RECA
REVISTA ELETRÔNICA DE COMPUTAÇÃO APLICADA
Vol. 6 Nro. 1 2025

Uni-FACEF Departamento de Computação

ANÁLISE DE DADOS PARA IDENTIFICAÇÃO DE PADRÕES E PREVISÃO DE DOENÇAS CARDÍACAS

Mariana Evangelista Reis Alves
Graduanda em Engenharia de Software – Uni-FACEF
marianaevangelista2004@gmail.com

Geraldo Henrique Neto
Mestre em Ciências – FMRP-USP
geraldo.henriqueneto@gmail.com

Resumo
Considerada a principal causa de morte no mundo, as doenças cardiovasculares (DCV) atingem mais pessoas anualmente do que qualquer outra enfermidade, de acordo com a Organização Pan-Americana da Saúde. Tendo em vista a prevalência destas doenças e a abundância de dados de saúde, ainda enfrentam-se desafios na aplicação de técnicas de análise de dados para diagnósticos e previsão precoces e precisos; por isso, o presente trabalho tem como objetivo analisar o uso de técnicas de análise de dados na área da saúde, com ênfase na identificação e previsão de doenças cardíacas, pois a prevenção e o diagnóstico precoce das doenças são muito importantes para reduzir a mortalidade e os custos relacionados ao tratamento dessa patologia. Para a realização do trabalho, foram utilizadas as ferramentas *Jupyter Notebook*, como um ambiente interativo para escrever códigos e visualizar dados, e a linguagem Python e suas bibliotecas, isso através da análise de dados de um *dataset* de doenças cardíacas retirado do site *UCI Machine Learning*. Utilizando técnicas de análise exploratória de dados e modelagem preditiva, como regressão linear, foi possível chegar à conclusão de que certas variáveis convergem para a importância em relação à variável alvo, enquanto outras divergem, sendo inconsistentes. E, através do modelo de regressão linear, avaliou-se que foi possível explicar 67,81% da variância do atributo alvo. Este trabalho também apresentou contribuições que corroboram com a literatura, como o fato de que os homens são mais propensos a terem doenças cardíacas em idades mais novas (50 a 60 anos de idade).

Palavras-chave: doenças cardiovasculares. análise de dados. previsão. regressão linear.

Abstract
Considered the leading cause of death worldwide, cardiovascular diseases (CVD) affect more people annually than any other illness, according to the Pan American Health Organization. Given the prevalence of these diseases and the abundance of health data, challenges are still faced in the application of data analysis techniques for early and accurate diagnoses and predictions; therefore, this study aims to analyze the use of data analysis techniques in the healthcare field, with an emphasis on the identification and prediction of heart diseases, since prevention and early diagnosis of these diseases are very important to reduce mortality and treatment-related costs of this condition. For the development of this work, the Jupyter Notebook tool was used as an interactive environment for writing code and visualizing data, along with the Python programming language and its libraries, through the analysis of a heart disease

dataset obtained from the UCI Machine Learning Repository. Using exploratory data analysis techniques and predictive modeling, such as linear regression, it was possible to conclude that certain variables converge in importance regarding the target variable, while others diverge, being inconsistent. Furthermore, through the linear regression model, it was evaluated that it was possible to explain 67.81% of the variance of the target attribute. This study also presented contributions that align with the literature, such as the fact that men are more likely to develop heart disease at younger ages.

Keywords: *cardiovascular diseases. data analysis. prediction. linear regression.*

1 Introdução

1.1 Contextualização do tema

De acordo com a Organização Pan-Americana da Saúde, as doenças cardiovasculares são a principal causa de morte no mundo, mais pessoas morrem anualmente por essas enfermidades do que por qualquer outra causa, destacando-se um cenário alarmante que necessita de estratégias eficazes para a sua prevenção, diagnóstico precoce e tratamento adequado. Com o avanço da tecnologia e o grande volume de dados gerados na área da saúde, a análise de dados têm revelado ser uma grande ferramenta na possibilidade de identificação de padrões e fatores de risco associados a essas doenças.

1.2 Problema de Pesquisa

Apesar da prevalência das doenças cardiovasculares e da abundância de dados de saúde, ainda há desafios na aplicação eficaz de técnicas de análise de dados para um diagnóstico e previsão precoces e precisos, especialmente considerando a complexidade e a heterogeneidade dos dados clínicos.

1.3 Objetivos Geral e Específico

O objetivo geral visa examinar a aplicação de técnicas de análise de dados no contexto da saúde, focando na detecção e previsão de doenças cardíacas.

Como objetivos específicos este trabalho visa apresentar os conceitos básicos sobre doenças cardíacas, observar a importância da análise de dados no contexto da saúde, explorar ferramentas e técnicas utilizadas na análise de dados e realizar experimentos práticos com um conjunto de dados pré-definido de doenças cardíacas, avaliando os resultados obtidos.

1.4 Justificativa

A prevenção e o diagnóstico precoce das doenças cardíacas são muito importantes para reduzir a mortalidade e os custos relacionados ao tratamento dessa patologia. A análise de dados permite que, meios eficazes de identificar padrões relevantes em registros clínicos, que possam passar despercebidos, sejam oferecidos a fim de que melhores resultados sejam obtidos para melhor tratar os pacientes. A importância deste trabalho se dá pela necessidade de unir os conhecimentos e dados

da área da saúde com a tecnologia e a análise de dados, promovendo resultados que poderão ser úteis para possíveis profissionais na área da saúde em tomadas de decisões precisas e eficientes.

1.5 Estrutura do trabalho

Este trabalho está dividido em seis seções. A seção 1 apresenta a introdução do tema, abordando a contextualização, problema, objetivos, justificativa e estrutura. A seção 2 trata-se do referencial teórico, que relata os conceitos básicos de doenças cardíacas, a análise de dados no contexto da saúde, as ferramentas e técnicas para análise de dados, aborda a mineração de dados, técnicas estatísticas descritivas e comenta sobre alguns trabalhos relacionados desta área. A seção 3 traz a metodologia que informa sobre o *dataset*, descrevendo-o, escrevendo sobre a sua origem e o dicionário dos dados, juntamente com as ferramentas utilizadas para a análise sendo Python e suas bibliotecas, relatando sobre ambientes computacionais, e as etapas da análise de dados, pré-processamento dos dados, limpeza e tratamento de dados faltantes e análise exploratória de dados. A seção 4 discorre sobre o desenvolvimento do trabalho em si, começando com uma análise descritiva, tendo em seguida as estatísticas básicas das variáveis, as relações entre as variáveis, uma visualização dos dados, os gráficos e mapas de correlação, os *insights* sobre o *dataset*, padrões identificando as variáveis mais relevantes na identificação de doenças cardíacas, as limitações da análise e a modelagem preditiva e avaliação de desempenho. A seção 5 mostra os resultados e a discussão contendo os principais resultados da análise. Por fim, A seção 6 com a conclusão, que informa o resumo das contribuições do trabalho e as recomendações para trabalhos futuros.

2 Referencial Teórico

Esta seção aborda tópicos importantes para o desenvolvimento do artigo científico que servirão como base para a compreensão do mesmo, como: conceitos básicos das doenças cardíacas, informações sobre análise de dados e técnicas estatísticas descritivas.

2.1 Conceitos básicos sobre doenças cardíacas

As doenças cardiovasculares (DCV) são um grupo de enfermidades que afetam o coração ou os vasos sanguíneos. Há diferentes tipos desta patologia, um exemplo é a doença cardíaca coronária que se refere aos problemas causados pelas artérias coronárias estreitadas, este é o tipo mais comum de enfermidade, tendo como primeiro sinal o ataque cardíaco (World Heart Federation, 2025).

Alguns dos sintomas das possíveis patologias cardíacas podem ser dor, aperto, pressão ou desconforto no peito, fraqueza ou dormência nas pernas e/ou braços, falta de ar, mudanças do ritmo cardíaco, dentre outros. Porém os sintomas podem variar de acordo com a doença (World Heart Federation, 2025).

O diagnóstico das DCV se dá através de uma série de exames feitos pelo médico responsável que também analisa os sintomas e checa o histórico do paciente. Alguns exemplos de exames que podem ser feitos para a detecção da

doença são teste de esforço, eletrocardiograma (ECG/EKG) e ecocardiograma (echo) (World Heart Federation, 2025).

Fatores como histórico familiar não podem ser mudados, porém aspectos de risco como pressão alta podem ser tratados com medicamentos, e os hábitos de vida do paciente podem ser melhorados, evitando dietas não saudáveis, fazendo atividades físicas e evitando o uso de tabaco e o uso nocivo do álcool para assim prevenir as doenças cardiovasculares (World Heart Federation, 2025).

“As doenças cardiovasculares são a principal causa de morte no mundo, mais pessoas morrem anualmente por essas enfermidades do que por qualquer outra causa” (OPAS, Organização Pan-americana da Saúde, 2025).

“Estima-se que 17,9 milhões de pessoas morreram por causa delas em 2019, representando 32 por cento de todas as mortes globais. Dessas mortes, 85 por cento foram devido a ataque cardíaco e derrame” (United Nations, 2025).

Devido ao grande impacto das DCVs na sociedade, torna-se de suma importância a realização de estudos correlacionados à área.

2.2 Análise de dados no contexto da saúde

Análise de dados é a ação de examinar minuciosamente dados brutos com a finalidade de obter conclusões sobre essas informações, tendo como objetivo converter dados que possam estar desorganizados para um formato limpo e de clara compreensão. O processo de análise de dados começa com perguntas a serem respondidas, através delas o analista consegue ter uma ideia clara do problema e saber qual direção necessita seguir (Bhatia, 2017, p. 166).

Na área da saúde a análise de dados também é utilizada, permitindo que certos pontos sejam beneficiados, como os exemplos a seguir.

Com os dados da área da saúde sendo analisados é possível que os cuidados preventivos sejam evidenciados em relação aos cuidados reativos. A tendência mostra que as pessoas tendem a buscar tratamentos depois que já estão com algum problema, levando-as aos procedimentos reativos, que são mais caros. Com o auxílio da análise de dados, os cuidados preventivos podem ser implementados provendo aos pacientes tratamentos anteriores às possíveis doenças e redução de custos (Raheja; Dubey; Chawda, 2017, p. 179).

Com a abordagem do estudo de dados na saúde, é factível que o tratamento seja baseado em evidências. Usando informações relevantes do histórico do paciente juntamente com o histórico médico é possível que seja avaliado os riscos de problemas pós-operatórios, mau funcionamento físico, reação a medicamentos e alergias. A princípio, estas questões geram custos financeiros para as instituições médicas, com a ajuda da análise de dados esses casos podem ser minimizados para diminuir os custos e aumentar a satisfação dos pacientes (Raheja; Dubey; Chawda, 2017, p. 179).

Mais um dos benefícios do processamento de dados pode ser compreendido como a personalização dos atendimentos. Levando em conta que frequentemente os pacientes têm problemas com a troca de médicos e ficam atados a refazer os mesmos exames e testes, perguntas e procedimentos, desperdiçando assim tempo, dinheiro e esforço, não gerando melhora ao paciente. Por meio da

avaliação de dados é admissível que os médicos gerem uma visão completa dos seus pacientes e mantenham isso (Raheja; Dubey; Chawda, 2017, p. 179).

Um exemplo de como a análise de dados foi utilizada se deu quando a *Google*, em uma tentativa de prever surtos de influenza, utilizou de pesquisas por termos relacionados à doença que serviriam como um padrão oculto, pois se as incidências de pesquisa por alergias aumentam durante a estação de alergias, as pesquisas sobre gripe aumentariam nas épocas de gripe. Este projeto ficou conhecido como Google Flu Trends (Tendências de Gripe Google) e ele teve um sucesso inicial em prever surtos de influenza com aproximadamente 97% de acurácia em relação ao Centro de Controle de Doenças (CDC) dos Estados Unidos em 2008, porém relatou falhas nas predições nos anos subsequentes. As palavras-chave das pesquisas não foram divulgadas, porém, alguns dos sintomas da gripe que são dores de cabeça, febre, tosse, vômito, coriza, dor de garganta, cansaço e dores nas articulações, podem ter sido pesquisadas por pessoas que não necessariamente estavam com influenza. Os dados do CDC mostraram que a maioria das pessoas que consultou os médicos cogitando terem gripe na verdade não tinham. Muitas buscas do *Google* não foram excluídas do modelo preditivo embora as pessoas não estivessem gripadas (Bari; Chaouchi; Jung, 2019).

2.3 Ferramentas e Técnicas para Análise de Dados

Com a ascensão dos dados e a sua crescente demanda, muitas ferramentas e técnicas foram criadas e desenvolvidas para facilitar o manuseio e o seu tratamento. Seguem algumas das ferramentas mais utilizadas na análise de dados.

O Microsoft Excel, oferece funções de planilha que são capazes de gerenciar e organizar grandes conjuntos de dados, incluindo também ferramentas gráficas e recursos de automatização de tarefas (Coursera, 2024).

O Tableau é um *software* de visualização de dados capaz de transformar conjuntos de dados em gráficos adicionando ferramentas aptas a executar funções analíticas avançadas, como exemplo a segmentação, análise de corte e análise preditiva (Coursera, 2024).

O SAS (*Statistical Analysis System*) trata-se de um conjunto de *softwares* de análise estatística utilizado para recuperar, relatar, analisar e visualizar dados. Ele combina uma variedade de ferramentas analíticas em um só lugar (Coursera, 2024).

O *Jupyter Notebook* é um ambiente interativo baseado na *Web* para compartilhar documentos computacionais, utilizado para escrever códigos, limpar e visualizar dados, aprendizado de máquina, e muitas outras coisas. Permitindo também que os usuários combinem as visualizações dos dados, código, dentre outros (Coursera, 2024).

Há duas linguagens mais comuns utilizadas na área da análise de dados, sendo elas o Python e o R.

O Python pode ser utilizado para simplificar, modelar, visualizar e analisar dados, muito popular devido aos seus recursos, às bibliotecas, como Pandas e Numpy, que oferecem uma gama de ferramentas para a análise.

E o R que é uma linguagem de programação de código aberto muito utilizado na análise estatística, visualização e manipulação de dados. Ele possui um foco na estatística, tornando-o adequado para cálculos (Coursera, 2024).

2.3.1 Análise de Dados

A análise de dados são procedimentos que exploram e analisam grandes volumes de dados na busca de padrões, erros e previsões. Ela está associada ao aprendizado de máquina, que é uma área da inteligência artificial voltada para o desenvolvimento de algoritmos, passível de aprender a partir do passado, usando dados de eventos que já ocorreram (Amaral, 2016).

A análise de dados é um elemento fundamental do KDD (*Knowledge Discovery in Databases*) – descoberta de conhecimento em banco de dados - que consiste no processo de conversão de dados brutos em informações úteis, das quais segue uma série de passos até o pós-processamento dos resultados da análise. Os passos são o pré-processamento de dados que englobam a seleção de recursos, redução de dimensionalidade e normalização, posteriormente vem a própria análise de dados e em seguida o pós-processamento que abrangem os padrões de filtragem, a visualização e a interpretação de padrões. A finalidade do pré-processamento é converter os dados de entrada que estão brutos passando pelos passos de limpeza, seleção de registros, dentre outros. Devido ao fato das muitas formas de coletar os dados, como por exemplo arquivos simples, planilhas ou tabelas relacionais, o pré-processamento dos dados talvez seja o passo mais demorado de todo o processo (Tan; Steinbach; Kumar, 2009).

Pode-se citar alguns desafios que motivaram o desenvolvimento da análise de dados:

escalabilidade que se refere aos algoritmos de análise de dados que lidam com conjuntos que estão cada vez maiores, podendo chegar a *petabytes*; alta *dimensionalidade* diz respeito a conjuntos que possuem centenas ou até milhares de atributos; *dados complexos* e *heterogêneos* relata-se a necessidade de técnicas de lidar com atributos heterogêneos e complexos (Tan; Steinbach; Kumar, 2009).

2.3.2 Técnicas Estatísticas Descritivas

“A Estatística é considerada um ramo da Matemática, que tem como um dos principais objetivos obter e analisar dados, determinar as correlações entre eles, proporcionando conclusões e previsões” (Morais, 2005).

“A estatística descritiva pode ser considerada como um conjunto de técnicas analíticas utilizadas para resumir o conjunto dos dados recolhidos numa dada investigação [...]” (Morais, 2005).

Temos como resultado os dados que devem ser interpretados para se produzir um resumo das informações. Os dados podem ser de dois tipos: qualitativos, que identificam alguma qualidade, categoria ou característica não apto a medir, mas classificável, que também são divididos em outras duas categorias, nominais sendo apenas uma descrição, tal como masculino e feminino, e ordinal que possuem uma ordem, como por exemplo baixo, médio e alto. Há também os dados quantitativos que representam números, informações e dados medíveis, tendo igualmente duas

classificações, discretos que são números redondos e inteiros, como por exemplo o número de irmãos de uma criança, e contínuos sendo números que podem ser quebrados como é o caso dos valores de uma área ou peso (Morais, 2005).

Algumas técnicas da estatística descritiva podem ser consideradas, como a amostragem e as medidas de estatística, que se seguem.

A utilização da amostragem muitas vezes é utilizada, pois em muitos casos não é possível realizar a pesquisa referente na população como um todo, quer seja por ela ser muito grande ou por falta de recursos, por isso, uma parte da população é selecionada e observada, a chamada amostra. Presume-se que quanto maior for a dimensão da amostra, melhores serão os resultados obtidos (Santos, 2018).

As medidas da estatística descritiva permitem condensar os dados da população ou amostra através de um só valor. Tem-se as medidas de tendência central que são ferramentas importantes para fornecer uma visão inicial ou um resumo da distribuição dos dados nas análises. Podemos verificar a média aritmética que é a soma de todos os valores dividido pelo total de valores observados, a moda que se classifica como o valor que mais se repete e a mediana, que ao se colocar os valores em ordem crescente ele deve estar no meio, em caso de uma quantidade ímpar, soma-se os dois valores do meio e divide-se por dois (Morais, 2005).

Estes são alguns dos tópicos presentes dentro da estatística descritiva.

2.4 Trabalhos Relacionados

Com a avaliação de outros trabalhos relacionados ao tema é possível verificar os estudos que já existem, e entender até onde já foi realizado, podendo assim identificar lacunas e/ou oportunidades para novos possíveis trabalhos.

Foram encontrados dois artigos que se referem ao mesmo tema deste trabalho, dos quais o primeiro tem como objetivo realizar a previsão de doenças cardiovasculares por algoritmos de aprendizagem de máquina, e verificar qual dentre os utilizados apresenta melhor capacidade de generalização, tanto interna quanto externamente (Marques et al., 2024). O segundo possui como objetivo apresentar um estudo detalhado sobre as diferentes técnicas de análise de dados que podem ser implantadas em sistemas automatizados (Bhatla ; Jyoti, 2012).

O primeiro artigo teve como metodologia treinar e validar os modelos em uma corte retrospectiva – um conjunto de dados históricos, coletados de forma sistemática, onde as condições de interesse, como as doenças, já ocorreram antes do início do estudo - utilizando dados da Faculdade de Medicina de Ribeirão Preto, como validação interna e para a validação externa utilizou-se os dados do Beth Israel Deaconess Medical Center (BIDMC), EUA através de oito algoritmos de aprendizados de máquina (Marques et al., 2024). O artigo que vem em seguida tem como principal metodologia examinar as publicações, periódicos e revisões no ramo da ciência da computação e engenharia, mineração de dados e doenças cardiovasculares (Bhatla ; Jyoti, 2012).

Analisando os resultados obtidos de ambos os artigos, o primeiro concluiu que apenas um dos oito algoritmos de aprendizado apresentou o melhor desempenho preditivo em validação interna e externa, possuindo também a melhor capacidade de generalização (Marques et al., 2024). O segundo chegou à conclusão

de que a rede neural – modelo de aprendizagem de máquina inspirado no funcionamento do cérebro humano, composta por nós que processam as informações e aprendem padrões a partir dos dados – apresentou maior precisão quando avaliada com 15 atributos, não excluindo o fato de que a Árvore de Decisão – outro modelo de aprendizagem – também teve um bom desempenho (Bhatla ; Jyoti, 2012).

3 Metodologia

Este tópico explica quais são as ferramentas que foram utilizadas no desenvolvimento do artigo, a descrição do *dataset* escolhido para este artigo e as etapas que serão seguidas durante o processo de análise dos dados.

3.1 Descrição do Dataset

O banco de dados Heart Disease contém 76 atributos, porém o *dataset* que será avaliado refere-se a um subconjunto de 14 deles contendo 303 instâncias de característica multivalorada, sendo os tipos de recurso categórico, inteiro e real (UCI Irvine Machine Learning Repository, 1988).

3.1.1 Origem dos dados

A base de dados tem como área de assunto saúde e medicina, os dados são provenientes de 4 bases de dados, sendo Cleveland, Hungria, Suíça e o VA Long Beach, sendo que o banco de dados de Cleveland é o único que foi usado por pesquisadores de ML (*Machine Learning*) até então, de acordo com as informações e os experimentos com o banco de dados de Cleveland se concentram em unicamente tentar distinguir a presença (valores 1, 2, 3, 4) da ausência (valor 0). Os nomes e números de segurança social dos pacientes, cuja informação está no banco de dados foram removidos e substituídos por valores fictícios (UCI Irvine Machine Learning Repository, 1988).

3.1.2 Dicionário de Dados

Os atributos do conjunto de dados analisado são: **age** (idade em anos), **sex** (sexo, 1 = Masculino; 0 = Feminino), **cp** (tipo de dor no peito (angina), valor 1: angina típica; valor 2: angina atípica; valor 3: dor não anginosa; valor 4: assintomático), **trestbps** (pressão arterial em repouso (em mm Hg na admissão ao hospital)), **chol** (colesterol sérico em mg/dl), **fbs** (açúcar no sangue em jejum > 120 mg/dl, 1 = verdadeiro; 0 = falso), **restecg** (resultados eletrocardiográficos em repouso, valor 0: normal; valor 1: com anormalidade da onda ST-T (inversões de onda T e/ou do ST ou depressão de 0,05 mV); valor 2: mostrando hipertrofia ventricular esquerda provável ou definitiva pelos critérios de 'Estes'), **thalach** (frequência cardíaca máxima alcançada), **enxang** (angina induzida pelo exercício, 1 = sim; 0 = não), **oldpeak** (depressão ST induzida pelo exercício em relação ao repouso), **slope** (a inclinação do segmento ST no pico do exercícios, valor 1: inclinação ascendente; valor 2: plano (sem inclinação); valor 3: inclinação descendente), **ca** (número de vasos principais (0-3) visualizados por fluoroscopia), **thal** (codificação usada para descrever resultados de exames relacionados ao fluxo de sangue no músculo cardíaco, 3 = normal; 6 = defeito fixo; 7 = defeito reversível) e **num** (diagnóstico de doença cardíaca (status da

doença obtido por angiografia), valor 0 = menos de 50% de estreitamento do diâmetro; valor 1 = mais de 50% de estreitamento do diâmetro) (UCI Irvine Machine Learning Repository, 1988).

3.2 Ferramentas utilizadas para a análise

Para a análise de dados ser realizada utiliza-se de algumas ferramentas. Estas serão apresentadas a seguir e utilizadas na realização deste trabalho.

3.2.1 Python e suas Bibliotecas

A linguagem de programação Python foi criada no início dos anos 90, por Guido van Rossum, sendo uma das mais utilizadas até os dias atuais, de acordo com o *ranking* realizado pelo instituto de engenheiros elétricos e eletrônicos (IEEE) em 30/08/2022 (Lima, 2022), além do índice TIOBE (uma empresa especializada em monitorar e avaliar a qualidade de *softwares*) (TIOBE, 2025) também ter revelado em 2025 que Python é a linguagem mais utilizada (Pilotto, 2025). Sendo considerada uma linguagem de alto nível, por ter uma sintaxe de simples e fácil compreensão que é amplamente utilizada em diversas áreas, como desenvolvimento *web*, ciência de dados e inteligência artificial (Python, 2025).

Esta linguagem possui diversas bibliotecas disponíveis para uso com diversas funções, aqui serão citadas algumas das existentes que também serão utilizadas neste trabalho.

A biblioteca Pandas, de acordo com a sua documentação, “[...] é uma ferramenta de análise e manipulação de dados de código aberto rápido, poderoso, flexível e fácil de usar, construído em cima da linguagem de programação Python” (PANDAS, 2025) que foi construída em 2008 permitindo a manipulação de dados numéricos de alto desempenho.

O Numpy é uma biblioteca voltada para operações matemáticas, *arrays* e matrizes. Sendo rápido e versátil nos conceitos de vetorização, oferecendo funções matemáticas abrangentes, tendo uma fonte aberta e possuindo sintaxe de alto nível, tornando-o acessível e produtivo (Numpy, 2025).

A Matplotlib “é uma biblioteca abrangente para criar estática, animada, visualizações interativas em Python”. Ela é voltada para a visualização de dados em formato de gráficos estatísticos, podendo criar figuras interativas e personalizar os gráficos com cores, legendas e títulos (Matplotlib, 2025).

Finalizando com a biblioteca Seaborn que de acordo com a sua biblioteca também é voltada para a visualização de dados Python que é baseada na Matplotlib, se integrando intimamente com as estruturas de dados do Pandas, fornecendo interface de alto nível para desenhos gráficos estatísticos (Seaborn, 2025).

3.2.2 Ambientes Computacionais

Os ambientes computacionais são espaços ou sistemas que podem ser físicos ou virtuais que foram criados para o desenvolvimento, teste, execução e para manter programas ou sistemas. Eles podem ser utilizados para desenvolver

softwares, criando e programando sistemas e/ou aplicativos, servem também para testar os códigos, garantindo que os erros sejam evitados e o sistema funcione corretamente, executando programas, dentre outros. Temos como exemplo o Visual Studio e o PyCharm que são ambientes de desenvolvimento e servem para a programação e desenvolvimento dos sistemas, há os ambientes de produção como os servidores em nuvem AWS e Azure onde o sistema real do usuário é executado para o usuário final, e dentre outros exemplos há o *Jupyter Notebook* que será o utilizado neste trabalho.

O *Jupyter Notebook* é voltado para ciência de dados, computação científica, jornalismo computacional, aprendizado de máquina, análise estatística, educação e programação. É um ambiente recente, desenvolvido originalmente como parte do projeto IPython, é de código aberto e pode suportar mais de 40 linguagens de programação como Python e R, partilhar os blocos de trabalho e gerar saída interativa em HTML, imagens, vídeos etc. Sendo um espaço de desenvolvimento interativo que suporta visualização de gráficos e tabelas, considerado um ambiente de desenvolvimento e análise por reunir as ferramentas que escrevem, testam, visualizam e documentam o código em um só lugar (Jupyter, 2025).

3.3 Etapas da Análise de Dados

Podemos classificar a análise de dados a partir de quatro tipos, a descritiva que descreve o que aconteceu usando técnicas de análise é possível gerar estatísticas como média de idades, distribuição de gênero, regiões com mais vendas dentre outros, ajudando a empresa a ter uma visão geral das informações necessárias.

Análise diagnóstica, relata o por que aconteceu, ao realizar esta análise a empresa descobre a causa raiz de um problema e toma as medidas corretivas, como exemplo a melhoria na infraestrutura de rede.

A análise preditiva, informa o que é provável de acontecer, esta forma utiliza técnicas como modelos de regressão ou algoritmos de aprendizado de máquina, podendo prever a probabilidade de algo específico acontecer, ou algum cliente fazer algo no futuro, permitindo que a empresa faça ajustes alocando recursos da melhor maneira possível.

Análise prescritiva, auxilia em qual decisão tomar, pode-se utilizar algoritmos de otimização para recomendar ações a serem tomadas levando em consideração fatores do contexto.

Os processos de análise de dados podem ser elencados em pré-processamento, limpeza e tratamento de dados faltantes e análise exploratória dos dados, os quais serão vistos a seguir.

3.3.1 Pré-processamento dos Dados

Nesta etapa, o pré-processamento dos dados, é onde os dados são organizados para assim poderem ir para as próximas etapas, aqui está presente a limpeza e o tratamento dos dados, para que os dados, ao serem analisados estejam claros, limpos, corretos e organizados, pois os dados chegam crus após serem coletados e precisam passar por esse pré-processamento para serem de fato

analisados e gerarem dados concisos e corretos, sem discrepâncias e erros (Grus, 2021)

3.3.2 Limpeza e Tratamento de Dados Faltantes

A limpeza e tratamento dos dados é para garantir que eles estejam no formato adequado para a análise ou modelagem, reduzindo assim as chances de erro.

Neste processo há algumas etapas que são executadas para que a limpeza seja efetivada de forma correta e eficiente, como a remoção dos dados ausentes, a substituição de dados ausentes/inconsistentes, a aplicação de transformações em dados numéricos, a conversão dos dados entre diferentes tipos de dados e a normalização dos mesmos (Grus, 2021).

Esta etapa tem como objetivo tornar os dados mais confiáveis e consistentes, possibilitando seu uso com maior qualidade e reduzindo a ocorrência de erros, conforme dito anteriormente (Grus, 2021).

3.3.3 Análise Exploratória de Dados

A análise exploratória está associada à exploração dos dados em busca de informações e *insights* cujo foco é visualizar e entender padrões, tendências e anomalias nos dados (Grus, 2021).

Muito comumente são utilizados as bibliotecas, as quais foram descritas anteriormente aqui, para investigar, visualizar e conhecer os dados (Grus, 2021).

A análise e a preparação de dados têm como propósito aprofundar a compreensão sobre bases de dados, suas variáveis e comportamentos. Aliadas à mineração de dados e à aprendizagem de máquina, essas etapas formam a base técnica do trabalho dos cientistas de dados. (Ferreira et al., 2021)

Nos dias atuais, o acesso aos mais diversos e variados tipos e formatos de dados tem aumentado, e para um projeto bem sucedido nesta área, é necessário tratar os dados brutos até que eles se tornem informações, o uso correto destas informações é transformado em conhecimento e ao utilizarmos o conhecimento a favor de uma tomada de decisão é gerada a sabedoria, sendo essas as etapas da transformação dos dados. A análise exploratória de dados tem como finalidade checar os dados para qualquer aplicação estatística, permitindo entender mais dos dados coletados e principalmente as existentes entre as variáveis analisadas. (Ferreira et al., 2021)

4 Desenvolvimento

Esta seção tem como objetivo informar o processo de desenvolvimento deste trabalho, através da análise descritiva, de estatísticas básicas das variáveis, das relações entre elas, da visualização dos dados através de gráficos e mapas, dentre outras coisas.

O *dataset* utilizado foi extraído do site “*UC Irvine Machine Learning Repository*” que refere-se à presença de doenças cardíacas nos pacientes, tendo sido publicados em 30/06/1988. Possui 303 instâncias (linhas) e 14 características (colunas), das quais estão especificadas no tópico 3.1.2

4.1 Análise Descritiva

Nesta seção serão identificadas algumas informações sobre os dados.

A Figura 1 retrata o começo do *dataset*, trazendo os primeiros 5 registros em linhas e colunas.

Figura 1 : Df.head

```
# Ver as 5 primeiras linhas
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0

Fonte: autoria própria

A Figura 2 descreve exatamente as linhas que têm ao menos um valor ausente, NaN, os quais, posteriormente, serão tratados e preenchidos.

Figura 2: Valor NaN

```
# Mostra todas as linhas que têm ao menos 1 valor ausente
df[df.isnull().any(axis=1)]
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
87	53	0	3	128	216	0	2	115	0	0.0	1	0.0	NaN	0
166	52	1	3	138	223	0	0	169	0	0.0	1	NaN	3.0	0
192	43	1	4	132	247	1	2	143	1	0.1	2	NaN	7.0	1
266	52	1	4	128	204	1	0	156	1	1.0	2	0.0	NaN	2
287	58	1	2	125	220	0	0	144	0	0.4	2	NaN	7.0	0
302	38	1	3	138	175	0	0	173	0	0.0	1	NaN	3.0	0

Fonte: autoria própria

Percebe-se que na coluna **ca** há quatro valores nulos, a solução encontrada para esta questão foi preencher estes valores com a mediana (o valor que fica no meio quando ordenado), pois é a mais consistente em relação a *outliers* (dados que estão muito distantes do padrão geral) fazendo sentido com este tipo de variável, numérica contínua com valores de 0-3.

Na coluna **thal** encontram-se 2 valores faltantes, a moda (valor que mais se repete) foi utilizado como solução, garantindo assim o preenchimento com o padrão mais comum no *dataset* utilizado, pois esta é uma variável categórica (variáveis que representam categorias ou grupos distintos, e não expressam quantidades numéricas mensuráveis. Elas indicam qualidade e não quantidade.), mantendo assim a coerência da variável.

4.1.1 Estatísticas Básicas das Variáveis

A Figura 3 e 4 ilustram as estatísticas básicas dos dados, contendo informações das quantidades de dados (count), das médias (mean), do desvio padrão (std), do mínimo (min) e máximo (max) e dos quartis, 25%, 50% e 75%. A Figura 3 será dividida em duas partes devido a sua dimensão, porém as informações provêm do mesmo lugar.

Figura 3: Descrição parte 1

```
# Resumo estatístico nas colunas numéricas
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000

Fonte: autoria própria

A Figura 5 retrata a contagem (count), a média (mean), o desvio padrão (std), o mínimo (min), os 25%, 50%, 75% e o máximo (max) respectivamente da nova coluna que foi adicionada ao *dataset* (**NovoNum**). A interpretação feita para chegar a conclusão da sua criação foi que no dicionário de dados a coluna original (num) deveria conter apenas números de 0 e 1, no entanto continha números de 0 a 4, o que não foi explicado pelo dicionário, por isso, a conclusão tomada foi de que os números maiores que 1 seriam convertidos para 1, pois isso significa que há uma condição existente de doença cardíaca.

Figura 4: Descrição parte 2

oldpeak	slope	ca	thal	num
303.000000	303.000000	303.000000	303.000000	303.000000
1.039604	1.600660	0.663366	4.722772	0.937294
1.161075	0.616226	0.934375	1.938383	1.228536
0.000000	1.000000	0.000000	3.000000	0.000000
0.000000	1.000000	0.000000	3.000000	0.000000
0.800000	2.000000	0.000000	3.000000	0.000000
1.600000	2.000000	1.000000	7.000000	2.000000
6.200000	3.000000	3.000000	7.000000	4.000000

Fonte: autoria própria

Figura 5: Nova coluna adicionada no *dataset*

NovoNum
303.000000
0.458746
0.499120
0.000000
0.000000
0.000000
1.000000
1.000000

Fonte: autoria própria

4.1.2 Relações entre Variáveis

As variáveis podem ser relacionadas entre si pela sua relevância como um indicador da doença arterial coronariana. A seguir é relatado a variável, a sua relevância e o motivo da sua relevância.

A variável **cp** (tipo de dor no peito) é considerada de alta relevância pois o tipo de angina é um indicador direto de isquemia (ocorre quando as artérias coronárias estão obstruídas, impedindo que o sangue chegue adequadamente ao músculo do coração (miocárdio)), há também a variável **thalach** (frequência cardíaca máxima) que por sua vez também possui relevância alta pois a frequência limitada durante o esforço pode indicar uma **dac** (doença arterial coronariana), o **exang**

(angina induzida pelo exercício) é de relevância alta porque está diretamente relacionada à presença de isquemia, explicada anteriormente, o **oldpeak** (depressão **st**) é considerada de alta relevância também devido às alterações do segmento ST serem sinais de esforço cardíaco anormal, o **ca** (vasos principais) são considerados de relevância muito alta pois ele visualiza diretamente as obstruções, o **thal** (fluxo de sangue no músculo cardíaco) de relevância alta devido a se referir ao exame nuclear do coração que identifica anomalias no fluxo de sangue, **chol**, **trestbps**, **fbs** (colesterol sérico, pressão arterial em repouso, açúcar no sangue em jejum, respectivamente) podem ser consideradas de relevância moderada, pois são fatores de risco mas não são diagnósticos diretos, o **restecg** (resultados eletrocardiográficos em repouso) é de relevância moderada pois detecta anormalidades elétricas no coração, e por fim **age** e **sex** (idade e sexo, respectivamente) também são considerados de relevância moderada pois são fatores demográficos que são importantes de avaliar em risco.

4.2 Visualização de Dados

A etapa de visualização de dados é essencial para a etapa de análise exploratória em análise de dados, permitindo interpretar de forma clara e visual o comportamento das variáveis presentes no conjunto de dados. Por meio de gráficos e representações visuais, é possível identificar padrões, relações entre variáveis, possíveis *outliers* e até mesmo inconsistências nos dados.

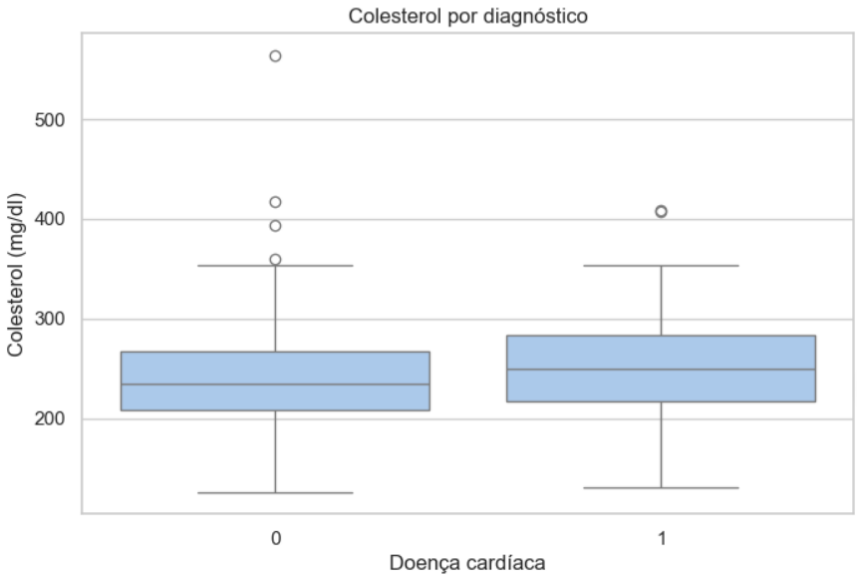
4.2.1 Gráficos e Mapas de Correlação e Insights sobre o Dataset

As Figuras de 6 a 16 são os gráficos e mapas de correlação das variáveis que posteriormente serão analisados e gerados *insights* sobre. Em cada figura há o bloco com o comando dado para a geração da imagem, juntamente com um comentário que explica sucintamente o que o gráfico ou mapa faz.

A figura 6 se refere a um gráfico de caixa (*boxplot*) que compara os níveis de colesterol (**chol**) entre os diferentes resultados de diagnóstico (**NovoNum**) verificando se os pacientes com doença têm colesterol mais alto que os saudáveis. Percebe-se que há alguns valores acima do valor limite estipulado pelo gráfico, que seriam acima de 360. Durante a análise exploratória da variável colesterol (**chol**) foram identificados alguns valores elevados, sendo o maior deles de 564, consideravelmente acima da média observada no conjunto de dados. Este valor é considerado um *outlier*, estatisticamente, no entanto optou-se por manter esses dados na base, uma vez que clinicamente, níveis tão elevados de colesterol podem estar associados a casos raros, porém possíveis de hipercolesterolemia familiar, especialmente em sua forma homozigótica, no qual os níveis podem ultrapassar 500 mg/dL. Portanto, a presença deste valor no *dataset* foi avaliada como uma possível representação de condição médica específica.

Figura 6: Colesterol x Diagnóstico

```
# 3. Boxplot colesterol x diagnóstico
plt.figure(figsize=(8,5))
sns.boxplot(x='NovoNum', y='chol', data=df)
plt.title("Colesterol por diagnóstico")
plt.xlabel("Doença cardíaca")
plt.ylabel("Colesterol (mg/dl)")
plt.show()
```



Fonte: autoria própria

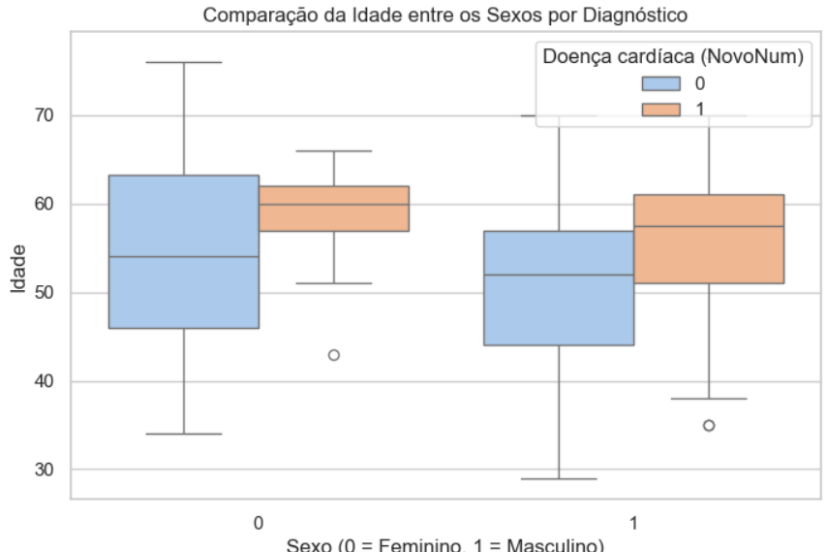
Alguns dos *insights* analisados a partir da visualização dos dados são, o colesterol deve ser avaliado juntamente com um conjunto de outros fatores, ele sozinho não é um bom preditor de DAC, de acordo com a figura 6, percebe-se que o intervalo dos valores de colesterol das pessoas que possuem a DAC é apenas um pouco mais elevado que o das pessoas que não possuem a doença.

A Figura 7 retrata a comparação entre a idade dos diferentes sexos dos pacientes, de acordo com o resultado, sendo útil para verificar a distribuição da idade entre homens e mulheres e a comparação da idade entre quem tem e quem não tem **dac** (doença arterial coronariana), dentro de cada sexo.

Através da comparação da idade entre os sexos por diagnóstico é possível perceber que a quantidade de homens que tem a DAC é maior que a de mulheres, o intervalo também é maior, o que indica há maior predisposição para homens na faixa de 50 a 60 anos de idade terem a doença, enquanto para mulheres esse intervalo é menor e a idade é maior, entre 57 a 62, o que indica que as mulheres têm mais predisposição a terem a doença quando mais velhas, principalmente depois da menopausa, devido à redução da produção dos hormônios, dado o qual pode ser confirmado através do trabalho de iniciação científica referenciado a seguir. (Vargas et al., 2019)

Figura 7: Idade e sexo por Diagnóstico

```
plt.figure(figsize=(8,5))
sns.boxplot(x='sex', y='age', hue='NovoNum', data=df)
plt.title("Comparação da Idade entre os Sexos por Diagnóstico")
plt.xlabel("Sexo (0 = Feminino, 1 = Masculino)")
plt.ylabel("Idade")
plt.legend(title="Doença cardíaca (NovoNum)", loc="upper right")
plt.show()
```

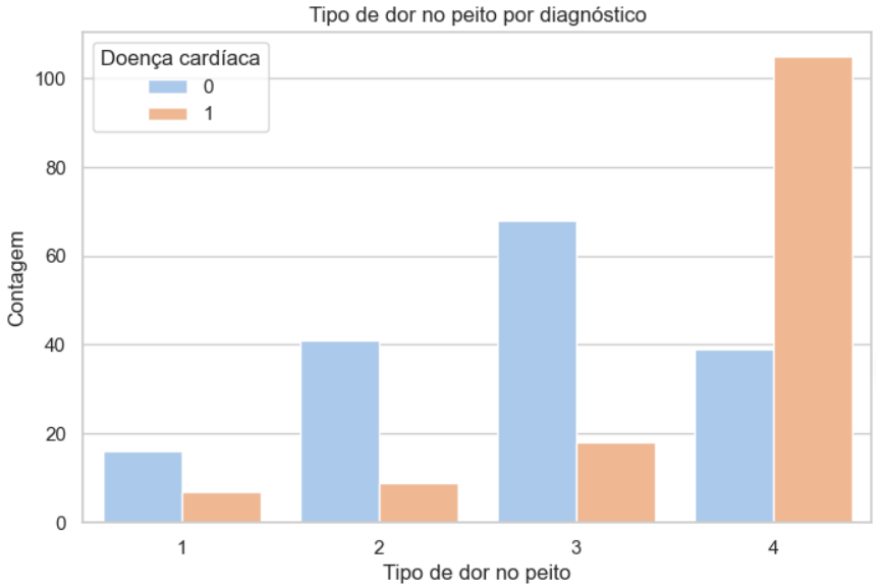


Fonte: autoria própria

A Figura 8 se refere a comparação do tipo de dor no peito por diagnóstico, revelando, através de um gráfico de barras, a quantidade de pessoas que possuem cada tipo de dor.

Figura 8: Tipo de angina

```
# 1. Frequência de tipos de dor no peito por diagnóstico
plt.figure(figsize=(8,5))
sns.countplot(x='cp', hue='NovoNum', data=df)
plt.title("Tipo de dor no peito por diagnóstico")
plt.xlabel("Tipo de dor no peito")
plt.ylabel("Contagem")
plt.legend(title='Doença cardíaca')
plt.show()
```



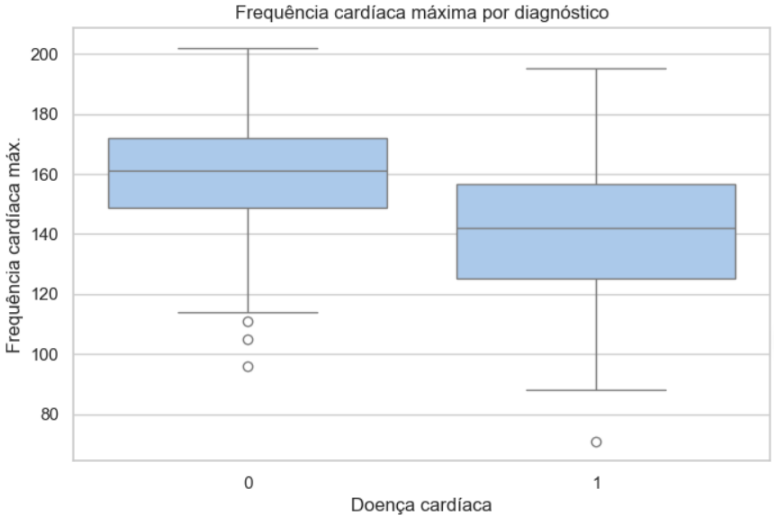
Fonte: autoria própria

Pelo gráfico de tipo de dor no peito por diagnóstico, Figura 8, pode-se avaliar que aqueles que apresentam dor no peito assintomático (**cp = 4**) aparecem com muito mais frequência entre os pacientes com doença cardíaca, enquanto aqueles que apresentam angina típica, ou atípica (**cp =1** ou **2**) tem menor associação com a DAC, indicando que pacientes assintomáticos, nesse aspecto, devem ser monitorados com atenção, pois podem esconder uma doença coronária silenciosa, e isso deve ser avaliado levando em consideração todos os aspectos possíveis do paciente.

A Figura 9 apresenta um gráfico *boxplot* que avalia a frequência cardíaca máxima por diagnóstico, o qual é capaz de verificar os valores das frequências cardíacas e seus limites, tais como suas médias, assim como a Figura 10 que representa as mesmas informações, porém em formato de histograma (**histplot**).

Figura 9: FCM por Diagnóstico 1

```
# 4. Frequência cardíaca máxima por diagnóstico
plt.figure(figsize=(8,5))
sns.boxplot(x='NovoNum', y='thalach', data=df)
plt.title("Frequência cardíaca máxima por diagnóstico")
plt.xlabel("Doença cardíaca")
plt.ylabel("Frequência cardíaca máx.")
plt.show()
```



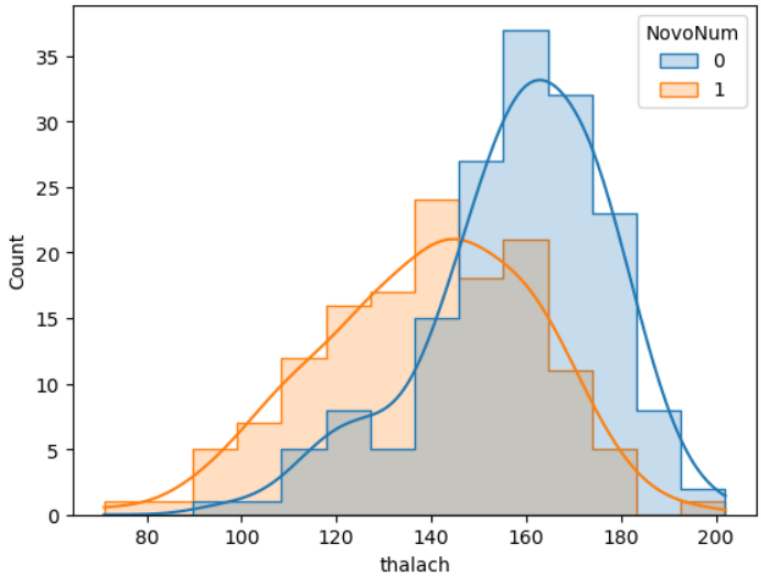
Fonte: autoria própria

Avaliando a frequência cardíaca máxima por diagnóstico nota-se que aqueles que foram diagnosticados com a DAC possuem a frequência cardíaca máxima diminuída, tanto a sua média, se comparada aos que não possuem a doença, quanto o intervalo, o que é um indicador de que aqueles sem a doença conseguem atingir frequências cardíacas mais elevadas durante o esforço, e a dificuldade de atingir tais frequências altas podem indicar limitações cardíacas e isquemia.

Figura 10: FCM por Diagnóstico 2

```
# Permite ver se pacientes com doença
# apresentam distribuições diferentes de frequência cardíaca máxima.
sns.histplot(data=df, x='thalach', hue='NovoNum', kde=True, element='step')
```

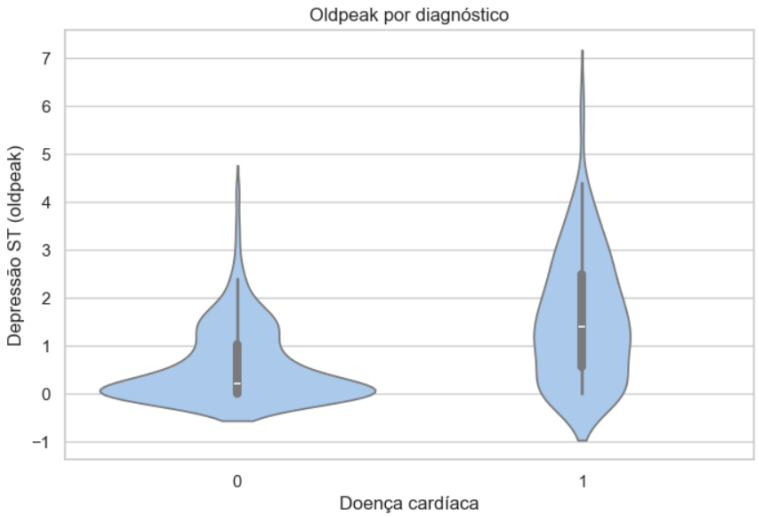
```
<Axes: xlabel='thalach', ylabel='Count'>
```



A Figura 11 representa um gráfico de violino que compara a depressão do segmento **st** (**oldpeak**) com o diagnóstico.

Figura 11: Depressão Seg. ST

```
# 5. oldpeak x diagnóstico
plt.figure(figsize=(8,5))
sns.violinplot(x='NovoNum', y='oldpeak', data=df)
plt.title("Oldpeak por diagnóstico")
plt.xlabel("Doença cardíaca")
plt.ylabel("Depressão ST (oldpeak)")
plt.show()
```



Fonte: autoria própria

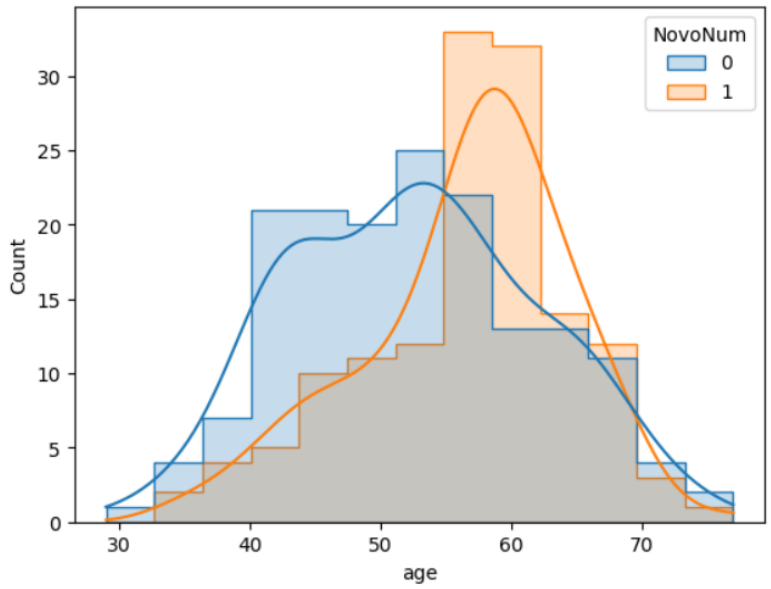
A respeito do gráfico que discorre acerca da depressão do segmento st (**oldpeak**), Figura 11, avalia-se que pacientes sem a DAC geralmente não apresentam depressão no segmento st significativa, pois a maior parte está concentrada em 0, mas para aqueles que possuem a doença, percebe-se a que depressão no segmento st se faz presente na sua maior parte em valores mais baixos, porém nestes que possuem a doença, os indicadores mostram que também há valores elevados.

A Figura 12 simboliza um histograma com a distribuição de idade, separando por cores conforme o resultado (**NovoNum**), checando assim se há uma faixa etária predominante entre os pacientes com e sem problemas cardíacos.

Figura 12: Idade por diagnóstico

```
# Cria um histograma com a distribuição da idade
#separando por cores conforme o resultado cardíaco
sns.histplot(data=df, x='age', hue='NovoNum', kde=True, element='step')
```

```
<Axes: xlabel='age', ylabel='Count'>
```



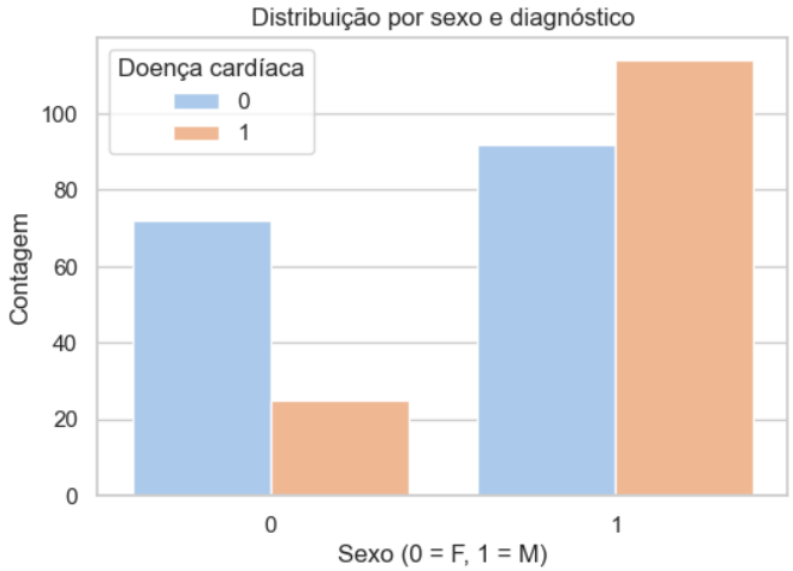
Fonte: autoria própria

Com o histograma de idade por diagnóstico, Figura 12, é possível perceber que há uma faixa etária na qual a existência da doença cardíaca prevalece, que neste caso é de 55 a 65 anos. Juntamente com o gráfico que retrata a distribuição de sexo e diagnóstico, Figura 19, onde através dele pode-se constatar que os homens tem mais pré-disposição a ter as doenças cardíacas do que as mulheres.

A Figura 13 indica um gráfico de barras que mostra a quantidade de pacientes por resultado separados por **sexo**, avaliando assim se a proporção de pacientes com problemas cardíacos é diferente entre homens e mulheres.

Figura 13: Quantidade por sexo e diagnóstico

```
# 6. Comparação por sexo
plt.figure(figsize=(6,4))
sns.countplot(x='sex', hue='NovoNum', data=df)
plt.title("Distribuição por sexo e diagnóstico")
plt.xlabel("Sexo (0 = F, 1 = M)")
plt.ylabel("Contagem")
plt.legend(title='Doença cardíaca')
plt.show()
```

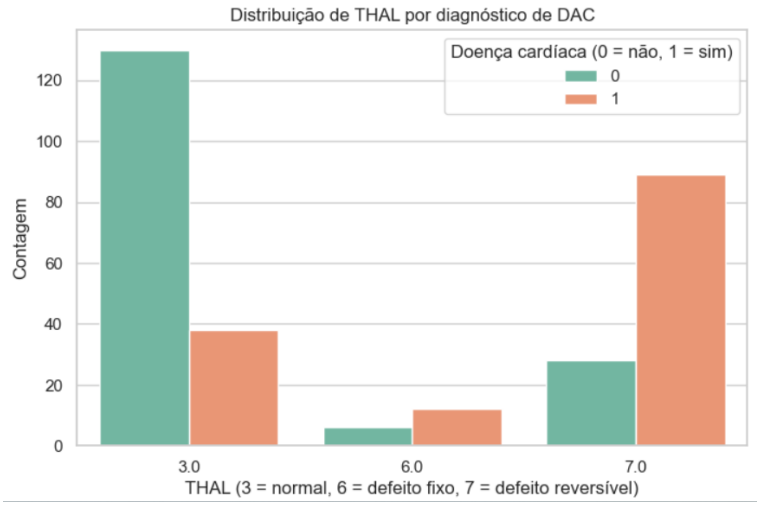


Fonte: autoria própria

A Figura 14 exibe a distribuição de **thal** por diagnóstico de DAC, através de um gráfico de barras, exibindo a quantidade de pacientes que tem cada uma das classificações da variável.

Figura 14: Distribuição do THAL

```
plt.figure(figsize=(8,5))
sns.countplot(x='thal', hue='NovoNum', data=df, palette='Set2')
plt.title("Distribuição de THAL por diagnóstico de DAC")
plt.xlabel("THAL (3 = normal, 6 = defeito fixo, 7 = defeito reversível)")
plt.ylabel("Contagem")
plt.legend(title="Doença cardíaca (0 = não, 1 = sim)")
plt.show()
```

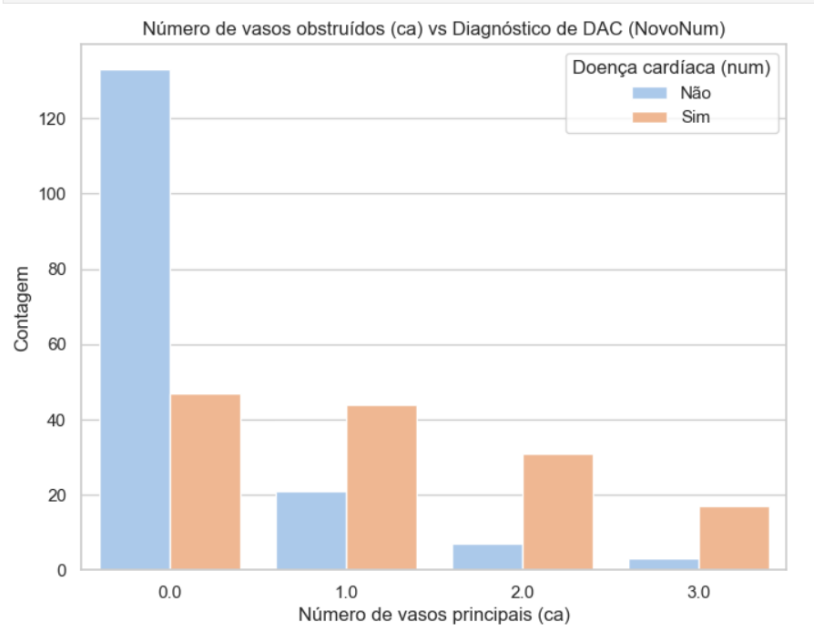


Fonte: autoria própria

O gráfico de distribuição de **thal** por diagnóstico de DAC, Figura 14, avalia-se que os resultados do tipo "defeito reversível" têm uma forte associação com a DAC o que se comprova com o gráfico gerado, que apresenta uma maior quantidade de pacientes que tem esse defeito e possui a DAC, pois os defeitos reversíveis mostram áreas do coração com o fluxo sanguíneo comprometido mas que em alguns casos podem ser reversíveis, como o nome diz, caso a obstrução não tenha sido completa ou se ainda houver fluxo sanguíneo colateral suficiente.

A Figura 15 exibe o número de vasos obstruídos por diagnóstico de DAC, através de um gráfico de barras, exibindo a quantidade de pacientes que tem cada uma das classificações da variável.

Figura 15: Número de vasos obstruídos



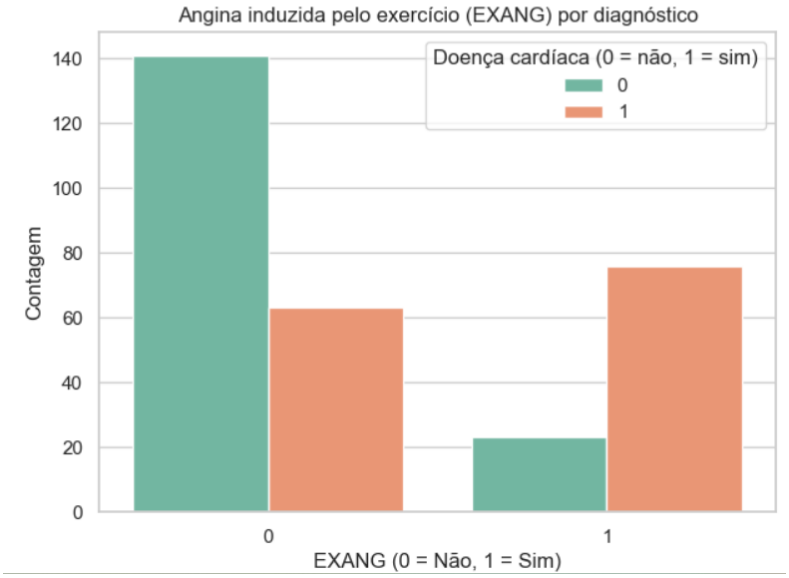
Fonte: autoria própria

Com referência ao gráfico de distribuição de **ca** por diagnóstico de doença cardíaca, Figura 15, revela-se que daqueles que não possuem a doença, também não há obstrução dos vasos, mas para aqueles que possuem a doença, se mostrou maior a quantidade de pessoas com apenas 1 vaso obstruído, quanto ao demais vasos, 2 ou 3, a quantidade decrementa, mas a prevalência ainda é maior daqueles que têm a doença e apresentam a obstrução dos vasos.

A Figura 16 exibe a quantidade de pacientes que relatam a angina induzida pelo exercício, por diagnóstico.

Figura 16: Angina induzida pelo exercício

```
plt.figure(figsize=(7,5))
sns.countplot(x='exang', hue='NovoNum', data=df, palette='Set2')
plt.title("Angina induzida pelo exercício (EXANG) por diagnóstico")
plt.xlabel("EXANG (0 = Não, 1 = Sim)")
plt.ylabel("Contagem")
plt.legend(title="Doença cardíaca (0 = não, 1 = sim)")
plt.show()
```



Fonte: autoria própria

Finalizando com a visualização do gráfico que apresenta a angina induzida pelo exercício por diagnóstico, Figura 16, onde através dele pode-se perceber que a maior parte dos pacientes não apresentam angina e não tem a doença, mas ainda assim há casos que fogem a essa maioria, dos quais precisam ser avaliados com cuidado

4.3 Padrões Identificados

Alguns dos padrões obtidos e identificados através dos *insights* e dos dados e gráficos gerados são apresentados em seguida.

Observou-se que a incidência de doenças cardíacas aumenta com a idade, especialmente em homens a partir dos 50 anos e em mulheres após os 57 anos, o que sugere uma predisposição etária e de gênero.

Pacientes com dor no peito assintomática (**cp=4**) apresentaram uma correlação notavelmente maior com o diagnóstico de doença cardíaca, indicando a importância de monitoramento mesmo na ausência de sintomas típicos.

A frequência cardíaca máxima (**thalach**) mostrou-se inversamente correlacionada com a presença de doença, com pacientes diagnosticados apresentando médias e intervalos de **thalach** significativamente menores.

Mesmo que a depressão do segmento ST mais elevada seja em menor densidade de volume, ela aparece em pacientes com DAC, já a maior densidade no volume aparece em pacientes sem doença marcados no valor 0, indicando que pacientes sem a doença não tenham depressão neste segmento.

Pacientes com resultado de “defeito reversível” apontam uma correspondência consideravelmente com a doença, indicando que há um dano no local, mas que ainda há a chance de ser reversível.

Percebe-se que a maioria daqueles que não possuem a doença cardíaca também não possuem obstrução dos vasos sanguíneos, porém pessoas que não tem vasos obstruídos também podem desenvolver a doença, e a quantidade de vasos obstruídos se reduz, levando a avaliar que há prevalência em obstruções de apenas um vaso, mas pode acometer mais de um.

4.3.1 Variáveis mais Relevantes na Identificação de Doenças Cardíacas

As variáveis mais preditivas da doença cardíaca neste conjunto de dados são aquelas que avaliam diretamente a função e perfusão do coração sobre estresse e/ou visualizam obstruções, como a **ca** (número de vasos obstruídos), que possui uma relevância altíssima, pois mede visualmente a obstrução das artérias, o que mostra uma forte correlação com o diagnóstico. O **oldpeak** (depressão do segmento **st**), representa alta relevância pois ele retrata as alterações no eletrocardiograma durante o esforço, que pode revelar casos de isquemia. **cp** (tipo de dor no peito), também de alta relevância, pois mostra que o tipo de dor (assintomática - **cp=4**) está fortemente relacionada a presença das doenças. **thalach** (frequência cardíaca máxima), alta relevância, é um indicativo da correlação do comprometimento da função cardíaca sob estresse. **exang** (angina induzida pelo exercício), alta relevância, através dos resultados obtidos pela variável, ela se provou eficaz pois a presença de angina durante o esforço está diretamente relacionada à isquemia cardíaca, e neste caso, ela se mostrou um bom indicador clínico, pois como foi mencionado, como o resultado foi a maioria negativo para dor e para a doença, então significa que quem não tem sintomas também não tem a doença, essa correspondência entre ausência de sintoma e diagnóstico fortalece a validade clínica da variável “**exang**” como um bom indicador de risco. **thal** (tipo de anormalidade no exame nuclear do coração), alta relevância, especialmente os casos com defeito reversível (**thal=7**) indicam fluxo sanguíneo comprometido em esforço e normal em repouso, típico de DAC. O gráfico mostra forte associação com o diagnóstico, Figura 20. Já as demais variáveis como colesterol, pressão e idade são importantes como fatores de risco, mas têm menor poder preditivo isolado para diagnosticar DAC com tamanha precisão.

4.4 Limitações da Análise

O tamanho e a origem do *dataset* contam como uma das limitações encontradas no decorrer deste trabalho, pois *o dataset* utilizado, embora robusto para fins de pesquisa, é proveniente de uma única fonte (Cleveland) e pode não ser totalmente representativo da diversidade populacional global, limitando a generalização dos resultados.

O número de atributos do *dataset* entra como mais uma das questões enfrentadas, pois a análise foi restrita a 14 atributos de um *dataset* original com 76, o que pode ter excluído informações potencialmente relevantes para uma compreensão mais completa.

O tratamento de *outliers* gera outra limitação, pois houve a decisão de manter *outliers* clinicamente plausíveis no colesterol, embora justificada, pode influenciar certas análises estatísticas.

4.5 Modelagem Preditiva e Avaliação de Desempenho

A presente seção tem como intuito apresentar o avanço da modelagem preditiva através da regressão linear, visando desvendar a relação entre as variáveis e o diagnóstico de doença cardíaca ora representado pela variável “**NovoNum**”.

Em primeira instância, um modelo de regressão linear simples foi ajustado, utilizando “**age**” como variável independente e “**NovoNum**” como dependente, levando à revelação de importantes interpretações de *insights*, como o coeficiente angular inclinação de 0,0123, o coeficiente linear intercepto de 0,2120, e R² coeficiente de determinação de 0,0498, representando respectivamente que, o resultado informa que a cada ano a mais na idade, há um acréscimo de 0,0123 na variável "**NovoNum**". Como "**NovoNum**" é binária (0 ou 1, indicando ausência ou presença de doença, respectivamente), isso pode ser entendido como um pequeno aumento na chance de desenvolver doença cardíaca conforme a idade avança; Do ponto de vista matemático, o valor de 0,2120 corresponde ao ponto em que o modelo cruza o eixo Y, ou seja, o valor atribuído a "**NovoNum**" quando a idade é igual a 0. Apesar de não ter relevância clínica — já que idade zero não se aplica ao estudo de doenças cardíacas — ele faz parte fundamental da equação linear; E o R² que indica a parcela da variação da variável dependente que pode ser explicada pela variável independente. Um R² de 0,0498 (cerca de 5%) é considerado baixo, mostrando que a idade, sozinha, consegue justificar apenas uma pequena parte da variação em "**NovoNum**". Isso significa que a idade, isoladamente, é um preditor fraco para a ocorrência de doença cardíaca, sendo necessário incluir outras variáveis no modelo para torná-lo mais robusto.

Na busca por um modelo mais completo e com maior poder preditivo, aplicou-se a regressão linear múltipla, utilizando todas as variáveis como independentes, exceto "NovoNum" (e a variável original "num"). Para assegurar que o modelo tivesse boa capacidade de generalização e reduzir o risco de *overfitting*, o conjunto de dados foi dividido de forma estratégica: 80% destinado ao treinamento e 20% reservado para o teste ("**test_size=0.2**"). Essa divisão é essencial para avaliar o desempenho do modelo em informações que ele ainda não havia processado. Avaliando este modelo de regressão linear múltipla, percebe-se a produção de resultados promissores, como o erro quadrático médio, MSE de 0,0803 que representa o erro médio das previsões do modelo, que quanto menor, melhor é o indicativo da performance preditiva, assim também como o R² de 0,6781 que se demonstra ser significativamente superior ao do modelo simples, o qual utilizava apenas a idade, demonstrando que o conjunto das variáveis independentes selecionadas é capaz de explicar aproximadamente 67,81% da variância de “**NovoNum**”, indicando um poder substancialmente maior.

A seguir é apresentado a interpretação dos coeficientes da regressão linear múltipla, revelando o impacto individual de cada variável na probabilidade de doença cardíaca, mantendo as outras constantes.

- “**sex**” de **0,1137**: ser homem (valor 1) aumenta a probabilidade de doença cardíaca em 0,1137 em comparação com ser mulher (valor 0). Este achado

condiz com literatura que constantemente aponta diferenças de gênero na prevalência e manifestação de doenças cardíacas.

- **“cp”** de **0,01740**: o tipo de dor no peito (especialmente de atípica para angina típica ou para assintomática) estão associados a um aumento na probabilidade de doença. Isso ressalta que os sintomas de dor torácica são importantes como indicadores.
- **“exang”** de **0,0571:** a presença de angina induzida por exercício (valor 1) eleva as chances de doença cardíaca, se mostrando um indicativo claro de limitação funcional e isquemia miocárdica.
- **“ca”** de **0,0326:** o número de vasos obstruídos (identificados por fluoroscopia) exerce um impacto positivo significativo na probabilidade de doença, o que é clinicamente esperado, pois indica a extensão da aterosclerose coronariana.
- **“thal”** de **0,0160:** o resultado do teste de estresse tálio (exame de imagem que avalia o fluxo sanguíneo para o coração sob condições de estresse e repouso) corresponde a uma influência positiva na probabilidade da doença, sendo um exame fundamental na cardiologia.

No entanto, a análise dos coeficientes também apresentou aspectos que demandam uma interpretação mais minuciosa, como se vê a seguir.

- **“fbs” - açúcar no sangue em jejum de -0,0815:** o coeficiente “**fbs**” foi negativo, podendo sugerir que, ao considerar outras variáveis de risco mais diretas no modelo, a variável "**fbs**" talvez não se mostre um fator de risco independente tão relevante quanto o esperado. Outra possibilidade é que existam interações complexas e não lineares com outras variáveis que o modelo linear não conseguiu captar totalmente. Assim, levanta-se a hipótese de que o impacto da glicemia em jejum na ocorrência de doença cardíaca possa ser influenciado por outras condições, ou que sua relação não seja estritamente linear dentro desse cenário multivariado.
- **"thalach"** de **-0,0005 e "oldpeak"** de **-0,0170:** os coeficientes negativos para "**thalach**" (frequência cardíaca máxima) e "**oldpeak**" (depressão **st**) são inesperados do ponto de vista clínico, pois um valor de frequência cardíaca máxima ou maior depressão de **st** estariam, em primeira instância, associados a uma maior probabilidade de doença, e não a uma menor. Esta aparente discrepância levanta a necessidade de uma análise mais aprofundada que pode indicar que:
    - Há interações complexas entre as variáveis que o modelo linear não consegue capturar completamente.
    - É necessário reavaliar com atenção tanto a escala dos valores quanto a interpretação da variável "**oldpeak**". Se "**oldpeak**" corresponde ao nível de depressão do segmento ST (em que valores mais altos indicam maior depressão e, consequentemente, maior isquemia), a presença de um coeficiente negativo seria inesperada. Isso exigiria uma análise mais aprofundada sobre a adequação do modelo linear para esse tipo de relação, ou até mesmo a consideração de modelos não-lineares.
    - A existência de multicolinearidade entre "**thalach**", "**oldpeak**" e outras variáveis pode estar influenciando de forma inadequada a interpretação de seus coeficientes individuais, algo que também foi observado anteriormente na etapa de Análise Exploratória de Dados (AED).

5. Resultados e Discussão

Esta seção apresentará os resultados gerais da análise exploratória de dados e da regressão linear e discutirá a relação de ambos.

A análise exploratória de dados e as regressões lineares simples e múltiplas contribuíram para uma compreensão adicional dos fatores que influenciam a ocorrência de doença arterial coronariana. A AED destacou tanto os padrões quanto às relações evidentes entre variáveis, enquanto a regressão linear se dedicou a quantificar essas relações e avaliar sua capacidade preditiva.

De maneira geral, a análise apontou que sexo, idade, tipo de dor no peito (**cp**), frequência cardíaca máxima (**thalach**), depressão do segmento ST (**oldpeak**), número de vasos afetados (**ca**), angina provocada pelo exercício (**exang**) e tipo de defeito no exame nuclear (**thal**) estão fortemente relacionados ao diagnóstico de DAC. A análise também revelou diferenças significativas entre os sexos: os homens apresentaram maior prevalência de DAC em idades mais jovens (50–60 anos), enquanto as mulheres foram mais afetadas em idades mais avançadas, especialmente após a menopausa. Além disso, mesmo na ausência de sintomas (**cp=4**), pacientes assintomáticos apresentaram maior tendência à doença, ressaltando a importância da vigilância clínica.

A regressão linear simples, que apenas previu a idade como variável preditora, obteve um baixo poder explicativo (R^2=0,0498). Isso fortaleceu a conclusão da análise de que a idade, apesar de ser um fator de risco significativo, não é suficiente para prever a presença da doença por si só.

Em contrapartida, a regressão linear múltipla, que avaliou todas as variáveis disponíveis, revelou resultados muito mais sólidos (R^2=0,6781). As variáveis **sexo**, **cp**, **exang**, **ca** e **thal** mostraram uma correlação positiva significativa com o diagnóstico de DAC nesse modelo, o que quantitativamente valida os resultados da AED. Essa concordância entre os métodos fortalece a evidência de que essas variáveis são essenciais para reconhecer a doença.

Contudo, alguns pontos de divergência apareceram. A AED havia apontado que a presença de DAC estava relacionada a valores mais altos de **oldpeak** e mais baixos de **thalach**. No entanto, na análise de regressão múltipla, os coeficientes dessas variáveis apresentaram sinais negativos, indicando uma relação inversa. Essa diferença pode ser justificada por elementos como:

- **Multicolinearidade**, pois variáveis inter-relacionadas podem alterar os coeficientes individuais do modelo;
- **Restrições do modelo linear**, que pode não ser capaz de identificar relações mais complexas ou não lineares;
- **A escala e a interpretação das variáveis**, particularmente no que diz respeito ao oldpeak, cujo valor clínico demanda cuidado na análise dos coeficientes.

Outra discordância surgiu em relação ao **fbs** (açúcar no sangue em jejum), que exibiu um coeficiente negativo na regressão múltipla. Apesar de ser tradicionalmente visto como um fator de risco cardiovascular, esse achado indica que, ao ajustar para variáveis mais diretamente relacionadas à função cardíaca, o efeito independente do **fbs** diminui. Isso pode sugerir que o modelo não captou algumas interações ou que são necessárias amostras mais representativas para entender melhor seu papel.

Assim, a combinação entre análise exploratória e regressão linear revela tanto convergências (como a importância de **ca**, **cp**, **exang**, **thal** e **sexo**) quanto divergências (principalmente nos coeficientes de **oldpeak**, **thalach** e **fbs**). Esses últimos foram destacados como variáveis mais relevantes na detecção de doenças cardíacas apenas por meio da análise exploratória, o que, posteriormente, com a regressão linear, mostrou-se contrário. Essa comparação destaca que a análise exploratória gera hipóteses e orientações iniciais, ao passo que a modelagem estatística se dedica a quantificar essas conexões, mesmo estando sujeita a restrições metodológicas. Em conjunto, essas abordagens proporcionam uma perspectiva mais abrangente e crítica do fenômeno de estudo deste artigo, facilitando uma melhor compreensão dos fatores que causam a doença cardíaca.

6. Conclusão

Foram feitas análises exploratórias dos dados e modelagem preditiva utilizando regressão linear e interpretação dos coeficientes, o que permitiu concluir que houveram convergências e divergências na comparação do resultados de ambos, onde as variáveis **ca**, **cp**, **exang**, **thal** e **sex** se mostraram relevantes para a detecção de doenças cardíacas tanto na análise exploratória quanto na regressão linear, enquanto as demais variáveis apresentaram discrepâncias, sendo coeficientes inconsistentes. Foi avaliado que depois do tratamento dos dados, ao realizar a avaliação do modelo de regressão linear múltipla, o coeficiente de determinação (R^2) apresentou que o conjunto das variáveis independentes selecionadas foi capaz de explicar 67,81% da variância do atributo alvo, indicando um poder preditivo substancialmente maior.

Examinar mais profundamente esses coeficientes considerados “inconsistentes” evidencia que interpretar modelos vai além da leitura direta dos números, sendo necessário relacioná-los constantemente ao conhecimento da área (neste caso, a cardiologia) e adotar uma postura crítica diante das premissas do modelo. Esse aspecto, em especial, mostra a complexidade da modelagem preditiva em contextos reais e reforça a importância de validar os resultados junto a especialistas do domínio.

Este trabalho traz contribuições que corroboram com aspectos preexistentes da literatura, como o fato de que os homens são mais propensos a terem doenças cardíacas em idades mais novas, levando em consideração que neste *dataset* há um maior número de homens. Revelou também que para as mulheres essa idade sobe, pois os casos mais ocorrentes acontecem após a menopausa, devido a certas disfunções hormonais. Ressalta a importância da atenção clínica para os sintomas de dor torácica como indicador, a presença de angina induzida pelo exercício, o número de vasos obstruídos e o resultado do teste de estresse tálio.

Para trabalhos futuros, recomenda-se a ampliação do *dataset*, de modo a incluir amostras mais recentes e representativas de diferentes perfis populacionais, o que poderia aumentar a robustez e a generalização dos resultados. Além disso, é relevante a exploração de modelos não-lineares, como algoritmos de aprendizado de máquina mais avançados, capazes de capturar relações complexas entre as variáveis que a regressão linear não contempla plenamente. Outro caminho promissor é olhar com mais atenção para as interações entre variáveis, especialmente aquelas que apresentaram resultados inesperados, buscando entender como elas se combinam na prática. Além disso, uma validação clínica junto a profissionais de saúde é

essencial, para garantir que os achados façam sentido também no dia a dia do atendimento médico. A inclusão de novos atributos, como informações sobre hábitos de vida, histórico familiar ou exames laboratoriais mais detalhados, também pode enriquecer bastante a análise. Por fim, um estudo mais cuidadoso das diferenças de gênero na manifestação da doença pode trazer *insights* importantes, já que homens e mulheres não são afetados da mesma forma ao longo da vida.

Referências

AMARAL, Fernando. **Aprenda Mineração de Dados**: Teoria e Prática. 1. ed. [S. l.]: Alta Books, 2016.

BARI, Anasse; CHAOUCHI, Mohamed; JUNG, Tommy. **Análise Prediitiva para Leigos 2019.** 2. ed. [S. l.]: Alta Books, 2019.

BHATIA, Manpreet Kaur. Data Analysis and its Importance**. International Research Journal of Advanced Engineering and Science**, [s. l.], v. 2, ed. 1, p. 166-168, 2017.

BHATLA, Nidhi; JYOTI, Kiran. An Analysis of Heart Disease Prediction using Different Data Mining Techniques. **International Journal of Engineering Research & Technology (IJERT),** [s. l.], v. 1, October 2012.

COURSERA. **7 Data Analysis Software Applications You Need to Know**. [S. l.], 15 dez. 2024. Disponível em: https://www.coursera.org/articles/data-analysis-software. Acesso em: 22 mar. 2025

FERREIRA , Rafael; MIRANDA , Leandro; PINTO , Rafael; PESSUTTO, Lucas; PEREIRA , Mariana; MARQUES, Leonardo; ANDRADE , Ana Luiza. **Preparação e Análise Exploratória de Dados**. [*S. l.*: *s. n.*], 2021

GRUS, Joel. **Data Science do Zero**: Noções fundamentais com Python. 2. ed. [*S. l.*]: Alta Books, 2021.

JUPYTER. **Jupyter**. [*S. l.*], 14 abr. 2025. Disponível em: https://jupyter.org/. Acesso em: 14 abr. 2025.

LIMA, Kaique. **Ranking IEEE | Veja as linguagens de programação mais populares de 2022**. [*S. l.*], 30 ago. 2022. Disponível em: https://canaltech.com.br/software/ranking-ieee-veja-as-linguagens-de-programacao-mais-populares-de-2022-224274/. Acesso em: 11 abr. 2025.

MARQUES, Paulo; CÉSAR, Hilton; SUZUKI, Katia; SCARPELINI, Sandro; MARQUES, João; COSTA, José; RAINER, Peter; KRAMER, Diether; JAUK, Stefanie; ROMÃO, Elen; SCHREMPF, Michael; SHIMIZU, Gilson. Machine learning-based risk prediction for major adverse cardiovascular events in a Brazilian hospital: Development, external validation, and interpretability. **PLOS ONE**, [s. l.], 11 out. 2024.

MATPLOTLIB. **Matplotlib**. [*S. l.*], 11 abr. 2025. Disponível em: https://matplotlib.org/. Acesso em: 11 abr. 2025.

MORAIS, Carlos. **Escalas de medida, estatística descritiva e inferência estatística**. 2005.

NUMPY. **NumPy**. [*S. l.*], 11 abr. 2025. Disponível em: https://numpy.org/. Acesso em: 11 abr. 2025.

OPAS, ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Doenças cardiovasculares**. [S. l.], 17 mar. 2025. Disponível em: https://www.paho.org/pt/topicos/doencas-cardiovasculares. Acesso em: 14 mar. 2025.

PANDAS. **Pandas**. [*S. l.*], 11 abr. 2025. Disponível em: https://pandas.pydata.org/about/. Acesso em: 11 abr. 2025.

PANDAS. **Pandas**. [*S. l.*], 11 abr. 2025. Disponível em: https://pandas.pydata.org/. Acesso em: 11 abr. 2025.

PILOTTO, Karina. **As linguagens de programação mais usadas**. [*S. l.*], 21 jan. 2025. Disponível em: https://www.caiena.net/blog/linguagens-de-programacao-mais-usadas. Acesso em: 11 abr. 2025.

PYTHON. **Python**. [*S. l.*], 11 abr. 2025. Disponível em: https://www.python.org/. Acesso em: 11 abr. 2025.

RAHEJA, Kavita; DUBEY, Ajay; CHAWDA, Rahul. Data Analysis and its Importance in Health Care**. International Journal of Computer Trends and Technology (IJCTT)**, [s. l.], v. 48, ed. 4, p. 176-180, Junho 2017. Disponível em: http://www.ijcttjournal.org. Acesso em: 19 mar. 2025.

SANTOS, Carla. **Estatística Descritiva**: Manual de Auto-Aprendizagem. 3. ed. [S. l.]: Edições Sílabo, 2018.

SEABORN. **Seaborn**. [*S. l.*], 11 abr. 2025. Disponível em: https://seaborn.pydata.org/. Acesso em: 11 abr. 2025.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining**: Mineração de dados. [S. l.]: Ciência Moderna, 2009.

TIOBE. **Sobre a organização TIOBE**. [*S. l.*], 11 abr. 2025. Disponível em: https://www-tiobe-com.translate.goog/about-us/?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=tc. Acesso em: 11 abr. 2025.

UCI IRVINE MACHINE LERANING REPOSITORY. **Heart Disease**. [*S. l.*], 30 jun. 1988. Disponível em: https://archive.ics.uci.edu/dataset/45/heart+disease. Acesso em: 14 abr. 2025.

UNITED NATIONS. **O fardo das doenças não transmissíveis**. [S. l.], 17 mar. 2025. Disponível em: https://www-un-org.translate.goog/en/observances/interventional-cardiology-day?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=wa. Acesso em: 14 mar. 2025.

VARGAS, Brenda; KERN, Karoline; SILVA, Ingrid; SILVEIRA, Eliane; MARTINS, Maria Isabel. **DOENÇAS CARDIOVASCULARES EM MULHERES NO CLIMATÉRIO/MENOPAUSA**. [*S. l.*], 10 set. 2019. Disponível em: http://www.conferencias.ulbra.br/index.php/sic/sic25/paper/view/13002. Acesso em: 9 set. 2025.

WORLD HEART FEDERATION. **What is cardiovascular disease?.** [S. l.], 17 mar. 2025. Disponível em: https://world-heart-federation.org/what-is-cvd/. Acesso em: 14 mar. 2025.