UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Patrick Regis reis dos Anjos, 2104320
[redacted]

Desenvolvimento de um modelo de classificação do risco de doenças cardíacas usando aprendizado de máquina

Guarujá – SP
2024

UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Desenvolvimento de um modelo de classificação do risco de doenças cardíacas usando aprendizado de máquina

Relatório Técnico-Científico apresentado na disciplina de Projeto Integrador para o curso de Bacharelado em Ciência de dados da Universidade Virtual do Estado de São Paulo (UNIVESP).

Guarujá – SP
2024

[redacted]; dos Anjos, Patrick Regis Reis. **Desenvolvimento de um modelo de classificação do risco de doenças cardíacas usando aprendizado de máquina.** Relatório Técnico-Científico. Bacharelado em Ciência de Dados – **Universidade Virtual do Estado de São Paulo**. Tutor: [redacted]. Polo Guarujá, 2024.

RESUMO

A detecção precoce do risco de doenças cardíacas continua a ser um desafio significativo na área da saúde. Com o aumento do poder computacional e a abundância de dados clínicos disponíveis, surgiram novas técnicas eficazes de aprendizado de máquina para predições baseadas em dados. O objetivo deste projeto foi desenvolver um modelo de classificação para prever o risco de doenças cardíacas, utilizando redes neurais profundas, e um dashboard para auxiliar na tomada de decisão baseada em dados. A Unidade de Saúde da Família [redacted], no Guarujá, forneceu o suporte técnico e teórico necessário para o desenvolvimento do modelo. O dataset utilizado foi extraído do repositório de aprendizado de máquina da Universidade da Califórnia – Irvine. A metodologia adotada para o desenvolvimento do modelo foi o KDD, que orientou a exploração e preparação dos dados, além da aplicação de técnicas analíticas para identificar padrões críticos. O resultado do projeto foi uma rede neural utilizando técnicas sofisticadas, como regularização, normalização e outras. O uso de redes neurais possibilitou capturar correlações complexas entre as variáveis, melhorando a acurácia na predição de risco. Os resultados obtidos indicam que o modelo pode fornecer suporte valioso no diagnóstico precoce de doenças cardíacas, ajudando a priorizar pacientes em risco através dos dados. A integração do dashboard visual se mostrou crucial para aumentar a compreensão dos dados, explicando suas correlações e facilitando a tomada de decisões.

PALAVRAS-CHAVE: Aprendizado de Máquina, Redes Neurais, Doenças Cardíacas, Visualização de Dados, Dashboard.

SUMÁRIO

ÍNDICE DE FIGURAS

1. INTRODUÇÃO

O diagnóstico precoce para doenças cardíacas ainda é um grande desafio no âmbito da saúde, devido à alta taxa de mortalidade associada a essas condições. A complexidade dos seus fatores de risco, que inclui variáveis demográficas, comportamentais e clínicas, torna a previsão dessas doenças uma tarefa complexa. A identificação precoce desses fatores pode ser responsável por melhorar significativamente a intervenção médica e reduzir a mortalidade, exigindo ferramentas analíticas cada vez mais sofisticadas para processar e interpretar dados clínicos volumosos.

O tema do presente trabalho surgiu a partir de conversas com funcionários do posto de saúde [redacted], no Guarujá. Durante as conversas com a comunidade do trabalho, foi identificado um desafio comum enfrentado pelos profissionais da saúde: a dificuldade em predizer com precisão o risco e doenças cardíacas com base nas informações clínicas disponíveis, dado a complexidade das variáveis envolvidas. Após uma breve sessão de brainstorming e posterior apresentação do tema aos funcionários do posto, esses expressaram grande interesse em uma ferramenta mais eficaz para análise de dados e classificação de risco de doenças cardíacas, o que levou a concepção deste projeto.

Neste projeto, desenvolvemos um modelo de aprendizado de máquina, útil para realizar a classificação do risco de doenças cardíacas na base de dados. Decidimos usar o conjunto de dados Heart Disease Dataset, da Universidade da Califórnia – Irvine (UCI). Esse conjunto de dados contém informações clínicas e demográficas importantes, que permitem a identificação de padrões e fatores de risco associados à doença cardíaca. A principal ideia que cerca o projeto é ser capaz de transformar os dados brutos no conjunto de dados em predições úteis para médicos e profissionais da saúde, facilitando a tomada de decisão. Utilizaremos um modelo computacional de redes neurais para o projeto, de maneira que o mesmo seja capaz de classificar os pacientes de maneira eficiente e confiável. A análise terá enfoque no conjunto de dados da UCI por ser um conjunto de dados que contém variáveis comuns à prática clínica, medições frequentemente coletadas em hospitais e estabelecimentos de saúde. Isso é importante, pois permite que o modelo de classificação desenvolvido seja mais generalizado e possa ser usado em conjunto com outras fontes de dados semelhantes.

O projeto está inserido no contexto da ciência de dados e análise preditiva na saúde. O tema foi escolhido motivado pela necessidade de ferramentas mais modernas para a identificação precoce dos riscos associados a doenças cardíacas. O objetivo geral do trabalho é a construção e validação de uma rede neural, além de um dashboard interativo capaz de facilitar a interpretação dos dados e contribuir para o processo de tomada de decisão pelos profissionais da saúde. O presente estudo busca aprimorar a predição e oferecer uma ferramenta útil para a prática clínica, além de promover uma abordagem na prevenção de doenças baseada em uma gestão de dados.

2. DESENVOLVIMENTO

2.1. OBJETIVOS

Objetivo Geral: Desenvolver uma rede neural capaz de classificar o nível de risco para doenças cardíacas através dos dados no Heart Disease Dataset da UCI e criar um dashboard interativo para facilitar a visualização de dados e auxiliar na tomada de decisão.

Objetivos Exploratórios:

- Conhecer e identificar os principais fatores de risco para doenças cardíacas presentes no dataset. Aqui é importante trabalhar com medidas relacionadas com o domínio do problema, como frequência cardíaca máxima, presença de dor no peito, açúcar no sangue em jejum, entre outras.
- Levantar material bibliográfico e referências sobre técnicas de aprendizado de máquina aplicadas à previsão de doenças cardíacas.
- Descobrir as melhores práticas e metodologias para a construção e validação de modelos preditivos de risco de doenças cardíacas.

Objetivos Descritivos:

- Descrever os métodos e técnicas de aprendizado de máquina utilizados para o desenvolvimento do modelo de classificação.
- Traçar os caminhos e etapas do desenvolvimento do modelo e do dashboard interativo.

Objetivos Explicativos:

- Analisar a eficácia do uso de redes neurais para encontrar insights e correlações nos dados.
- Verificar o impacto da solução proposta acionada na comunidade.
- Explicar o processo de desenvolvimento e o funcionamento dos artefatos do trabalho.

2.2 JUSTIFICATIVA E DELIMITAÇÃO DO PROBLEMA

Atualmente, a análise do risco de doenças cardíacas frequentemente depende de métodos tradicionais que podem ser insuficientes para lidar com a complexidade e a variedade dos dados clínicos disponíveis. Esses métodos muitas vezes não integram dados de forma eficiente, dificultando a interpretação precisa e a tomada de decisões informadas.

A pergunta que norteia o projeto nesse contexto é: como podemos desenvolver uma ferramenta que utilize aprendizado de máquina para classificar de forma o risco de doenças cardíacas com evidências baseadas em dados e apresentar essas informações de maneira intuitiva e acessível, de modo a tornar-se um projeto útil para tomada de decisão para médicos e funcionários da saúde?

A necessidade de uma solução eficaz é impulsionada pela crescente demanda por diagnósticos mais precisos e pela dificuldade em integrar e interpretar dados clínicos complexos. Dados esses fatores, foi pensado em construir uma solução utilizando um modelo preditivo baseado em aprendizado de máquina e um dashboard interativo para apresentar os dados. Esta abordagem proporcionará uma ferramenta valiosa para a prática clínica, oferecendo uma maneira mais eficiente e informada de avaliar o risco de doenças cardíacas e melhorar a gestão da saúde na [redacted].

2.3 FUNDAMENTAÇÃO TEÓRICA

Para solucionar o problema foi decidido criar uma solução integrada que utilize técnicas avançadas de análise e visualização de dados. O objetivo é desenvolver um modelo de aprendizado de máquina para prever o risco de doenças cardíacas e uma plataforma de visualização interativa para facilitar a interpretação dos resultados.

O primeiro passo para o desenvolvimento dessa solução é a coleta e preparação dos dados, que é realizado através do processo de Knowledge Discovery in Databases (KDD). KDD é uma abordagem sistemática para descobrir informações valiosas a partir de grandes conjuntos de dados. O KDD pode ser dividido em algumas fases O KDD é caracterizado por um processo estrutural para descoberta de conhecimento em bancos de dados volumosos, denominado ciclo de vida dos dados. Esse processo é dividido em:

- Seleção: Escolha dos dados relevantes para análise.
- Pré Processamento: Limpeza e tratamento dos dados. Resume-se na remoção de inconsistências e preenchimento de dados faltantes.
- Transformação: Converter os dados em formatos apropriados para as técnicas de mineração. Diz-se sobre usar técnicas como normalização, agrupamentos, entre outras.
- Mineração de dados (Data Mining): Aplicação de algoritmos para elucidar padrões ocultos ou extrair modelos nos dados.
- Interpretação e avaliação: Análise dos dados para extração de conhecimento útil e validação dos modelos.

O KDD (Knowledge Discovery in Databases) é uma metodologia comprovada para a extração de conhecimento útil de grandes volumes de dados. Ele oferece uma abordagem sistemática para transformar os brutos em informações valiosas, além de ser amplamente reconhecida por sua eficácia em garantir a confiabilidade dos dados e a robustez no desenvolvimento de modelos. Santos (2018) usa o processo de KDD em seu trabalho para identificar se existem padrões na estimativa local de evapotranspiração com características

semelhantes, com intuito de generalizar um modelo para localidades similares sem estações de medição.

> Buscou-se, neste estudo inicial, verificar a existência de padrões regionais para a estimativa da evapotranspiração, com o uso de KDD. Através da regressão linear, chegou-se a um modelo de cálculo da evapotranspiração¸ao para uma localidade (Resende-RJ), que foi usado em outras localidades de latitudes similares, para verificar se os resultados estariam próximos dos valores históricos, e em localidades de latitudes não similares, para verificar se os resultados teriam uma correlação bem mais baixa em relação aos valores históricos da localidade (SANTOS, 2018, p. 5 – 6).

Realizada a coleta, pré-processamento e a transformação dos dados, podemos passar para o processo de Data Mining. Segundo Stainer (2006), enquanto o KDD refere-se ao processo completo de descoberta de conhecimento útil em bases de dados, o Data Mining é a etapa do KDD que envolve a aplicação de técnicas para extrair padrões e modelos dos dados. O aprendizado de máquina, utilizado no projeto, é um campo da inteligência artificial que utiliza algoritmos para permitir que computadores aprendam com os dados, permitindo a eles a capacidade de realizar previsões ou tomar decisões sem intervenção humana. As duas abordagens mais importantes do aprendizado de máquina são: aprendizado supervisionado e não supervisionado.

No aprendizado supervisionado, o modelo é treinado com dados rotulados, com respostas já conhecidas, para prever a saída para novos dados. O modelo aprende com as respostas já conhecidas e generaliza para predizer novas respostas, baseado no conhecimento adquirido com os dados no processo de aprendizado. Já o aprendizado não supervisionado é um tipo de aprendizado onde o modelo é treinado com dados que não possuem rótulos. O objetivo é simplesmente identificar padrões ou estruturas ocultas nos dados de forma autônoma.

> No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Em geral, cada exemplo é descrito por um vetor de valores de características, ou atributos, e o rótulo da classe associada. O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham o rótulo da classe. Para rótulos de classe discretos, esse problema é conhecido como classificação e para valores contínuos como regressão (MONARD; BARANAUSKAS, 2003, p. 32).

Para melhorar a precisão do nosso modelo, discutimos sobre o aprendizado profundo, uma subcategoria do aprendizado de máquina que utiliza redes neurais artificiais complexas com várias camadas profundas para ser capaz de modelar dados com alta complexidade. Cada neurônio na rede neural simula a função de um neurônio biológico, processando e transmitindo informações. Optamos pelo aprendizado profundo pois como as redes neurais profundas são particularmente eficazes em capturar padrões complexos nos dados, essas podem melhorar significativamente a performance do modelo de classificação e torná-lo mais confiável e acurado.

> Em sua forma mais geral, uma rede neural é um sistema projetado para modelar a maneira como o cérebro realiza uma tarefa particular, sendo normalmente implementada utilizando-se componentes eletrônicos ou é simulada por propagação em um computador digital. Para alcançarem bom desempenho, as redes neurais empregam uma interligação maciça de células computacionais simples, denominadas de "neurônios" ou unidades de processamento (HAYKIN, 2001, apud FLECK et al, 2016, p. 48).

Após treinar o modelo para adaptar-se aos dados, temos que avaliar sua eficácia usando métricas de validação. Essas métricas são fundamentais para medir o desempenho do modelo e incluem:

- A acurácia é o percentual de previsões corretas.
- Precisão é a proporção de verdadeiros positivos em relação ao total de positivos previstos.
- Recall é a proporção de verdadeiros positivos em relação ao total de positivos reais.

Para tornar os resultados compreensíveis e de fácil acesso, criamos um dashboard interativo utilizando técnicas de visualização de dados. Visualização de dados refere-se à representação gráfica das informações, o que facilita a interpretação dos resultados e a tomada de decisões.

> Definimos o conceito de Visualização de Dados como a técnica de transformar um conjunto complexo de dados em visualizações gráficas de modo a constituir uma representação visível dos dados que estavam "invisíveis" e que passam a ser manipulados por algoritmos em sistemas computacionais para a estruturação de um conteúdo (RODRIGUES, 2019, p. 97).

Um dashboard é um recurso visual que apresenta gráficos, tabelas e painéis interativos, explorando os dados de forma dinâmica e intuitiva. O dashboard do projeto será criado utilizando uma biblioteca da linguagem Python, chamada Plotly. Plotly é uma biblioteca de visualização de dados interativa, amplamente utilizada para criar gráficos dinâmicos. Plotly é multilinguagem, disponível em Python, Javascript, R e outras linguagens de programação, além de ser uma biblioteca especialmente útil em projetos que exigem visualizações responsivas e ricas em detalhes.

O dataset da UCI Heart Disease é amplamente reconhecido por sua eficiência e confiabilidade para estudos na construção de modelos para predição de doenças cardíacas. Um estudo publicado por Alfadli e Almagrabi (2023) demonstra a capacidade do conjunto de alcançar uma alta acurácia, precisão e recall utilizando o mesmo subconjunto de variáveis do projeto. A robustez do conjunto também foi comprovada em pesquisa, oferecendo uma forte estrutura para desenvolver modelos preditivos no domínio dos dados médicos. Além disso Mukherjee e Sharma explicam sobre a eficiência do uso das redes neurais utilizando o dataset da UCI, validando um modelo de alta eficiência.

> Nós treinamos o modelo usando o conjunto de dados mencionado acima e testamos com 30 dados de pacientes não vistos anteriormente pelo modelo. O modelo tem uma acurácia promissora de 97% em dados de pacientes ainda não vistos. Conforme definido, derivamos a acurácia: Acurácia = (vn + vp)/(vn+vp+fn+fp) = (12+17)/(12+17+1+0) = 0,97 (resultado do modelo) – onde vn é o número de verdadeiros negativos, vp é o de verdadeiros positivos e fn e fp são respectivamente os números de falsos negativos e falsos positivos. O denominador é igual ao total (MUKHERJEE; SHARMA, p. 402 – 405p, 2019, tradução nossa).

2.4 APLICAÇÃO DAS DISCIPLINAS ESTUDADAS NO PROJETO INTEGRADOR

As disciplinas Algoritmos e Programação de Computadores I e II foram fundamentais para introduzir os conceitos básicos de programação e estrutura de dados. Esta foi ministrada em Python, a linguagem que escolhemos para o desenvolvimento deste projeto. Python se mostrou uma escolha adequada, tanto pela sua versatilidade quanto pela vasta quantidade de bibliotecas que oferece para aprendizado de máquina e manipulação de dados.

As disciplinas de Cálculo desempenharam um papel essencial na construção do modelo de redes neurais. O entendimento de derivadas parciais, por exemplo, é indispensável para compreender o funcionamento do algoritmo de retropropagação, utilizado no treinamento das redes neurais para ajustar os pesos das conexões internas. Esse conhecimento matemático forneceu a base necessária para implementar e otimizar o processo de aprendizagem do modelo. Estatística e Probabilidade foram igualmente importantes para a validação do modelo. Essas disciplinas nos ensinaram a interpretar métricas estatísticas fundamentais, como a matriz de confusão, e a calcular indicadores de performance, como precisão, recall e F1-score.

As disciplinas de Mineração de Dados e Aprendizado de Máquina foram cruciais para a preparação e análise do dataset utilizado. A Mineração de Dados nos capacitou a realizar uma análise prévia e preparar os dados, procedimento indispensável para melhorar a performance do modelo. Já Aprendizado de Máquina ofereceu uma base sólida sobre conceitos fundamentais de modelos supervisionados, além de técnicas específicas para treinar e avaliar modelos, o que foi diretamente aplicado no desenvolvimento do classificador.

As disciplinas Redes Neurais e Redes Profundas forneceram o arcabouço teórico necessário para o desenvolvimento, otimização e validação da rede neural desenvolvida no projeto. O conhecimento adquirido nessas disciplinas permitiu explorar diferentes arquiteturas e métodos de otimização para obter melhores resultados.

Para a construção do dashboard, as disciplinas Visualização de Dados e Desenvolvimento Web foram especialmente importantes. Visualização de Dados ajudou a escolher gráficos e componentes visuais que pudessem transmitir os resultados de maneira clara e intuitiva. Além disso, Desenvolvimento Web foi essencial na formatação e construção visual do dashboard, utilizando HTML e CSS para organizar a interface e garantir uma

apresentação visual atraente e de fácil navegação. Essa combinação de habilidades possibilitou a criação de um painel interativo e informativo, que auxilia na análise dos resultados e no entendimento dos dados.

2.4 METODOLOGIA

O desenvolvimento do projeto começou com a proposta de criar uma solução voltada para a melhoria das ferramentas de classificação de doenças baseadas em dados. Após um período de pesquisa e análise sobre a iniciativa, decidiu-se que o foco principal seria o domínio das doenças cardíacas. Um dos integrantes do grupo já havia trabalhado anteriormente em um posto de saúde local, [redacted], o que facilitou a descoberta da comunidade do trabalho e o contato com os profissionais da área que participaram na concepção e avaliação do projeto.

Iniciamos o período de idealização do projeto conversando com alguns funcionários do posto de saúde, conhecidos do integrante, especificamente o gerente da unidade (atualmente gerente substituto) e a médica da saúde da família da área laranja, uma das quatros áreas que a unidade abrange. Foi feita uma coleta por meio de entrevistas semiestruturadas, nas quais os profissionais compartilharam suas experiências e necessidades. Essa coleta de dados foi essencial para identificar as demandas e os principais requisitos técnicos que norteiam um bom desenvolvimento para o modelo de classificação. Os profissionais demonstraram interesse na ideia do projeto de classificação, ressaltando a dificuldade de integrar dados clínicos de pacientes com análises preditivas eficientes.

Apesar de fora do escopo do atendimento realizado pelo posto de saúde na maioria dos casos, o interesse na ferramenta se deu não somente pela natureza preditiva da ferramenta a ser desenvolvida, que se atrelada a um conjunto de dados generalizados seria de grande eficiência para outros conjuntos de dados similares, mas também pela oportunidade de tomada de decisão em tempo real baseado em dados mais eficiente, já que o projeto poderia explicar as correlações entre as variáveis presentes nos dados, ressaltando visualmente os fatores mais importantes que levam ao risco de doença cardíaca.

Para o desenvolvimento do projeto utilizamos o Heart Disease Dataset do repositório de aprendizado de máquina da universidade da Califórnia – Irvine (UCI). É um dataset amplamente reconhecido pela comunidade acadêmica por fornecer dados clínicos importantes para a análise cardíaca. Para nosso trabalho, o dataset é importante principalmente por ser generalista e trabalhar com dados úteis e medidas facilmente realizadas em hospitais. O dataset contém informações de pacientes, como idade, sexo, pressão arterial, níveis de

colesterol, resultados de exames de estresse, entre outros fatores de risco, que são essenciais para avaliação da probabilidade de risco de doença cardíaca.

Originalmente o dataset é composto por 76 atributos, mas apenas 14 deles são amplamente utilizados em estudos. Os atributos no nosso dataset são:

- Idade (age): A idade do paciente em anos.
- Sexo (sex): Um valor binário que indica o sexo do paciente.
- Dor no peito (cp): Indica o tipo de dor no peito apresentado. Pode ser assintomático, angina típica, angina atípica e dor não cardíaca.
- Pressão arterial em repouso (trestbps): Pressão arterial sistólica enquanto em repouso, medida em milímetros de mercúrio (mmHg).
- Colesterol Sérico (chol): Nível de colesterol total no sangue (mg/dl).
- Nível de glicose em jejum (fbs): Valor binário que indica se o paciente tem diabetes. Glicose em jejum > 120 mg/dl.
- Eletrocardiograma em Repouso (restecg): Resultado do eletrocardiograma, podendo assumir os valores: Normal, anormalidade da onda ST-T e hipertrofia ventricular esquerda (0, 1 ou 2).
- Frequência Cardíaca Máxima (thalach): A frequência cardíaca máxima que foi alcançada durante um teste de estresse. Indica capacidade cardiovascular.
- Angina Induzida por Exercício (exang): Valor binário que indica se o paciente apresenta dor no peito durante o exercício.
- Depressão Induzida por Exercício (oldpeak): Um valor que mede a depressão do segmento ST durante o teste de estresse, indicando possíveis anormalidades.
- Pico de Estresse (slope): Inclinação do segmento ST durante o exercício. Pode ser descendente, plana ou ascendente (0, 1 ou 2).
- Número de Vasos Coloridos por Fluoroscopia (Ca): A quantidade de vasos sanguíneos visíveis na fluoroscopia (0-3), que pode indicar a presença de obstruções.
- Talassemia (Thal): Um valor categórico que indica a presença de talassemia. Normal, fixa ou reversível (0, 1 ou 2)
- Classificação da Doença Cardíaca (target): Um valor que indica a presença ou ausência de doenças cardíacas, variando de 0 (sem doença) a 4 (doença grave).

O projeto seguiu com a metodologia Knowledge Discovery in Databases (KDD) para realizar a análise exploratória dos dados. KDD é um processo de descoberta de conhecimento em bases de dados, constituído das seguintes etapas:

- Seleção dos dados: Como dito anteriormente, utilizamos os dados do Heart Disease Dataset da UCI, contendo informações relevantes para nosso modelo de classificação.
- Pré-processamento e Transformação: Realiza-se a limpeza dos dados e o preparo para os algoritmos de aprendizado de máquina. Não foi necessário realizar grandes mudanças pois o dataset da UCI já é disponibilizado tratado. Nesta etapa trabalhamos principalmente com a separação do conjunto em dados para treino, teste e validação. Além disso, foi realizada também a codificação das variáveis categóricas e a normalização do conjunto de dados, técnicas importantes para garantir um melhor desempenho e acurácia no modelo de classificação.
- Mineração de dados: Selecionamos os algoritmos de aprendizagem de máquina capazes de solucionar nosso problema. No caso do projeto, foi decidido trabalhar com as redes neurais como abordagem para construção do modelo computacional de classificação. A escolha de redes neurais foi pautada na sua capacidade de resolver problemas complexos e sua versatilidade para resolver uma ampla variedade de problemas.
- A Interpretação e Avaliação: Aqui interpretamos os dados, conhecendo seu significado implícito. Além disso, utilizamos métricas de avaliação para determinar a eficácia do modelo. As métricas utilizadas no projeto foram: acurácia, precisão e F1-Score.
- Visualização dos Dados: Apresentação visual dos resultados para facilitar a comunicação e interpretação. Foi desenvolvido um dashboard interativo, painel de informações, para apresentar visualmente os resultados do projeto. O dashboard foi construído utilizando a tecnologia Plotly. Plotly é uma biblioteca da linguagem Python que permite a criação de gráficos interativos de maneira facilitada e eficiente.

A solução desenvolvida foi um modelo de classificação de aprendizado profundo, capaz de prever o risco de doenças cardíacas, treinado e validado utilizando do dataset Heart

Disease Dataset da UCI. A escolha por redes neurais foi baseada em sua capacidade de modelar relações complexas entre os dados e identificar padrões ocultos, que métodos mais simples de aprendizado de máquina podem não conseguir capturar. As redes neurais são eficazes quando trabalhamos com modelos de classificação, porque lidam bem com a multiplicidade de variáveis e podem aprender profundamente o nível de relevância de cada uma. A arquitetura em camadas atreladas aos modelos profundos permite à rede neural encontrar interações entre os atributos clínicos e gerar uma previsão precisa do risco de doença cardíaca.

O foco do projeto foi em trazer eficácia nas predições do modelo, para que essa possa ser usada por profissionais da saúde na tomada de decisão. O dashboard interativo do projeto foi desenvolvido para apresentar os resultados de maneira mais clara, auxiliando na interpretação e comunicação dos dados.

Após o desenvolvimento do modelo, a solução foi testada com base nas métricas de avaliação apresentadas acima, utilizando um conjunto de dados separados para validação. Apesar da construção bem-sucedida do modelo e os novos métodos de otimização, seus parâmetros de avaliação continuaram abaixo do esperado. Utilizando o dataset do projeto, o modelo foi capaz de prever corretamente somente 65% dos dados, aproximados. Na tentativa de mitigar o problema, utilizamos o mesmo método de avaliação para um novo conjunto de dados gerado a partir do original, com a variável alvo somente oscilando entre 0 e 1. Na prática, reduzimos o problema para uma classificação binária, que diz se há ou não presença de doença cardíaca. No novo dataset, foi obtido uma precisão de 85%, indicando um modelo robusto e útil para resolver esse problema reduzido.

Apesar de não alcançarmos as metas de desempenho previstas, ainda existem estratégias viáveis para mitigar esse problema. Após uma análise detalhada da arquitetura da rede, concluiu-se que o desempenho inesperado não decorre da construção do modelo em si, mas sim da escassez de dados no dataset utilizado. Esse fator limita a capacidade do modelo de generalizar com eficácia, prejudicando sua performance. Há alternativas para enfrentar essa limitação, como a coleta de novos dados ou o uso de técnicas de data augmentation, que ampliariam a variedade de padrões no conjunto de dados existente. No entanto, para manter o foco e a viabilidade do presente trabalho, optou-se por não expandir o dataset. Essa decisão

está embasada em restrições de tempo e recursos, além do objetivo de validar a robustez do modelo com os dados disponíveis.

O feedback dos profissionais elogiou o uso dos gráficos para explicar o comportamento dos gráficos, apontando para um significante aumento na compreensão dos dados. O principal elogio foi na capacidade de compreender as correlações entre os dados, o que permitiu confirmar com base em dados, e de maneira visual, as variáveis com maior necessidade de monitoria para definir o risco de doença cardíaca. Após análise, descobriu-se que as variáveis mais influentes na pesquisa foram: frequência cardíaca máxima atingida sob stress físico (thalach), dor no peito (cp), nível de colesterol (chol) e talassemia (thal).

Com base no feedback dos profissionais de saúde, ficou evidente que o uso dos gráficos foi crucial para aprimorar a compreensão dos dados. A representação visual facilitou a identificação das correlações entre as variáveis, permitindo uma análise mais precisa e objetiva. O desenvolvimento do modelo persiste sendo incentivado, principalmente dado o desempenho do novo modelo binário.

3. RESULTADOS PARCIAIS

O projeto iniciou com uma série de visitas à comunidade do trabalho, a [redacted], onde foi possível obter insights valiosos sobre a viabilidade do trabalho e os desafios a serem enfrentados no desenvolvimento, com objetivo final de criar uma ferramenta de utilidade ampla na predição do grau de doença cardíaca. Um dos membros do grupo já possuía contato prévio com o posto de saúde local, o que facilitou a aproximação com os profissionais que seriam responsáveis por validar o modelo e confirmar sua utilidade prática. Durante as visitas, realizou-se entrevistas semi estruturadas, utilizando formulários prontos e também diálogo livre. As entrevistas foram conduzidas de forma a permitir que os participantes compartilhassem suas experiências e dificuldades no manuseio dos dados clínicos, na tentativa de receber informação suficiente que justificasse o desenvolvimento do projeto. Algumas perguntas comuns realizadas nas entrevistas foram:

- Como é feito o acompanhamento de pacientes com fatores de risco para doenças cardíacas na sua unidade?
- Quais informações clínicas dos pacientes são coletadas atualmente para monitorar o risco de doenças cardíacas?
- Você acredita que os dados coletados atualmente são suficientes para prever o risco de doenças cardíacas de forma eficaz? Se não, que informações adicionais seriam úteis?
- Existem atualmente ferramentas tecnológicas que auxiliam na análise preditiva de doenças na sua unidade? Se sim, quais são elas e como são usadas?

Através dessa abordagem, foi possível levantar pontos importantes sobre o manejo dos dados. Constou-se não somente uma dificuldade na organização dos dados para tomada de decisão como também uma falta de ferramentas eficientes que possam integrar informações dos pacientes com facilidade. Além disso, algumas variáveis importantes na predição do risco de doença cardíaca, como resultados de exames complexos, não estavam no escopo de coleta do posto de saúde. Essa coleta de informações foi crucial para alinhar o projeto com as necessidades reais e fornecer a base para o desenvolvimento de uma ferramenta que poderia

oferecer previsões baseadas em dados e auxiliar na tomada de decisões clínicas mais informadas.

A escolha do dataset do projeto, o Heart Disease Dataset da Universidade da Califórnia – Irvine (UCI), foi motivada por ser amplamente utilizado na pesquisa científica e principalmente por utilizar algumas variáveis comuns às coletas de dados realizados no posto. Foi então cuidadosamente planejado para transformar o dataset em uma ferramenta que fosse capaz de generalizar as instâncias de dados para outros conjuntos. Apesar de a classificação de risco de doenças cardíacas não ser o foco principal do atendimento primário prestado pelo posto, os profissionais demonstraram grande interesse no potencial da solução proposta. Eles viram a oportunidade de utilizar uma ferramenta de análise preditiva para facilitar a gestão de riscos e melhorar o atendimento de pacientes com condições crônicas. Com base nesse feedback inicial e o apoio dos profissionais para determinar os fatores de risco importantes para nossa análise, foi possível definir melhor o escopo do projeto e identificar os requisitos fundamentais para o desenvolvimento da solução.

Para iniciarmos o projeto utilizando a metodologia KDD, primeiro selecionamos os dados. Como já citado anteriormente, usaremos o conjunto de dados Heart Disease Dataset da UCI. O dataset da UCI inclui dados de várias fontes de pesquisa, como Cleveland, Hungary, Switzerland e Long Beach. Para o projeto, nós utilizaremos somente o conjunto de dados de Cleveland, o mais utilizado em pesquisas e estudos devido a alguns fatores como a qualidade da coleta de dados, a validação científica e o foco nas variáveis relevantes.

Para dar início ao desenvolvimento do nosso modelo, precisamos carregar o conjunto de dados. Isso pode ser feito utilizando bibliotecas como pandas para ler arquivos CSV ou Excel, ou diretamente através de APIs de dados. Pandas é uma biblioteca python para manipulação e análise de dados. Pandas oferece estruturas de dados chamados Dataframes, eficientes e flexíveis para trabalhar com dados tabulares, como planilhas ou bases de dados. Com o dataset carregado vamos processá-lo para que fique no formato ideal para os algoritmos de aprendizado de máquina e modelos de visualização de dados.

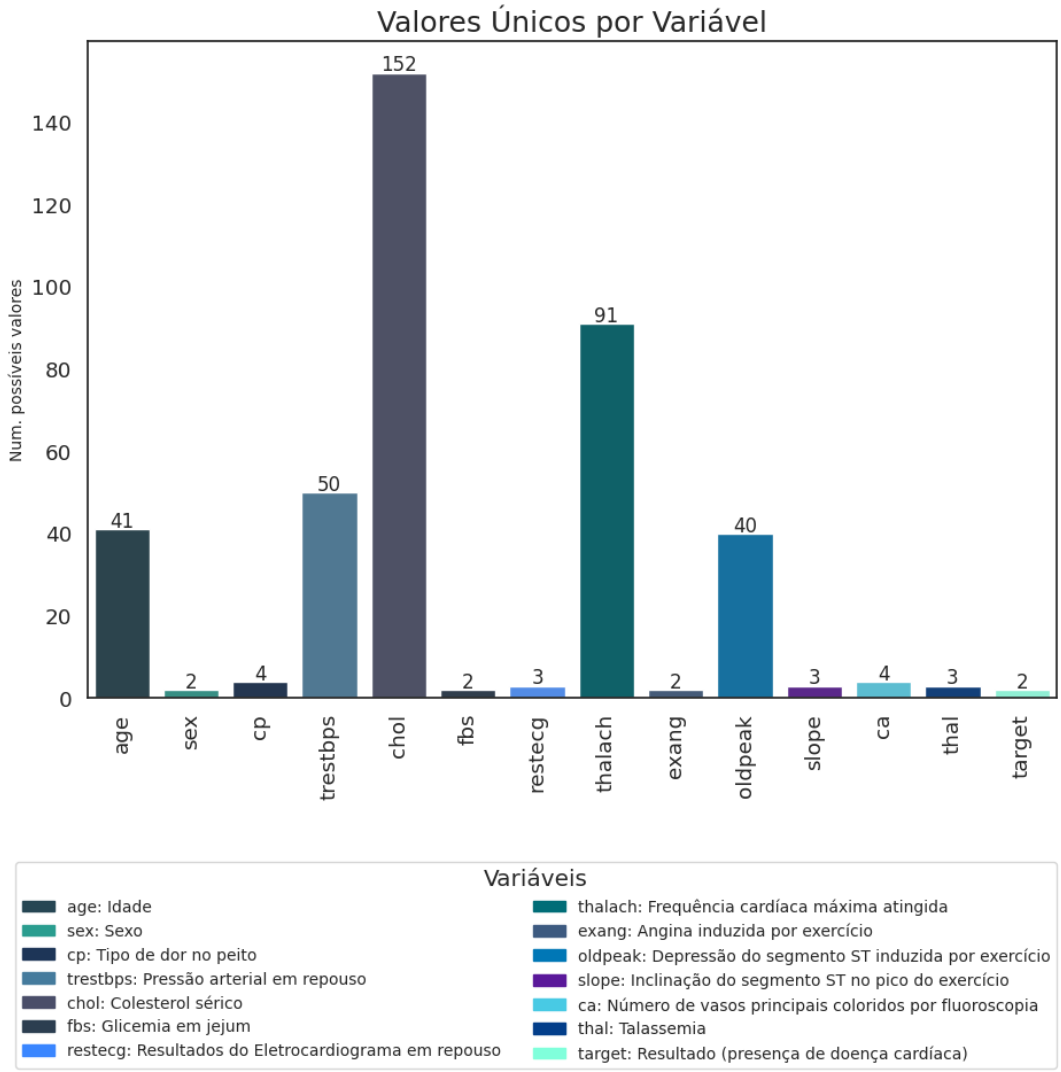
Primeiramente, vamos separar features e label no dataset. As features, variáveis independentes, são os atributos que o modelo usará para fazer previsões, enquanto a label, variável dependente, é o valor que o modelo tentará prever — que pode ser explicado pelas features no conjunto de dados. No caso do problema de classificação de doenças cardíacas, as

features são fatores como idade, colesterol, pressão arterial, entre outros. Já a label no dataset é a variável num que é um número de 0 a 4 indicando a gravidade da doença cardíaca, sendo que 0 é a ausência dessa.

Em seguida realizaremos uma análise exploratória dos dados do conjunto, para podermos compreender melhor a estrutura geral dos dados, a descrição das variáveis e identificar as correlações entre os atributos disponíveis. Foi necessário realizar pouco pré-processamento, uma vez que o dataset de Cleveland já é amplamente conhecido por ser um dos mais completos e bem tratados na literatura. A limpeza básica, como remoção de valores faltantes ou dados inconsistentes, já foi previamente realizada, permitindo que pudéssemos focar diretamente na exploração das variáveis e menos nas etapas preliminares de tratamento.

Primeiro, vamos realizar a contagem de elementos únicos presentes em cada coluna do dataset. Utilizando um gráfico de barras, podemos visualizar rapidamente a variabilidade dos dados em cada atributo. As barras são coloridas de acordo com a contagem de valores únicos, destacando colunas com maior diversidade em comparação às com menor variabilidade. Esse tipo de análise permite rapidamente identificar o tipo de dado de cada variável, fornecendo informação valiosa de como tratá-la. Pode-se visualizar no gráfico que a variável “sex” tem somente dois valores únicos, homem e mulher. Já a variável “chol” tem muitos valores únicos, o que indica uma variável distribuída em um intervalo contínuo.

Figura 1 - Gráfico de Valores Únicos por Variável

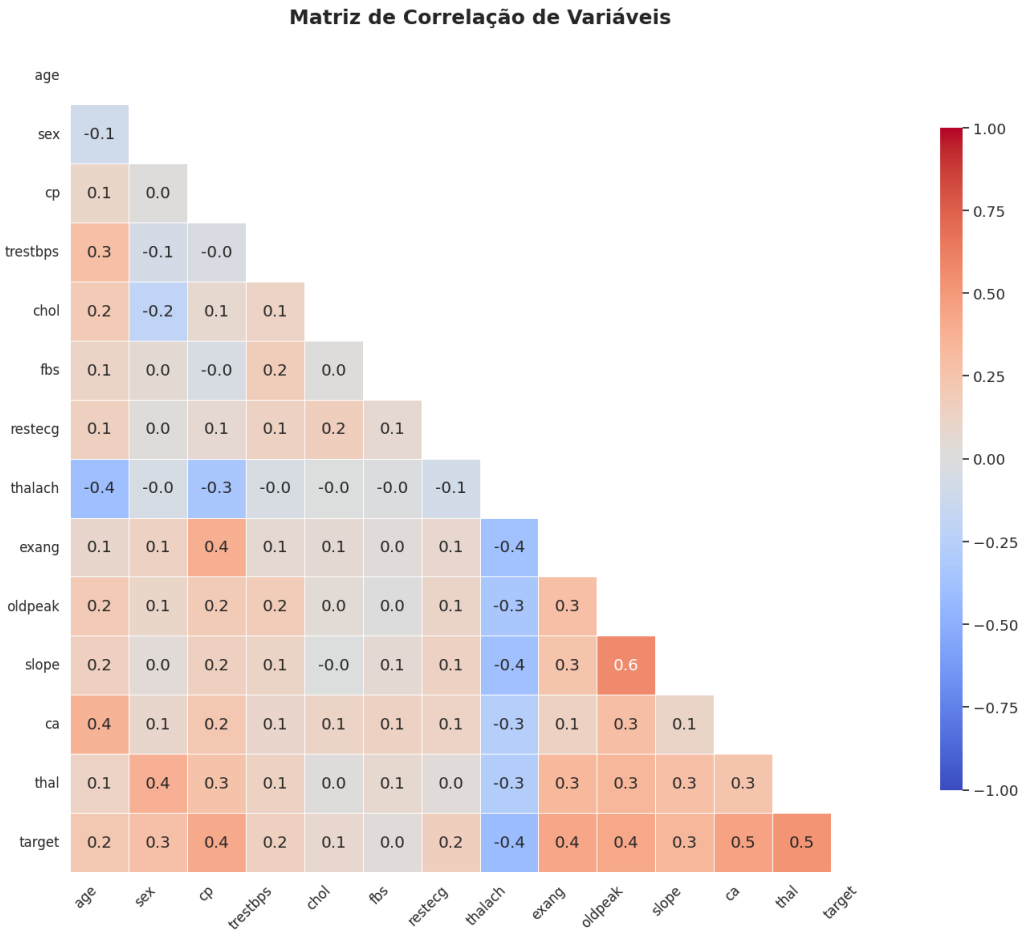


Fonte: Elaborada pelo autor, 2024.

A análise inicial de correlação entre as variáveis demonstrou algumas informações cruciais para o entendimento geral dos dados. Foi realizado uma análise com gráfico de correlação para compreender o comportamento entre variáveis. A correlação é uma medida estatística que indica a força e a direção de uma relação linear entre duas variáveis, sendo que este varia de -1 até 1. As variáveis que apresentaram maior correlação com o desfecho de doenças cardíacas incluíram:

- Idade (age): Pacientes mais velhos tendem a ter maior risco de desenvolver doenças cardíacas, conforme esperado.
- Colesterol sérico (chol): Níveis elevados de colesterol demonstraram uma correlação positiva com o risco de doenças cardíacas.
- Frequência cardíaca máxima no teste de stress (thalach): Uma menor frequência cardíaca máxima alcançada durante exercícios físicos foi associada a um maior risco de problemas cardíacos.
- Tipo de dor no peito (cp): A dor no peito de tipo apresentou alta correlação com a presença de doenças cardíacas.
- Angina induzida por exercício (exang): Pacientes que relataram angina induzida por exercícios físicos apresentaram maior probabilidade de doenças cardíacas.

Figura 2- Gráfico de Correlação das Variáveis

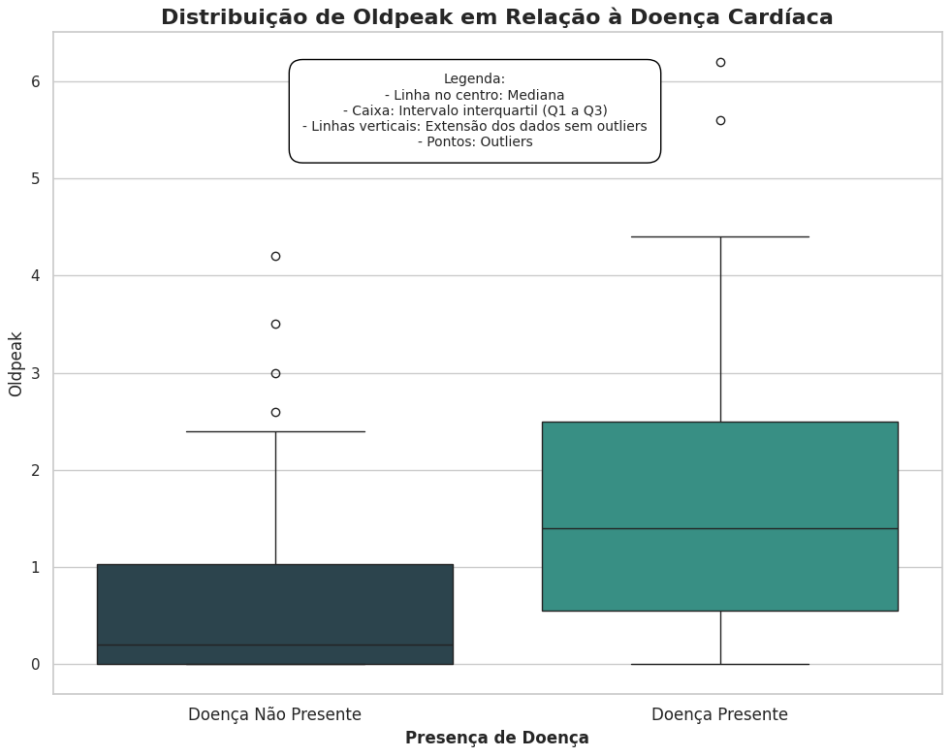


Fonte: Elaborado pelo autor, 2024.

Realizado uma análise inicial para explicar as variáveis contidas no conjunto, foi possível observar alguns padrões e características interessantes, como por exemplo:

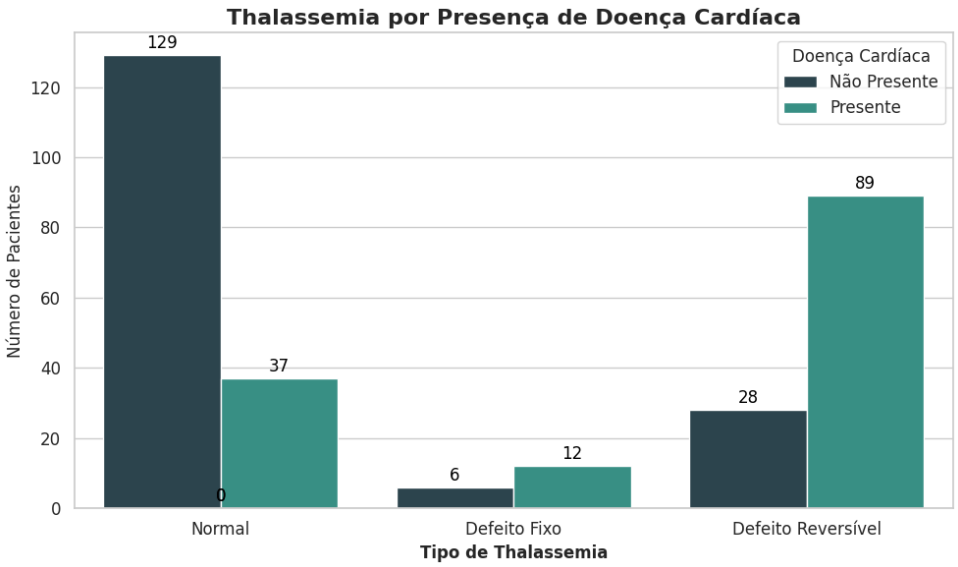
- As variáveis *thal* (anormalidade talâmica) e *oldpeak* (depressão do segmento ST induzida por exercício) foram identificados como características de alta influência na previsão de doenças cardíacas. Houve um maior número de instâncias com doença cardíaca com talassemia do tipo reversível e valores maiores de *oldpeak* também indicam uma maior presença de doença cardíaca. Para explicar visualmente a variável *oldpeak* utilizamos um gráfico de boxplot, uma ferramenta gráfica que resume a distribuição de um conjunto de dados e facilita a visualização de cinco estatísticas principais: o mínimo, o primeiro quartil, a mediana, o terceiro quartil e o máximo, além de destacar possíveis outliers, ou seja, valores atípicos.
- A distribuição de gênero no dataset mostrou que, enquanto homens têm maior prevalência de doenças cardíacas, mulheres tendem a ser diagnosticadas com a condição em estágios mais avançados de idade.
- Há uma maior prevalência de doenças cardíacas em pacientes com dor no peito classificada como assintomática Isso sugere que, embora esses pacientes não apresentem sintomas clássicos de dor no peito, eles estão em um grupo de maior risco para doenças cardíacas. A presença de uma condição assintomática pode, portanto, ser um indicador crucial para a detecção precoce da doença, sendo uma variável importante de se monitorar.
- É possível dizer, através dos dados, que pacientes com uma idade mais avançada são mais suscetíveis à presença de doença cardíaca, o que já era esperado na análise inicial.

Figura 3- Gráfico Depressão Segmento ST por Presença de Doença



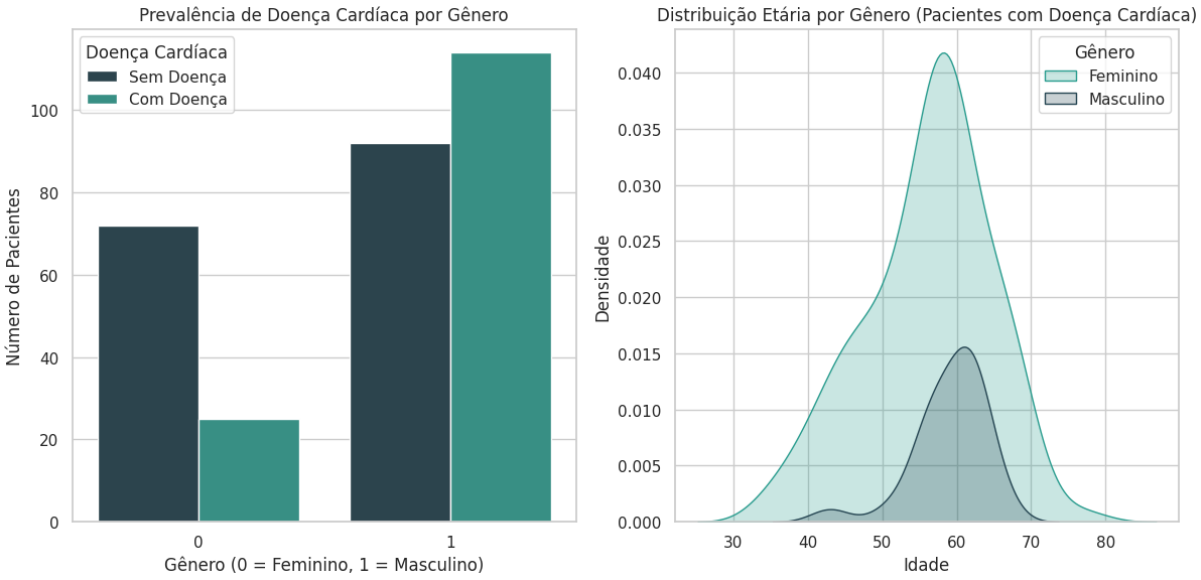
Fonte: Elaborado pelo autor, 2024.

Figura 4- Gráfico Tipo de Talassemia por Presença de Doença



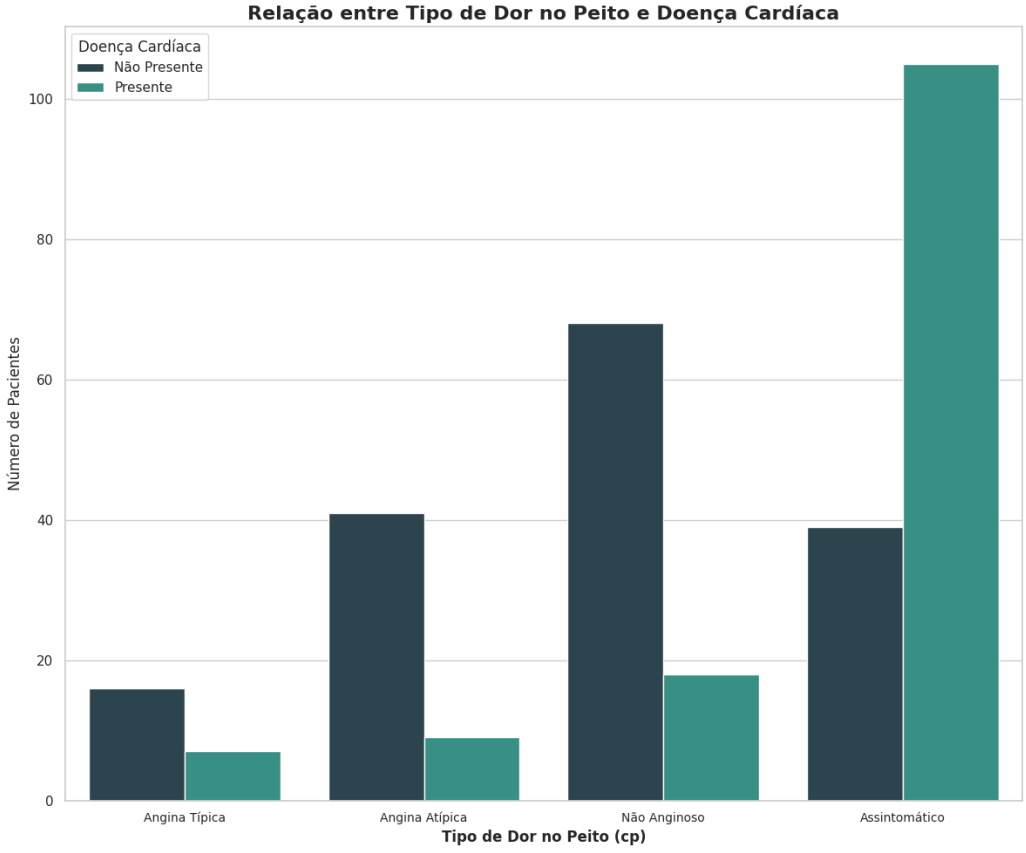
Fonte: Elaborado pelo autor, 2024.

Figura 5- Gráfico Doença x Gênero e Distribuição Etária por Gênero para Doença Positiva



Fonte: Elaborado pelo autor, 2024.

Figura 6- Gráfico Presença de Doença por Tipo de Dor no Peito



Fonte: Elaborado pelo autor, 2024.

Com essa análise visual, é possível interpretar os dados de forma mais eficiente, permitindo identificar padrões ocultos e correlações importantes que facilitam a compreensão do conjunto de dados. A auxilia na detecção de anomalias e na tomada de decisões informadas, tornando o processo de análise mais ágil e acessível. Com mais conhecimento em mãos, partimos para o desenvolvimento do modelo.

Para o melhor desempenho do modelo, realizaremos duas transformações nos dados: One Hot Encoding e Normalização. Se o dataset contém variáveis categóricas (por exemplo, sexo, tipo de dor no peito), elas precisam ser convertidas em uma forma que o modelo possa entender. One Hot Encoding é uma técnica de pré-processamento de dados usada para converter variáveis categóricas em um formato numérico que possa ser utilizado por algoritmos de aprendizado de máquina. Essa técnica transforma categorias em um formato binário, onde cada categoria única é representada por um vetor com 1 em uma posição e 0 nas outras posições. Essa transformação evita que os algoritmos tratem variáveis categóricas como variáveis ordinais, prevenindo inferências erradas e preservando a natureza qualitativa das variáveis categóricas. No conjunto de dados nós temos a variável cp (dor no peito), que é qualitativa e possui 4 valores possíveis: 1. angina típica, 2. angina atípica, 3. dor não anginosa, 4. assintomático. Utilizando o one hot encoding, teríamos:

Tabela 1 – Exemplo de One Hot Encoding

Angina Típica	Angina Atípica	Dor Não Anginosa	Assintomático
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Fonte: Autor, 2024

Já a normalização é um passo importante para redes neurais, especialmente quando as features possuem escalas diferentes. A normalização de dados é o processo de ajustar os valores de diferentes variáveis para uma escala comum, sem distorcer as diferenças nos intervalos de valores, evitando que valores muito fora da curva, em escala diferente por exemplo, interfiram no aprendizado do modelo. A normalização transforma os dados para que

fiquem dentro de uma faixa, no nosso caso de-1 e 1. Essa técnica de transformação garante que nenhuma variável domine as outras em termos de magnitude.

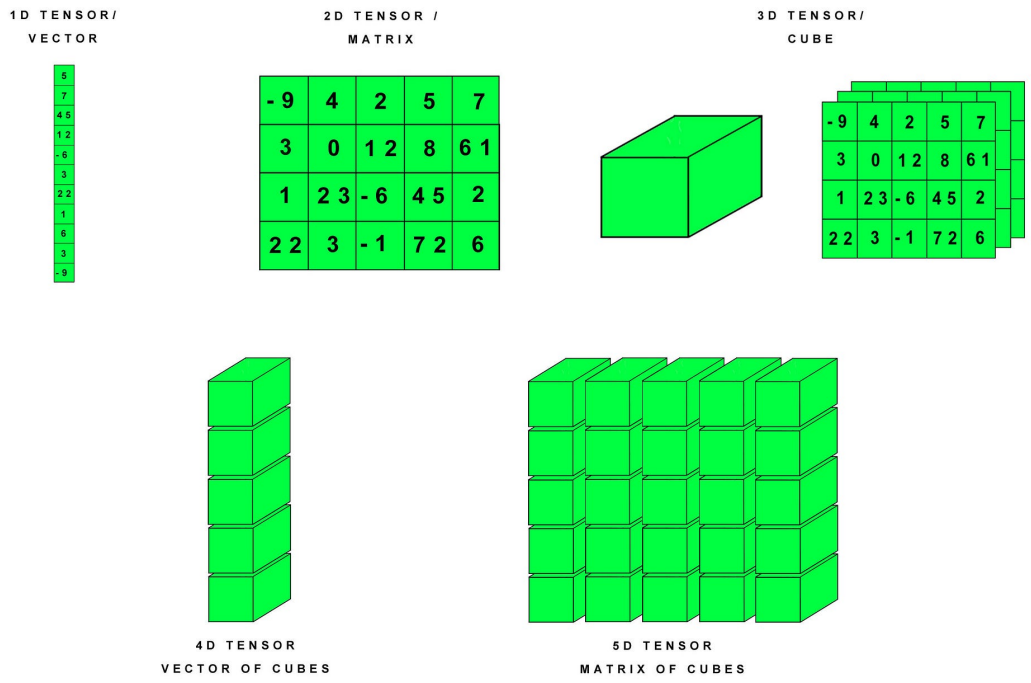
Ambas as transformações foram realizadas com a biblioteca Scikit-learn. Scikit-learn é uma biblioteca de código aberto em Python amplamente utilizada para aprendizado de máquina e análise de dados. Ela é capaz de fornecer diversas ferramentas para trabalhar com aprendizado de máquina, seja algoritmos de aprendizagem de máquina ou ferramentas de pré-processamento. O scikit-learn é muito usado por sua simplicidade de uso e integração com outras bibliotecas populares, como NumPy e o próprio Pandas, sendo ideal para prototipagem rápida em projetos de aprendizagem de máquina.

Em seguida, vamos dividir o nosso conjunto de dados em conjuntos de treino, teste e validação. A separação dos dados em conjuntos de treino, teste e validação é uma prática essencial em aprendizado de máquina para garantir que os modelos sejam treinados e avaliados de forma eficaz e generalizável, capaz de prever rótulos para dados não vistos no treinamento. Essa separação foi feita também utilizando o scikit-learn.

- O conjunto de treino é utilizado para ajustar os parâmetros do modelo, permitindo que ele aprenda padrões com os dados. Esse é o maior subconjunto e tem a função de prover para o modelo uma base sólida suficiente para entender as relações entre as variáveis de entrada e saída. Como usaremos redes neurais, os parâmetros do modelo são os pesos das conexões sinápticas, responsáveis por controlarem a forma como os sinais são transmitidos e influenciam o comportamento do modelo.
- O conjunto de validação é usado durante o treinamento para afinar os hiperparâmetros do modelo. Como estamos usando redes neurais, estamos lidando com hiperparâmetros como: quantidade de camadas da rede, quantidade de neurônios por camada, taxa de aprendizado, entre outros.
- O conjunto de testes é utilizado para avaliar o desempenho do modelo pós treinamento. Ele contém dados que o modelo não viu durante o processo de treino, permitindo medir a capacidade de generalização do modelo em dados não antes observados. Isso ajuda a identificar se o modelo está sofrendo de overfitting, está memorizando os dados de treino em vez de aprender padrões.

Em seguida, para obter cálculos mais eficientes com nosso modelo, convertemos nossos dados para tensores utilizando a biblioteca Keras. Keras é uma biblioteca de alto nível para aprendizado de máquina que funciona como uma interface amigável para o Tensor Flow, que é uma biblioteca de código aberto fundamental para trabalhar com tensores. Tensores são o bloco fundamental de dados. Eles são uma estrutura de dados que generaliza matrizes e vetores para várias dimensões. Um tensor pode ser pensado como um contêiner multidimensional de números, vantagem de serem otimizados para cálculos em GPU, o que é extremamente útil no treinamento de modelos grandes e complexos.

Figura 7- Tensores Multidimensionais



Fonte: DataCamp, 2018.

Finalmente, vamos iniciar o processo de construção da rede neural do projeto. Uma Rede Neural Artificial (RNA) é um modelo computacional inspirado no funcionamento do cérebro humano, projetado principalmente para reconhecer padrões e aprender com dados.

Consiste fundamentalmente de várias unidades interconectadas chamadas neurônios, responsáveis por processar informação de maneira similar a neurônios biológicos encontrados no corpo humano. Elas são a base do aprendizado profundo, onde modelos complexos com várias camadas aprendem representações profundas de dados para resolver problemas de alta complexidade. Vale ressaltar que somente é considerado aprendizado profundo se uma RNA tiver pelo menos duas camadas ocultas, sendo essas camadas que contêm neurônios responsáveis por processar os dados de entrada.

O neurônio artificial, também conhecido como neurônio MCP (McCulloch e Pitts, 1943) é a unidade básica de uma rede neural. Cada neurônio recebe uma ou mais entradas, dados ou saídas de outros neurônios em uma rede, as quais são ponderadas por pesos. As entradas ponderadas são somadas e opcionalmente um termo de bias é adicionado à soma. O valor resultante passa por uma função de ativação, que determina se o neurônio será ativado. A saída pode então ser enviada para outros neurônios nas camadas seguintes, permitindo o fluxo de informações pela rede. A rede neural aprende a resolver problemas complexos ajustando os pesos das conexões entre os neurônios. Esses pesos determinam a importância de cada entrada em relação à saída esperada, e seu ajuste, durante o treinamento, permite que a rede identifique padrões e tome decisões com base nos dados.

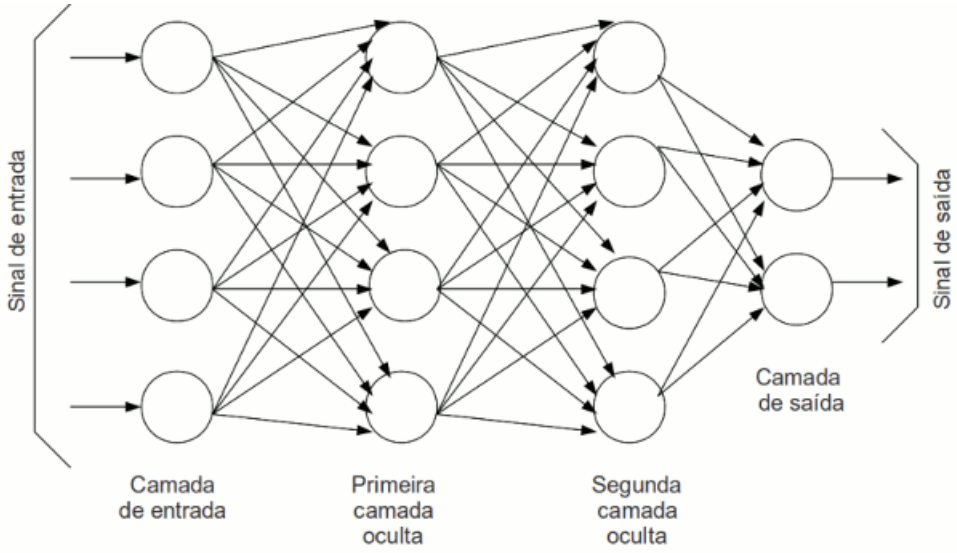
O neurônio artificial MCP, realiza uma operação linear sobre as entradas, pesos multiplicados pelas entradas, e aplica uma função de ativação para gerar a saída. Na prática, um único neurônio é capaz de modelar funções lineares: ele define uma reta em um plano, separando os dados em duas regiões. Para resolver problemas mais complexos e não lineares, redes neurais são compostas por camadas ocultas de neurônios com funções de ativação não lineares. Dessa forma, a combinação dessas camadas permite que a rede aprenda fronteiras de decisão não lineares, podendo modelar relações complexas entre as variáveis de entrada e produzir soluções mais sofisticadas.

Essencialmente, uma rede neural é composta por várias camadas de neurônios, onde cada camada é conectada à próxima. Essas camadas podem ser definidas por:

- Camada de Entrada: É a camada que recebe os dados, cada neurônio da camada representa uma característica ou atributo do dado. No nosso caso, nossa camada de entrada teria 13 neurônios, um para cada atributo diferente.

- Camadas Ocultas: Camadas intermediárias entre a camada de entrada e a camada de saída. Contém neurônios com funções não lineares, responsáveis por modelar relações complexas.
- Camada de Saída: Camada que gera a classificação com base no processamento interno da rede. O número de neurônios de saída depende da tarefa, no nosso caso são 5 pois temos uma variável alvo que varia de 0 a 4.

Figura 8- Arquitetura de Rede Neural Unidirecional



Fonte: Monolito Nimbus, 2017

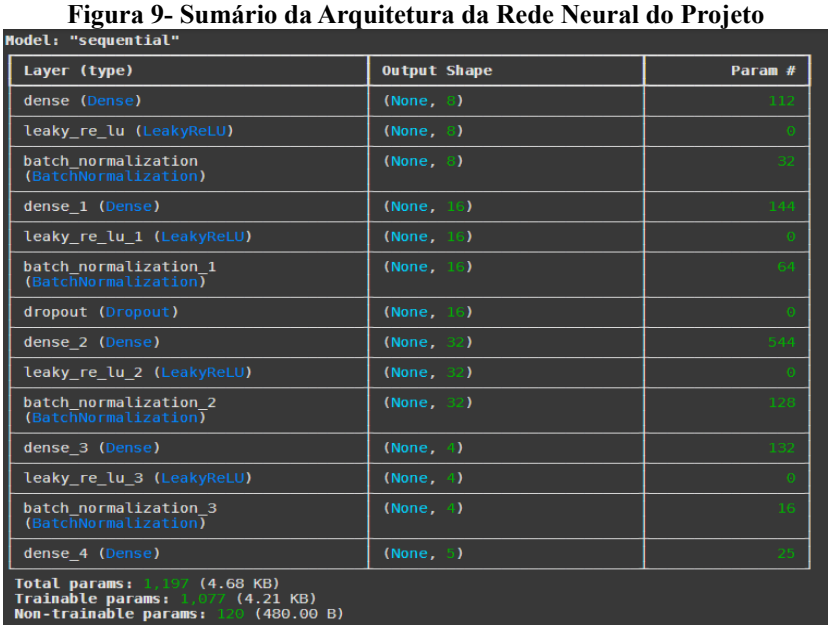
A rede neural do projeto tem a seguinte arquitetura:

- Camada de entrada: Ela aceita 13 entradas (as características de entrada do seu dataset) e gera 8 saídas. A camada aprende 8 conjuntos de pesos, um para cada entrada, que será ajustado durante o treinamento.
- A segunda camada conecta 8 neurônios da primeira camada a 16 neurônios. Aqui, a rede está expandindo a capacidade de modelagem, permitindo que ela capture mais complexidade nos dados.
- Entre a segunda e a terceira camada há camada de Dropout. Dropout é uma técnica de regularização para diminuir o sobreajuste do modelo aos dados.

- A terceira camada conecta 16 neurônios da segunda camada a 32 neurônios. Essa camada adicional permite à rede aprender representações ainda mais complexas dos dados.
- Esta é a camada de saída, que conecta 32 neurônios da terceira camada a 5 neurônios. Cada neurônio na camada de saída representa um rótulo, classe, que a rede pode prever.

A função de ativação na camada oculta é a Leaky ReLU. A Leaky ReLU é uma função não linear que decide como os neurônios de uma rede neural reagem a diferentes entradas de dados. Ela é uma variação da função ReLU. A função ReLU (Rectified Linear Unit) é uma função de ativação que transforma valores negativos em zero e mantém os valores positivos inalterados. Por definir os valores negativos como zero, a ReLU cria uma ativação esparsa, o que significa que menos neurônios são ativados simultaneamente, resultando em uma melhora na eficiência do modelo, resultando em uma melhor generalização em tarefas de classificação. A ReLU pode sofrer do problema de “neurônios mortos”, onde unidades podem ficar permanentemente inativas se o gradiente se torna zero. Já a função Leaky ReLU mitiga o problema permitindo que os valores negativos tenham um pequeno gradiente, em vez de serem simplesmente zero. Isso ajuda a manter as unidades ativas, mesmo que os valores de entrada sejam negativos, melhorando a capacidade de aprendizado da rede. Quando recebidos por um neurônio forem negativos, a função Leaky ReLU permitirá passagem somente de parte do sinal, enquanto a ReLU padrão cessaria o sinal por completo.

Figura 9- Sumário da Arquitetura da Rede Neural do Projeto



Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 8)	112
leaky_re_lu (LeakyReLU)	(None, 8)	0
batch_normalization (BatchNormalization)	(None, 8)	32
dense_1 (Dense)	(None, 16)	144
leaky_re_lu_1 (LeakyReLU)	(None, 16)	0
batch_normalization_1 (BatchNormalization)	(None, 16)	64
dropout (Dropout)	(None, 16)	0
dense_2 (Dense)	(None, 32)	544
leaky_re_lu_2 (LeakyReLU)	(None, 32)	0
batch_normalization_2 (BatchNormalization)	(None, 32)	128
dense_3 (Dense)	(None, 4)	132
leaky_re_lu_3 (LeakyReLU)	(None, 4)	0
batch_normalization_3 (BatchNormalization)	(None, 4)	16
dense_4 (Dense)	(None, 5)	25

Total params: 1,197 (4.68 KB)
Trainable params: 1,077 (4.21 KB)
Non-trainable params: 120 (480.00 B)

Fonte: Elaborado pelo autor, 2024.

Compõem também na configuração da nossa rede a definição da função de perda. A função de perda quantifica a diferença entre as previsões do modelo e os valores reais nos rótulos dos dados. Essencialmente é uma medida capaz de quantificar o erro cometido pelo modelo e o objetivo durante o treinamento é minimizar essa função, ou seja, fazer com que o modelo se torne o mais preciso possível. Duas funções de perda bem comuns em modelos de rede neural são:

- Erro Quadrático Médio (MSE): Usada em problemas de regressão, calcula a média dos quadrados das diferenças entre as previsões e os valores reais.
- Entropia Cruzada: Utilizada para problemas de classificação múltipla. A função de perda mede a diferença entre as previsões do modelo e os rótulos verdadeiros. É mais sensível a erros em classes menos frequentes, tornando-a uma escolha popular para problemas de classificação. Entropia Cruzada é a função de perda escolhida para nossa rede neural, já que estamos trabalhando com um problema de classificação múltipla.

Para atualizar os pesos em uma rede neural profunda, utilizamos o algoritmo de retropropagação, o principal algoritmo utilizado para o treinamento de redes neurais. Ele

funciona em duas etapas principais: a passagem para frente (forward pass) e a passagem para trás (backward pass). Na passagem para frente, os dados de entrada percorrem a rede, camada por camada, até gerar uma saída na camada final. A saída é comparada com o valor real (rótulo) usando a função de perda, entropia cruzada no nosso caso.

O erro é medido com base na diferença entre a predição da rede e o valor real. Este erro é então propagado de volta através da rede para ajustar os pesos. Esse passo é chamado de passagem para trás, onde a retropropagação começa pela última camada e caminha para trás, até a primeira camada oculta.

O algoritmo calcula os gradientes da função de perda em relação a cada peso da rede usando a regra da cadeia do cálculo diferencial. Isso é feito por meio da derivada parcial da função de perda em relação às ativações e pesos de cada camada. Após calcular os gradientes, esses são utilizados para atualizar os pesos da rede para reduzir o erro. Isso é feito através de um algoritmo de otimização, sendo o mais comum deles o Stochastic Gradient Descent (SGD). O algoritmo de otimização é o recurso utilizado para ajustar os pesos de acordo com o erro calculado pela função de custo, sendo ele o algoritmo capaz de minimizar o erro e melhorar a precisão da rede. Esse processo então é repetido de maneira iterativa, onde cada nova repetição é chamada de época.

Para nossa rede neural, utilizaremos o algoritmo de otimização Adam (2014). Adam significa Adaptive Moment Estimation, é um algoritmo de otimização moderno e eficiente. Ele ajusta a taxa de aprendizado para cada parâmetro individual, acumulando a média dos gradientes e a média dos quadrados dos gradientes em cada iteração. Isso permite que o Adam possa adaptar dinamicamente a taxa de aprendizado de acordo com o histórico de atualizações, resultando em uma convergência mais rápida e estável

O algoritmo Adam (Adaptive Moment Estimation) é um método de otimização popular para redes neurais que combina as vantagens de duas técnicas anteriores: o Momentum e o RMSprop. Ele ajusta a taxa de aprendizado para cada parâmetro individualmente, acumulando a média dos gradientes e a média dos quadrados dos gradientes em cada iteração. Isso permite que o Adam adapte dinamicamente a taxa de aprendizado com base no histórico das atualizações, resultando em uma convergência mais rápida e estável. Introduzido em 2014, o Adam é amplamente utilizado por sua eficiência em grandes datasets

e alta dimensionalidade. Com isso definido, finalizamos a construção da arquitetura da nossa rede neural de classificação.

Para melhorar ainda mais a capacidade de generalização do modelo, aplicamos nas camadas da rede neural as técnicas de regularização L2 e dropout. A regularização L2, também chamada de penalidade de peso, é uma técnica que adiciona uma penalidade ao valor dos pesos das conexões na função de perda do modelo. Na prática, a regularização L2 incentiva a rede a manter pesos menos, reduzindo a complexidade do modelo e auxiliando a evitar que o modelo memorize os dados.

Já o dropout é uma técnica de regularização onde durante o treinamento, neurônios da rede são aleatoriamente desligados. As conexões e a saída desses neurônios não são atualizadas temporariamente, o que impede que o modelo dependa de maneira excessiva de neurônios específicos. O dropout força o modelo a aprender representações mais generalizadas dos dados, forçando sua adaptação e reduzindo o sobreajuste.

Para aprofundar e aumentar a robustez da análise, desenvolvemos também um segundo modelo preditivo, com o objetivo de realizar uma classificação binária . O enfoque desse modelo é abordar exclusivamente o problema de classificar a presença ou não de doença cardíaca. Para uso desse modelo, transformamos os rótulos dos dados em valores binários, sendo que 1 representa presença de doença cardíaca e 0 indica sua ausência. O objetivo desse modelo é aprofundar a análise abordando um problema mais simples, servindo de objeto de comparação e validando a acurácia em um contexto de diagnóstico direto. Esse modelo binário utilizou a função sigmoide como função de ativação na camada de saída. A função sigmoide é frequentemente utilizada em modelos de classificação binária por converter qualquer valor de entrada em uma probabilidade, entre 0 e 1. Esse novo modelo também substitui a Entropia Cruzada pela Entropia Cruzada Binária, uma função de perda para problemas de classificação mais adequada para nosso novo conjunto de dados.

Sumarizando, a rede neural desenvolvida no projeto é uma arquitetura de aprendizado profundo supervisionado, com capacidade para classificar o risco de doenças cardíacas utilizando os dados do conjunto Heart Disease Dataset, da universidade da Califórnia - Irvine. Para o otimizar o modelo, os dados no conjunto foram normalizados para eliminar problemas com escala e aplicamos o one hot encoding na variável alvo, para que essa seja tratada adequadamente como uma variável categórica. Dividimos o conjunto de dados em treino,

teste e validação, garantindo que tenhamos dados suficientes para a rede aprender e validar seus dados, evitando o sobreajuste. A arquitetura do modelo é composta de múltiplas camadas ocultas com com funções de ativação Leaky ReLU, uma variação da função ReLU que impede o fenômeno dos “neurônios mortos”, otimizando a classificação. A função de perda da rede é a entropia cruzada, padrão em problemas de classificação, e a função de otimizadora é a Adam, uma evolução do algoritmo de retropropagação utilizado para treinar redes de entrada-saída com múltiplas camadas. A arquitetura foi validada com o método o conjunto de validação, resultando em um modelo capaz de realizar previsões robustas para auxiliar em diagnósticos médicos. Criamos também um segundo modelo com uma rede neural simplificada para classificação binária, focada exclusivamente em detectar a presença ou ausência de doença cardíaca. Nesse modelo, transformamos os rótulos dos dados em valores binários e utilizamos a função sigmoide como função de ativação na camada de saída, já que ela converte qualquer valor de entrada em uma probabilidade entre 0 e 1. Também usamos a função de perda Entropia Cruzada Binária, mais adequada para o novo modelo, criando assim um modelo de classificação binária capaz de predizer a presença de doenças cardíacas com alta precisão.

O dashboard do projeto foi desenvolvido utilizando a linguagem Python, com o auxílio das bibliotecas Plotly e Dash, amplamente empregadas para a criação de visualizações de dados interativos e interfaces de usuário. Plotly é uma biblioteca que possibilita a construção de gráficos dinâmicos e de alta qualidade, oferecendo suporte para diversos tipos de visualizações fundamentais para a análise exploratória e interpretação dos dados. Dash fornece uma estrutura para criar dashboards interativos de forma eficiente e com um design personalizável, além de possibilitar a criação de um layout modular, configurado em linhas e colunas para organizar os gráficos de acordo com sua importância e tornar a interface mais clara e objetiva. A integração de ambas as bibliotecas no dashboard proporcionou uma ferramenta interativa que atende às necessidades do projeto, oferecendo uma visualização completa e personalizada dos dados para apoiar a análise e tomada de decisão.

4. CONSIDERAÇÕES FINAIS

O projeto integrador visou o desenvolvimento de um modelo preditivo de classificação de risco para doenças cardíacas, utilizando dados clínicos do dataset Heart Disease da UCI e algoritmos de aprendizado de máquina. Os resultados do trabalho foram relevantes tanto em termos de precisão do modelo quanto na usabilidade e clareza do painel interativo, que permite uma análise mais aprofundada dos fatores de risco envolvidos. O uso de visualizações interativas contribuiu para o entendimento de como certas variáveis, como idade, frequência cardíaca máxima, talassemia, entre outras, impactam o risco de desenvolvimento de doenças cardíacas.

Ao longo do projeto, foram observadas tanto contribuições como limitações. Em termos de contribuições, a solução criada se destacou pela aplicabilidade no apoio à tomada de decisão clínica, uma vez que o modelo demonstrou alta eficácia na categorização binária de doença cardíaca e eficácia moderada na classificação do grau de tal doença. Além disso, a visualização interativa dos dados possibilita que profissionais envolvidos compreendam e explorem de forma prática as tendências e correlações presentes nos dados.

Contudo, entre as limitações, destaca-se a necessidade de uma base de dados maior para melhorar a robustez e a generalização do modelo preditivo. A classificação do risco de doença cardíaca teve precisão de aproximadamente 65%, valor abaixo do esperado durante o desenvolvimento. Isso ocorre pela complexidade dos dados no conjunto, necessitando de mais amostras válidas para aumentar a precisão do modelo. Felizmente a classificação binária somente, que indica presença ou ausência de doença cardíaca, atingiu uma marca de aproximadamente 80% de precisão, sendo válidas do como um modelo robusto e eficiente no âmbito do projeto.

Para garantir a integridade dos dados e manter o foco do trabalho, optamos por não utilizar técnicas de data augmentation para gerar novos conjuntos de amostras sintéticas. A decisão foi baseada na prioridade de desenvolver uma arquitetura de rede neural suficientemente robusta, capaz de lidar com dados reais sem a necessidade de expansão artificial do conjunto de treinamento. Buscamos aumentar a eficácia do modelo em contextos práticos, onde ele será aplicado a dados genuínos, refletindo condições reais.

Por fim, acredita-se que este projeto traz uma contribuição significativa para os profissionais da área da saúde, oferecendo uma ferramenta que apoia a prevenção e o acompanhamento de doenças cardíacas. Em termos de impacto social, o desenvolvimento desta solução representa um avanço no uso de tecnologias acessíveis e intuitivas em saúde, permitindo o acesso a informações e previsões mais claras, possibilitando uma abordagem preventiva e informada no tratamento de seus pacientes. Assim, o projeto cumpre seus objetivos iniciais e, apesar dos desafios encontrados, estabelece uma base sólida para futuras melhorias.

5. BIBLIOGRAFIA

ALFADLI, Khadijah Mohammad; ALMAGRABI, Alaa Omran. Feature-Limited Prediction on the UCI Heart Disease Dataset. Computers, Materials & Continua, v. 74, n. 3, 2023.

FLECK, Leandro et al. Redes neurais artificiais: Princípios básicos. Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47-57, 2016.
SANTOS, William Hamilton dos. Estudo da base de dados abertos E-Saúde da prefeitura de Curitiba usando técnicas de mineração de dados. 2018.

MUKHERJEE, Soumonos; SHARMA, Anshul. Intelligent heart disease prediction using neural network. International Journal of Recent Technology and Engineering, v. 30, n. 5, 2019.

RODRIGUES, Adriana Alves et al. Visualização de dados no cenário da data science: práticas de laboratórios de inovação guiados por dados. 2019.

Tensores: https://www.datacamp.com/tutorial/investigating-tensors-pytorch

https://www.monolitonimbus.com.br/redes-neurais-artificiais/