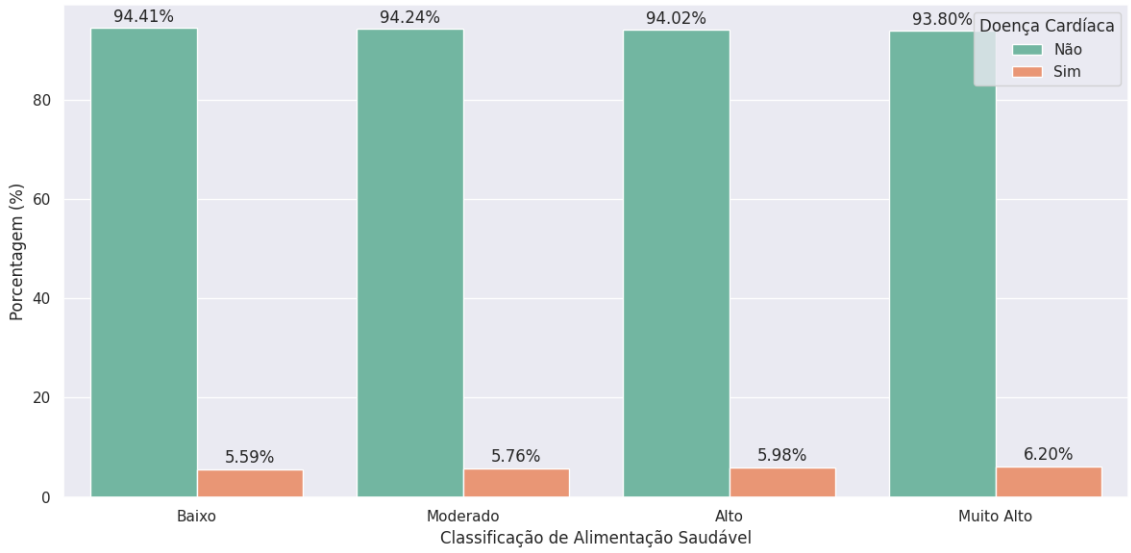
(12,41% versus 5,40%), indicando que fatores de saúde mental merecem atenção ao se investigar o risco cardíaco.

Por fim, entre indivíduos com histórico de câncer, 13,86% apresentaram diagnóstico de doença cardíaca, valor superior aos 5,59% entre aqueles sem câncer, reforçando a relevância clínica da variável, mesmo considerando a proporção menor de casos na amostra.

Em suma, os dados da Tabela 6 destacam a importância de variáveis comportamentais, clínicas e de histórico de doenças prévias como elementos relevantes para compreender o perfil de risco de doença cardíaca, fornecendo subsídios para as etapas subsequentes de modelagem e análise preditiva.

A Figura 5 apresenta a distribuição dos participantes com e sem diagnóstico de doença cardíaca, estratificada segundo as categorias de alimentação saudável.

Figura 5 – Distribuição da doença cardíaca segundo categorias de alimentação saudável na amostra.



Fonte: Elaborado pelos autores.

Observa-se que a maioria dos participantes, independentemente da categoria de alimentação, não apresenta diagnóstico de doença cardíaca. A análise visual do gráfico sugere que a categoria *Alto* apresenta a menor proporção relativa de casos positivos para doença cardíaca, enquanto a categoria *Moderado* demonstra maior concentração de diagnósticos positivos. Tal padrão aponta para uma possível associação entre melhor qualidade alimentar e menor risco cardiovascular, corroborando a literatura que destaca a importância de hábitos alimentares saudáveis na prevenção de doenças crônicas. Assim, a variável *Alimentação Saudável* será incluída nas próximas etapas da modelagem preditiva.

Dando continuidade à análise dos hábitos alimentares dos participantes, foi investigada

a variável **Alimentação Não Saudável**, categorizada em quatro níveis: “Baixo”, “Moderado”, “Alto” e “Muito Alto”. A Tabela 7 apresenta a distribuição dos participantes segundo essa variável.

Tabela 7 – Distribuição da doença cardíaca segundo níveis de alimentação não saudável na amostra da PNS (2019).

Nível de Alimentação Não Saudável	Não tem DC	Tem DC	Total
Baixo	26.565 (93,21%)	1.934 (6,79%)	28.499
Muito Alto	15.251 (95,54%)	711 (4,46%)	15.962
Moderado	14.105 (94,05%)	892 (5,95%)	14.997
Alto	10.999 (94,80%)	603 (5,20%)	11.602
Total	66.920	4.140	71.060

Fonte: Elaborado pelos autores.

Observa-se que a maioria dos indivíduos se concentra nas categorias “Baixo” (40,11%) e “Moderado” (21,10%), seguidas pelas categorias “Muito Alto” (22,46%) e “Alto” (16,33%).

Embora a categoria “Baixo” apresente o maior número absoluto de participantes com diagnóstico positivo para doença cardíaca (1.934), é importante considerar o percentual relativo dentro de cada grupo. Nesse aspecto, a prevalência de doença cardíaca é maior entre os indivíduos classificados como “Moderado” (5,95%), seguido de “Baixo” (6,79%), “Alto” (5,19%) e “Muito Alto” (4,46%).

Esse comportamento, à primeira vista contraintuitivo, sugere que outros fatores podem estar interferindo na relação entre alimentação não saudável e risco cardiovascular, como estilo de vida, nível de atividade física, presença de comorbidades e autocuidado. Dessa forma, a variável será considerada nas próximas etapas de modelagem, respeitando sua possível interação com outros atributos do conjunto de dados.

Dando continuidade à análise dos fatores associados à Doença Cardíaca, foi incluída a variável **Índice de Massa Corporal (IMC)**, categorizada segundo as faixas tradicionalmente adotadas na literatura médica: Abaixo do Peso, Peso Normal, Sobrepeso, Obesidade Grau I, Obesidade Grau II e Obesidade Grau III.

A Tabela 8 apresenta a distribuição dos casos de Doença Cardíaca de acordo com as diferentes categorias de IMC. Observa-se que a maior parte dos participantes encontra-se nas faixas de **Peso Normal** (38,34%), **Sobrepeso** (37,96%) e **Obesidade Grau I** (15,93%). As demais categorias representam proporções menores da amostra.

Tabela 8 – Distribuição dos casos de Doença Cardíaca segundo categorias do IMC.

Categoria de IMC	Não tem Doença Cardíaca	Tem Doença Cardíaca	Total
Abaixo do Peso	1.401 (94,92%)	75 (5,08%)	1.476
Peso Normal	25.866 (94,93%)	1.381 (5,07%)	27.247
Sobrepeso	25.398 (94,17%)	1.572 (5,83%)	26.970
Obesidade I	10.538 (93,07%)	785 (6,93%)	11.323
Obesidade II	2.874 (92,14%)	245 (7,86%)	3.119
Obesidade III	843 (91,14%)	82 (8,86%)	925
Total	66.920	4.140	71.060

Fonte: Elaborado pelos autores.

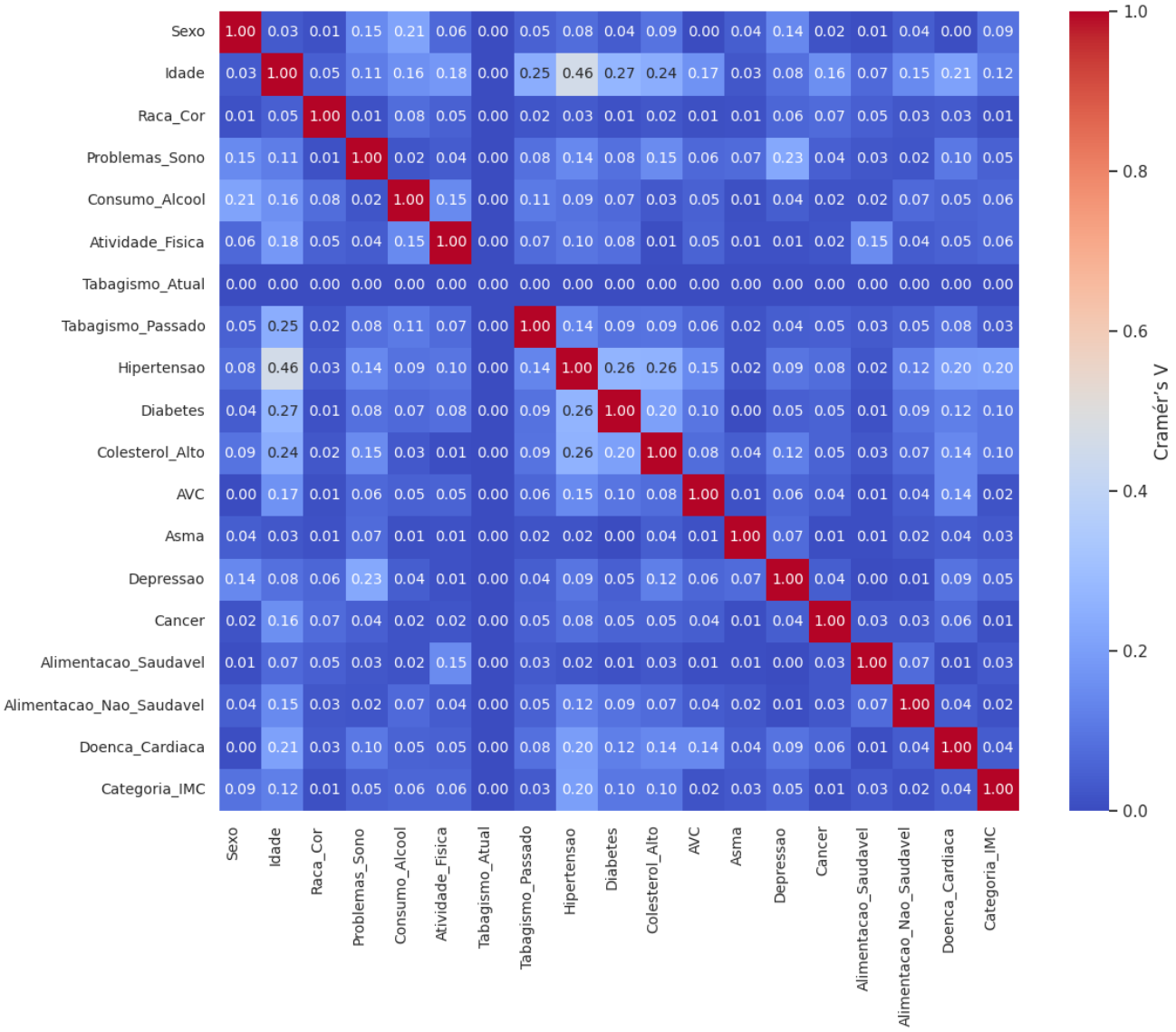
Em relação à presença de Doença Cardíaca, verifica-se um padrão crescente nas proporções de casos positivos à medida que o IMC se eleva. Na faixa de **Peso Normal**, 5,07% dos indivíduos relataram diagnóstico da doença. Já entre os indivíduos classificados com **Obesidade Grau I**, essa proporção sobe para 6,93%. O comportamento se repete nas faixas mais elevadas: **Obesidade Grau II** com 7,85% e **Obesidade Grau III** com 8,86% dos casos positivos.

Esse padrão evidencia uma tendência de aumento da prevalência de doença cardíaca conforme o aumento do IMC. Dessa forma, a variável *IMC*, especialmente em sua forma categórica, será considerada nas próximas etapas de modelagem preditiva, dada sua relevância clínica e estatística.

A Figura 6 apresenta a matriz de associação entre as variáveis, mensurada pelo coeficiente de Cramér's V. Considerando a variável-alvo **Doença Cardíaca**, observa-se que as associações mais expressivas ocorrem com **Idade** ($V = 0{,}21$), **Hipertensão** ($V = 0{,}20$), **Colesterol Alto** ($V = 0{,}14$), **AVC** ($V = 0{,}14$), **Diabetes** ($V = 0{,}12$) e **Problemas de Sono** ($V = 0{,}10$).

Embora de intensidade fraca a moderada, essas associações reforçam o papel reconhecido de fatores clínicos — como idade avançada, hipertensão, colesterol elevado e diabetes — no risco cardiovascular. Adicionalmente, condições como problemas de sono e histórico de AVC, apesar de apresentarem correlações mais discretas, também serão considerados nessa pesquisa e podem contribuir de forma complementar em análises preditivas multivariadas.

Figura 6 – Matriz de associação entre variáveis (Cramér's V).



Fonte: Elaborado pelo autor.

6.2 Construção, Avaliação e Comparação dos Modelos

Neste estudo, foram construídos modelos de Regressão Logística, K-Nearest Neighbors (KNN) e Random Forest para a previsão da variável alvo. Inicialmente, os modelos foram desenvolvidos considerando todas as variáveis disponíveis, com exceção de Tabagismo Atual, que não apresentou correlação alguma com a variável alvo. Por conta disso, a mesma foi descartada.

Em seguida, para identificar os fatores mais fortemente associados à presença de Doença Cardíaca, foi utilizada a estatística V de Cramer. Todas as variáveis com associação igual

ou superior a 0,10 foram selecionadas. Com base nessa análise, as seguintes variáveis foram escolhidas para a construção de novos modelos, conforme apresentado na Tabela 9.

Tabela 9 – Associação entre variáveis preditoras e Doença Cardíaca (V de Cramer).

Variável	**V de Cramer**
Idade	0,21
Hipertensão	0,20
Colesterol alto	0,14
AVC	0,14
Diabetes	0,12
Problemas de sono	0,10

Fonte: Elaborado pelo autor.

Os resultados indicam que **idade** e **hipertensão** apresentam a maior associação com a presença de Doença Cardíaca, refletindo seu papel crucial como fatores de risco. Colesterol alto, AVC e diabetes mostraram associações moderadas, enquanto problemas de sono apresentaram menor correlação, mas ainda podem influenciar o risco cardiovascular.

Estudos prévios corroboram esses resultados. A hipertensão arterial é amplamente reconhecida como um dos principais fatores de risco para Doença Cardíaca, estando associada a maior incidência de eventos cardiovasculares, incluindo infarto e insuficiência cardíaca (OLIVEIRA et al., 2022; MALTA et al., 2022). A idade, por sua vez, está relacionada a alterações fisiológicas que aumentam a vulnerabilidade do sistema cardiovascular, sendo um preditor consistente de morbimortalidade em diversos estudos populacionais (CARLUCCI et al., 2013).

A dislipidemia, medida aqui pelo colesterol alto, tem sido associada à aterosclerose e ao desenvolvimento de doença arterial coronariana (XAVIER et al., 2013), enquanto o histórico de AVC reflete a presença de comorbidades vasculares que podem agravar o risco cardíaco (NASCIMENTO; GOMES; SARDINHA, 2011). Diabetes mellitus, por sua vez, contribui para danos micro e macrovasculares, elevando a probabilidade de eventos cardíacos (MALTA et al., 2022).

Embora problemas de sono tenham apresentado menor associação, evidências sugerem que distúrbios crônicos do sono, como a insônia, podem agravar fatores metabólicos e cardiovasculares, constituindo um fator adicional de risco (SOFI et al., 2014).

Os resultados obtidos pelo V de Cramer confirmam a literatura, destacando, nesse conjunto de dados, idade e hipertensão como os fatores mais fortemente associados à Doença Cardíaca, enquanto colesterol alto, AVC e diabetes desempenham papéis complementares, e problemas de sono podem contribuir de forma secundária para o risco cardiovascular.

Vale destacar que essas associações refletem os dados disponíveis na amostra da Pesquisa Nacional de Saúde (PNS) 2019. Ou seja, os fatores identificados como mais fortemente associados à Doença Cardíaca neste estudo — idade, hipertensão, colesterol alto, AVC, diabetes e problemas

de sono — representam o padrão observado nesta população específica, podendo variar em outros contextos ou coortes populacionais.

Para os modelos de Regressão Logística e KNN, foram aplicadas técnicas de pré-processamento, incluindo balanceamento de dados (SMOTE), padronização das variáveis numéricas (StandardScaler) e codificação de variáveis categóricas (OneHotEncoder). Já para o Random Forest, por ser um modelo menos sensível a distância, foi aplicado somente o balanceamento dos dados.

Os modelos foram então comparados entre si para identificar aquele com melhor desempenho preditivo, considerando as métricas de avaliação propostas nesse estudo.

6.2.1 Regressão Logística

6.2.1.1 Modelo de Regressão Logística sem Seleção de Variáveis

A Tabela 10 apresenta o desempenho do modelo de regressão logística na predição de Doença Cardíaca. Foram avaliadas as métricas de precisão (*precision*), sensibilidade (*recall*), F1-score e suporte para cada classe, além da acurácia global e da área sob a curva ROC (AUC-ROC).

Tabela 10 – Desempenho da Regressão Logística na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,96	0,71	0,82	20.076
Com Doença Cardíaca (1)	0,10	0,54	0,18	1.242
Acurácia		0,70		
AUC-ROC		0,6822		

Fonte: Elaborado pelo autor.

Mesmo com o ajuste de hiperparâmetros[1] e o balanceamento da base de dados, o modelo apresenta forte discrepância entre as classes em termos de precisão e F1-score. A classe majoritária (indivíduos sem Doença Cardíaca) mantém alta precisão (0,96) e F1-score (0,82), enquanto a classe minoritária (indivíduos com Doença Cardíaca) ainda apresenta baixa precisão (0,10) e F1-score (0,18), embora o recall de 0,54 indique que mais da metade dos casos positivos foi corretamente identificada.

A acurácia global de 0,70, embora razoável, não reflete adequadamente o desempenho do modelo na identificação de casos positivos, reforçando a necessidade de analisar métricas específicas por classe. O AUC-ROC de 0,6822 demonstra que o modelo consegue discriminar moderadamente entre indivíduos com e sem Doença Cardíaca, mas que sua capacidade preditiva ainda é limitada.

[1] Hiperparâmetros são parâmetros externos ao modelo que não são aprendidos diretamente dos dados, mas são configurados para controlar o processo de aprendizado. Sua otimização melhora o desempenho do modelo.

Em síntese, a regressão logística, mesmo com ajustes finos de hiperparâmetros e pré-processamento adequado, se mostra eficiente na identificação de indivíduos saudáveis, mas apresenta limitações na predição de casos positivos.

6.2.1.2 Modelo de Regressão Logística com Seleção de Variáveis

O modelo de Regressão Logística, construído com seleção de variáveis, foi avaliado quanto à sua capacidade de classificar indivíduos com e sem Doença Cardíaca, e seu desempenho é apresentado na Tabela 11

Tabela 11 – Desempenho do modelo de Regressão Logística com Seleção de Variáveis na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,98	0,70	0,82	20.076
Com Doença Cardíaca (1)	0,14	0,76	0,23	1.242
Acurácia		0,70		
AUC-ROC		0,7859		

Fonte: Elaborado pelo autor.

A análise das métricas evidencia que o modelo apresenta desempenho fortemente desigual entre as classes. A classe majoritária (indivíduos sem Doença Cardíaca) mantém elevada precisão (0,98), mas seu recall (0,70) indica que ainda há um percentual significativo de falsos negativos, ou seja, indivíduos com Doença Cardíaca que foram classificados incorretamente como saudáveis.

Já a classe minoritária (indivíduos com Doença Cardíaca) apresenta recall de 0,76, o que significa que o modelo consegue capturar a maioria dos casos positivos, uma melhora significativa em relação aos modelos anteriores. No entanto, a precisão extremamente baixa (0,14) evidencia que muitas predições positivas são incorretas, resultando em um grande número de falsos positivos. O F1-score de 0,23 reflete esse desequilíbrio entre recall e precision.

A acurácia global de 0,70, embora razoável, não reflete adequadamente a capacidade do modelo em identificar casos positivos, enquanto o AUC-ROC de 0,7859 indica que o modelo tem boa capacidade de discriminação geral entre indivíduos com e sem Doença Cardíaca.

Em síntese, o modelo, mesmo com pré-processamento, balanceamento via SMOTE e ajuste de hiperparâmetros, apresenta forte sensibilidade para a classe minoritária, sendo útil, por exemplo, como uma possível ferramenta de triagem, mas sua baixa precisão limita a confiabilidade das predições individuais.

6.2.2 KNN

6.2.2.1 Modelo KNN sem Seleção de Variáveis

O modelo KNN sem seleção de variáveis foi avaliado quanto à sua capacidade de classificar indivíduos com e sem Doença Cardíaca. O desempenho do modelo é apresentado na Tabela 12.

Tabela 12 – Desempenho do modelo KNN sem seleção de variáveis na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,95	0,82	0,88	20.076
Com Doença Cardíaca (1)	0,10	0,32	0,15	1.242
Acurácia		0,79		
AUC-ROC		0,6157		

Fonte: Elaborado pelo autor.

A análise das métricas indica que o modelo apresenta desempenho desigual entre as classes. A classe majoritária (indivíduos sem Doença Cardíaca) apresenta elevada precisão (0,95) e F1-score (0,88), enquanto a classe minoritária (indivíduos com Doença Cardíaca) mantém baixa precisão (0,10) e F1-score (0,15), com recall de 0,32, indicando que aproximadamente um terço dos casos positivos foi corretamente identificado.

A acurácia global de 0,79, embora aparentemente satisfatória, não reflete adequadamente a capacidade do modelo em detectar casos positivos, e o AUC-ROC de 0,6157 evidencia discriminação apenas moderada entre as duas classes.

Em síntese, o modelo KNN sem seleção de variáveis apresenta desempenho adequado para a classe majoritária, mas ainda insuficiente para capturar corretamente os indivíduos com Doença Cardíaca. Sua baixa sensibilidade para a classe positiva aumenta a probabilidade de falsos negativos, ou seja, casos da doença que são incorretamente classificados como saudáveis.

6.2.2.2 Modelo KNN com Seleção de Variáveis

O modelo KNN com seleção de variáveis foi avaliado quanto à sua capacidade de classificar indivíduos com e sem Doença Cardíaca. O desempenho do modelo é apresentado na Tabela 13.

Tabela 13 – Desempenho do modelo KNN com seleção de variáveis na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,95	0,98	0,96	20.076
Com Doença Cardíaca (1)	0,22	0,09	0,13	1.242
Acurácia		0,93		
AUC-ROC		0,6750		

Fonte: Elaborado pelo autor.

O alto recall na classe majoritária demonstra que o modelo é excelente para identificar pessoas saudáveis. No entanto, o baixo recall da classe minoritária (0,09) evidencia que o modelo gera um alto número de falsos negativos, falhando em capturar a maioria dos indivíduos com Doença Cardíaca.

Para a classe minoritária (1), observa-se um desempenho muito limitado: recall de apenas 0,09 e F1-score de 0,13 indicam que o modelo consegue capturar apenas uma pequena fração dos casos positivos, apesar do aumento da precisão para 0,22 em relação ao modelo sem seleção de variáveis.

A acurácia global de 0,93, embora elevada, não reflete o baixo desempenho na detecção da classe minoritária, enquanto o AUC-ROC de 0,6750 demonstra capacidade moderada de discriminação geral entre indivíduos com e sem Doença Cardíaca.

O modelo se mostra eficiente na classificação da classe majoritária, mas não apresenta desempenho satisfatório na identificação de indivíduos com Doença Cardíaca, limitando sua aplicabilidade em cenários reais.

6.2.3 Random Forest

6.2.3.1 Modelo Random Forest sem Seleção de Variáveis

O modelo Random Forest sem seleção de variáveis foi desenvolvido para avaliar sua capacidade de classificar indivíduos com e sem Doença Cardíaca. Esta técnica de aprendizado de máquina, conhecida por explorar múltiplas árvores de decisão em conjunto, permite capturar padrões complexos nos dados e fornecer uma visão robusta sobre a discriminação entre casos positivos e negativos. Os resultados obtidos estão apresentados na Tabela 14.

Tabela 14 – Desempenho do modelo Random Forest sem seleção de variáveis na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,95	0,91	0,93	20.076
Com Doença Cardíaca (1)	0,10	0,16	0,12	1.242
Acurácia		0,86		
AUC-ROC		0,7044		

Fonte: Elaborado pelo autor.

A análise das métricas evidencia que o modelo apresenta alto desempenho na classe majoritária (0), com recall de 0,91 e F1-score de 0,93, demonstrando boa capacidade de identificar indivíduos sem Doença Cardíaca. Entretanto, o desempenho para a classe minoritária (1) é muito limitado, com recall de 0,16 e F1-score de 0,12, indicando que a maioria dos casos positivos não foi corretamente identificada. Esta baixa sensibilidade para a classe positiva aumenta a probabilidade de falsos negativos, ou seja, casos da doença que são incorretamente classificados como saudáveis. A precisão da classe 1, de apenas 0,10, também revela um grande número de falsos positivos.

A acurácia global de 0,86 é relativamente alta, mas não reflete adequadamente a capacidade do modelo em detectar casos positivos, enquanto o AUC-ROC de 0,7044 indica que o modelo possui capacidade moderada de discriminação geral entre indivíduos com e sem Doença Cardíaca.

6.2.3.2 Modelo Random Forest com Seleção de Variáveis

Para avançar na análise, o modelo Random Forest com seleção de variáveis foi avaliado quanto à sua capacidade de classificar indivíduos com e sem Doença Cardíaca, considerando apenas as variáveis mais correlacionadas com a variável alvo. Esta abordagem permite investigar como a redução do número de preditores impacta a eficiência do algoritmo em capturar padrões relevantes e discriminar corretamente entre casos positivos e negativos. Os resultados obtidos estão apresentados na Tabela 15.

A análise das métricas evidencia que o modelo mantém alto desempenho na classe majoritária (0), com recall de 0,73 e F1-score de 0,83, demonstrando boa capacidade de identificação de indivíduos sem Doença Cardíaca. Entretanto, esse recall mais baixo em relação ao modelo anterior indica que há um aumento no número de falsos positivos.

Para a classe minoritária (1), observa-se melhora no recall (0,60) em relação ao modelo sem seleção de variáveis, indicando que uma parcela maior dos casos positivos foi corretamente identificada. No entanto, a precisão ainda é limitada (0,12), resultando em um número considerável de falsos positivos, e o F1-score permanece baixo (0,20), refletindo o trade-off entre sensibilidade e precisão. A baixa sensibilidade para a classe positiva significa que o modelo ainda produz uma

taxa elevada de falsos negativos, ou seja, falha em identificar uma parte significativa dos casos da doença.

Tabela 15 – Desempenho do modelo Random Forest com seleção de variáveis na predição de Doença Cardíaca.

Classe	**Precision**	**Recall**	**F1-Score**	**Suporte**
Sem Doença Cardíaca (0)	0,97	0,73	0,83	20.076
Com Doença Cardíaca (1)	0,12	0,60	0,20	1.242
Acurácia		0,72		
AUC-ROC		0,6773		

Fonte: Elaborado pelo autor.

A acurácia global de 0,72 reflete parcialmente a capacidade do modelo em classificar corretamente a população geral, enquanto o AUC-ROC de 0,6773 indica capacidade moderada de discriminação entre indivíduos com e sem Doença Cardíaca.

O modelo melhora a detecção da classe minoritária em comparação ao modelo completo, mas ainda apresenta limitações importantes na identificação de casos positivos, exigindo cautela em sua aplicação na vida real.

6.2.4 Comparação dos Modelos

A fim de avaliar de forma abrangente o desempenho dos modelos construídos, esta seção apresenta uma comparação entre Regressão Logística (RL), KNN e Random Forest (RF), considerando versões com e sem seleção de variáveis. Foram consolidadas métricas de *precision*, *recall*, F1-score, acurácia e AUC-ROC para ambas as classes, permitindo identificar quais modelos apresentam melhor capacidade de discriminação e sensibilidade para a detecção de casos positivos de Doença Cardíaca.

Tabela 16 – Comparação do desempenho dos modelos na predição de Doença Cardíaca.

Modelo	**Seleção de Variáveis**	**Classe**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
RL	Não	0	0,96	0,71	0,82	0,6822
RL	Não	1	0,10	0,54	0,18	
RL	Sim	0	**0,98**	0,70	0,82	**0,7859**
RL	Sim	1	0,14	**0,76**	**0,23**	
KNN	Não	0	0,95	0,82	0,88	0,6157
KNN	Não	1	0,10	0,32	0,15	
KNN	Sim	0	0,95	**0,98**	**0,96**	0,6750
KNN	Sim	1	**0,22**	0,09	0,13	
RF	Não	0	0,95	0,91	0,93	0,7044
RF	Não	1	0,10	0,16	0,12	
RF	Sim	0	0,97	0,73	0,83	0,6773
RF	Sim	1	0,12	0,60	0,20	

Fonte: Elaborado pelos autores.

A partir da Tabela 16, observa-se que:

- **Classificação da classe majoritária (0)**: Todos os modelos apresentam alta precisão e F1-score, com KNN com seleção de variáveis atingindo o maior recall (0,98) e F1-score (0,96). Isso demonstra que qualquer modelo é eficiente na identificação de indivíduos sem Doença Cardíaca.
- **Classificação da classe minoritária (1)**: A Regressão Logística com seleção de variáveis apresenta o maior recall (0,76), capturando a maioria dos casos positivos, embora a precisão seja baixa (0,14), resultando em falsos positivos. O Random Forest com seleção de variáveis também apresenta recall relativamente elevado (0,60), mas com precisão menor (0,12). KNN, mesmo com seleção de variáveis, apresenta recall muito baixo (0,09), sendo pouco efetivo para detectar casos positivos.
- **Acurácia e AUC-ROC**: A acurácia global tende a ser mais influenciada pela classe majoritária, enquanto a AUC-ROC mostra capacidade moderada de discriminação em todos os modelos. A RL com seleção de variáveis apresenta a maior AUC-ROC (0,7859), indicando melhor separação entre classes.

Considerando o objetivo de identificar corretamente indivíduos com Doença Cardíaca, a **Regressão Logística com seleção de variáveis** se mostra a mais adequada, devido ao seu maior recall e AUC-ROC, tornando-se uma ferramenta promissora para triagem clínica, apesar da baixa precisão individual. Os demais modelos apresentam desempenho satisfatório na classe majoritária, mas são menos confiáveis na detecção da classe minoritária.

6.3 Relevância e Limitações na Detecção de Doenças Cardíacas

A aplicação de modelos estatísticos e de aprendizado de máquina na predição de Doença Cardíaca demonstra relevância ao fornecer uma ferramenta complementar para a identificação de indivíduos em risco e a priorização de intervenções clínicas. Técnicas como Regressão Logística, Random Forest e K-Nearest Neighbors permitem capturar padrões complexos nos dados.

Apesar disso, diversas limitações devem ser consideradas. O desbalanceamento entre classes impacta diretamente a capacidade de detecção de casos positivos, mesmo quando estratégias de balanceamento são aplicadas. Além disso, a baixa precisão em alguns modelos, que resulta em um número considerável de falsos positivos, exige cautela em sua aplicação prática. Embora a classificação incorreta de um indivíduo saudável como doente (falso positivo) não represente um risco imediato à saúde, ela pode levar a custos desnecessários e ansiedade para o paciente, sendo um aspecto crítico em um contexto de triagem. A disponibilidade e qualidade dos dados também influenciam a robustez dos modelos, e preditores não contemplados, como a ausência de uma variável sobre histórico familiar de doenças cardiovasculares, podem reduzir

a capacidade de generalização dos resultados. A complexidade do comportamento humano, fatores psicossociais e condições ambientais, muitas vezes não capturados nos conjuntos de dados, também constituem desafios à precisão das predições.

Em suma, embora os modelos ofereçam insights valiosos e potencial para apoio à tomada de decisão clínica, seu uso deve ser interpretado com cautela, reconhecendo as limitações inerentes aos dados e à modelagem.

6.4 Trabalhos Futuros

Apesar dos esforços em pré-processamento, balanceamento de dados via SMOTE e ajuste de hiperparâmetros, os modelos avaliados apresentaram limitações na identificação da classe minoritária (indivíduos com Doença Cardíaca). Dessa forma, diversas estratégias podem ser exploradas em trabalhos futuros para aprimorar o desempenho e a robustez das previsões.

Além do SMOTE, outras abordagens de balanceamento podem ser testadas, como ADASYN (Adaptive Synthetic Sampling), Borderline-SMOTE e combinação de over- e under-sampling. A utilização de técnicas de ensemble voltadas a classes desbalanceadas, como Balanced Random Forest ou EasyEnsemble, também pode contribuir para a melhoria da detecção de casos positivos.

Experimentar diferentes métodos de normalização, padronização e transformação de variáveis pode impactar positivamente o desempenho dos algoritmos. A criação de novas variáveis derivadas, interações entre preditores ou agregações baseadas em conhecimento clínico podem fornecer informações adicionais relevantes para a predição.

A construção de modelos híbridos ou de ensembles (stacking, blending, voting) pode combinar as forças de diferentes algoritmos, como Random Forest, KNN e Regressão Logística, potencializando o recall da classe minoritária sem comprometer significativamente a precisão. Testar métodos de aprendizado profundo (deep learning) em conjunto com técnicas tradicionais de machine learning também é uma possibilidade para exploração futura.

Testes adicionais em diferentes conjuntos de dados, bem como a validação cruzada estratificada ou por blocos temporais, podem oferecer insights sobre a estabilidade e generalização dos modelos. Além disso, a incorporação de dados longitudinalmente coletados permitiria avaliar predições de risco ao longo do tempo.

Considerar variáveis adicionais relacionadas a fatores psicossociais, comportamentais e ambientais pode aumentar a capacidade de identificação de indivíduos em risco. Integrações com dados de hábitos de sono, níveis de estresse, sedentarismo ou histórico familiar podem enriquecer o modelo e aproximar as análises da realidade clínica.

Trabalhos futuros podem focar em otimizar métricas específicas de interesse, como recall da classe minoritária ou F1-score balanceado, utilizando abordagens de cost-sensitive learning

ou ajustando limiares de decisão para priorizar a detecção de indivíduos com Doença Cardíaca.

7 CONCLUSÃO

O presente estudo teve como objetivo geral construir um modelo capaz de prever a presença de doença cardíaca e identificar os principais fatores de risco associados, utilizando técnicas estatísticas e de aprendizado de máquina. Para atingir esse objetivo, explorou-se o comportamento das variáveis clínicas, comportamentais e demográficas, além de ajustar modelos preditivos de classificação, incluindo Regressão Logística, Random Forest e K-Nearest Neighbors (KNN).

Os resultados obtidos demonstraram que, mesmo com pré-processamento, balanceamento via SMOTE e ajuste de hiperparâmetros, os modelos apresentaram desempenho limitado na identificação da classe minoritária (indivíduos com doença cardíaca). A Regressão Logística com seleção de variáveis mostrou maior sensibilidade para a detecção de casos positivos, mas com baixa precisão, enquanto os modelos Random Forest e KNN apresentaram melhor desempenho na classificação da classe majoritária, porém ainda insuficiente na identificação de indivíduos com a condição.

A análise dos fatores de risco indicou que idade, hipertensão, diabetes, colesterol alto e AVC estão associados à presença de doença cardíaca, corroborando achados da literatura nacional e internacional. Problemas de sono apresentaram menor associação, mas evidências sugerem que distúrbios crônicos do sono podem constituir um fator adicional de risco.

Embora os modelos ajustados não tenham atingido a performance ideal, o estudo evidencia a aplicabilidade de técnicas de aprendizado de máquina para apoiar a triagem e a identificação de padrões em saúde cardiovascular.

Como trabalhos futuros, recomenda-se:

- Explorar outras técnicas de balanceamento de classes, como ADASYN e undersampling adaptativo, para melhorar a detecção da classe minoritária;
- Avaliar métodos de pré-processamento alternativos, incluindo engenharia de atributos e transformação de variáveis contínuas;
- Testar abordagens de ensemble e combinação de modelos, como Voting Classifier ou Stacking, para potencializar o desempenho preditivo;
- Investigar a integração de dados longitudinais ou clínicos adicionais, incluindo exames laboratoriais, para aumentar a robustez dos modelos;
- Desenvolver métricas customizadas de avaliação para cenários de triagem, considerando o impacto clínico de falsos negativos.

Este trabalho demonstra a viabilidade de aplicar métodos estatísticos e de aprendizado de máquina na predição de doença cardíaca, aponta as limitações atuais e sugere caminhos para refinamento e aprofundamento em pesquisas futuras.

REFERÊNCIAS

BERGAMINI, M. et al. Mapping risk of ischemic heart disease using machine learning in a brazilian state. *PLoS ONE*, v. 15, n. 12, p. e0243558, 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0243558>. Citado na página 28.

BICHARA, J. L. et al. Index of social vulnerability and mortality from ischemic heart disease and cerebrovascular diseases in brazil from 2000 to 2021. *Arquivos Brasileiros de Cardiologia*, v. 122, n. 8, p. e20240428, 2024. Citado na página 18.

BRASIL, M. d. S. *Pesquisa Nacional de Saúde 2013: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal – Brasil, grandes regiões e unidades da federação*. Rio de Janeiro: IBGE, 2014. Acesso em: 09 ago. 2025. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv91110.pdf>. Citado na página 15.

BRIJITH, A. *Data Preprocessing for Machine Learning*. 2023. International Center for AI and Cyber Security Research and Innovations (CCRI), Asia University. Disponível em: <https://www.researchgate.net/publication/375003512_Data_Preprocessing_for_Machine_Learning>. Acesso em: 9 ago. 2025. Citado 3 vezes nas páginas 21, 22 e 28.

CARLUCCI, E. M. d. S. et al. Obesidade e sedentarismo: fatores de risco para doença cardiovascular. *Ciência & Saúde Coletiva*, Brasília, v. 18, n. 11, p. 3171–3180, 2013. Acesso em: 03 ago. 2025. Disponível em: <https://bvsms.saude.gov.br/bvs/artigos/ccs/obesidade_sedentarismo_fatores_risco_cardiovascular.pdf>. Citado 2 vezes nas páginas 19 e 46.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. *Semi-supervised Learning*. Cambridge: MIT Press, 2006. Citado na página 20.

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Citado na página 24.

DELPINO, F. M. et al. Predicting all-cause mortality with machine learning among brazilians aged 50 and over: results from the brazilian longitudinal study of ageing (elsi-brazil). *npj Aging*, v. 11, p. 22, 2025. Disponível em: <https://doi.org/10.1038/s41514-025-00210-7>. Citado na página 28.

ELLIOTT, P. et al. Classification of the cardiomyopathies: a position statement from the european society of cardiology working group on myocardial and pericardial diseases. *European Heart Journal*, v. 35, n. 39, p. 2733–2741, 2014. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/17916581/>. Citado na página 14.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Acesso em: 11 ago. 2025. Disponível em: <https://doi.org/10.1016/j.patrec.2005.10.010>. Citado 2 vezes nas páginas 27 e 28.

FLORESTI, F. Cerca de 400 mil pessoas morreram em 2022 no brasil por problemas cardiovasculares. fev 2024. Acesso em: 25 jun. 2025. Disponível em: <https://revistapesquisa.fapesp.br/cerca-de-400-mil-pessoas-morreram-em-2022-no-brasil-por-problemas-cardiovasculares/>. Citado 2 vezes nas páginas 14 e 18.

GIL, A. C. *Como elaborar projetos de pesquisa*. 6. ed. São Paulo: Atlas, 2019. Citado na página 29.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge: MIT Press, 2016. Citado na página 20.

HAIR, J. F. et al. *Análise multivariada de dados*. 6. ed. Porto Alegre: Bookman, 2009. Citado na página 32.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, Radiological Society of North America, v. 143, n. 1, p. 29–36, 1982. Acesso em: 11 ago. 2025. Disponível em: <https://doi.org/10.1148/radiology.143.1.7063747>. Citado 2 vezes nas páginas 27 e 28.

HASANAH, U.; SOLEH, A. M.; SADIK, K. Effect of random under sampling, oversampling, and smote on cardiovascular disease prediction. *JMSK*, v. 10, n. 2, p. 123–135, 2024. Acesso em: 9 ago. 2025. Disponível em: <https://journal.unhas.ac.id/index.php/jmsk/article/view/35552>. Citado na página 22.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009. Citado 2 vezes nas páginas 22 e 23.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009. Citado 2 vezes nas páginas 23 e 24.

IBGE. *Pesquisa Nacional de Saúde 2019: informações sobre domicílios, características demográficas, socioeconômicas e de saúde*. Rio de Janeiro: IBGE, 2019. Acesso em: 09 ago. 2025. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101703.pdf>. Citado na página 15.

JAMES, G. et al. *An Introduction to Statistical Learning*. New York: Springer, 2013. Citado 6 vezes nas páginas 19, 22, 24, 25, 26 e 27.

JANUARY, C. T. et al. 2019 aha/acc/hrs focused update of the 2014 aha/acc/hrs guideline for the management of patients with atrial fibrillation. *Circulation*, v. 140, n. 2, p. e125–e151, 2019. Disponível em: <https://doi.org/10.1161/CIR.0000000000000665>. Citado na página 14.

KRITTANAWONG, C. et al. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, v. 69, n. 21, p. 2657–2664, 2017. Acesso em: 21 jun. 2025. Citado na página 15.

LIU, Y. et al. Trends analysis of the global burden of hypertensive heart disease from 1990 to 2021: a population-based study. *BMC Public Health*, v. 25, n. 23389, 2025. Acesso em: 03 ago. 2025. Disponível em: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-025-23389-6>. Citado na página 18.

MALTA, D. C. et al. Hipertensão arterial e fatores associados: Pesquisa nacional de saúde, 2019. *Revista de Saúde Pública*, São Paulo, v. 56, n. 24, 2022. Acesso em: 03 ago. 2025. Disponível em: <https://www.scielo.br/j/rsp/a/mncyrfyzjH77bgymWfSBCkK/?format=pdf&lang=pt>. Citado 4 vezes nas páginas 15, 18, 21 e 46.

MANSUR, A. d. P.; FAVARATO, D. Mortalidade por doenças cardiovasculares no brasil e na região metropolitana de são paulo: atualização 2011. *Arquivos Brasileiros de Cardiologia*, v. 99, n. 2, p. 755–761, august 2012. Acesso em: 21 jun. 2025. Citado na página 14.

MITCHELL, T. M. The discipline of machine learning. Pittsburgh, 2006. Citado na página 19.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012. Citado na página 20.

NASCIMENTO, J. S. d.; GOMES, B.; SARDINHA, A. H. d. L. Fatores de risco modificáveis para as doenças cardiovasculares em mulheres com hipertensão arterial. *Enfermería Global*, v. 23, p. 102–110, 2011. Acesso em: 03 ago. 2025. Disponível em: <https://enfispo.es/descarga/articulo/8802047.pdf>. Citado 2 vezes nas páginas 19 e 46.

NECIOSUP-BOLAñOS, B. R.; CIEZA-MOSTACERO, S. E. A review of machine learning for heart disease prediction: the heart of artificial intelligence. *International Journal of Advanced Computer Science and Applications*, v. 15, n. 12, 2024. Disponível em: <https://thesai.org/Downloads/Volume15No12/Paper_8-The_Heart_of_Artificial_Intelligence_A_Review.pdf>. Citado 2 vezes nas páginas 21 e 27.

OLIVEIRA, G. M. M. et al. Posicionamento sobre a saúde cardiovascular da mulher – 2022. *Arquivos Brasileiros de Cardiologia*, v. 119, n. 5, p. 815–882, nov 2022. Acesso em: 03 ago. 2025. Citado 2 vezes nas páginas 18 e 46.

Organização Pan-Americana da Saúde. *Doenças cardiovasculares continuam sendo principal causa de morte nas Américas*. Washington, DC: [s.n.], 2021. Acesso em: 03 ago. 2025. Disponível em: <https://www.paho.org/pt/noticias/29-9-2021-doencas-cardiovasculares-continuam-sendo-principal-causa-morte-nas-americas>. Citado na página 19.

PONIKOWSKI, P. et al. 2016 esc guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, v. 37, n. 27, p. 2129–2200, 2016. Citado na página 14.

SHIMIZU, G. Y. et al. Machine learning-based risk prediction for major adverse cardiovascular events in a brazilian hospital: Development, external validation, and interpretability. *PLoS ONE*, v. 19, n. 10, p. e0311719, 2024. Disponível em: <https://doi.org/10.1371/journal.pone.0311719>. Citado 4 vezes nas páginas 20, 21, 27 e 28.

SHISHEHBORE, F.; AWAN, Z. *Enhancing cardiovascular disease risk prediction with machine learning models*. 2024. Preprint. Available at: <https://arxiv.org/abs/2401.17328>. Citado na página 28.

SOFI, F. et al. Insomnia and risk of cardiovascular disease: a meta-analysis. *European Journal of Preventive Cardiology*, v. 21, n. 1, p. 57–64, 2014. Epub 2012 Aug 31. Citado na página 46.

World Health Organization. Cardiovascular diseases (cvds). *Website*, 2021. Acesso em: 25 jun. 2025. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>. Citado na página 14.

World Heart Federation. *World Heart Report 2023*. 2023. Acesso em: 21 jun. 2025. Disponível em: <https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf>. Citado 2 vezes nas páginas 14 e 18.

XAVIER, H. T. et al. V diretriz brasileira de dislipidemias e prevenção da aterosclerose. *Arquivos Brasileiros de Cardiologia*, v. 101, n. 4, p. 1–20, october 2013. Acesso em: 21 jun. 2025. Citado 2 vezes nas páginas 14 e 46.

ZHANG, Y.; LIU, X.; WANG, L. Reducing bias in coronary heart disease prediction using smote-enn. *PLOS ONE*, v. 18, n. 8, p. e0327569, 2023. Acesso em: 9 ago. 2025. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0327569>. Citado na página 22.