Predição de Doenças Cardiovasculares: Uma Abordagem de Aprendizado de Máquina Utilizando Dados Clínicos de Pacientes

Felipe de Carvalho Frenhan
Instituto de Ciência e Tecnlogia
Universidade Federal de São Paulo
Bacharelado em Ciência e Tecnologia
São José dos Campos - SP
Email: carvalho.felipe@unifesp.br

Mariana Furtado dos Santos
Instituto de Ciência e Tecnlogia
Universidade Federal de São Paulo
Bacharelado em Ciência e Tecnologia
São José dos Campos - SP
Email: mariana.furtado@unifesp.br

I. Resumo

Este artigo explora a utilização de técnicas de aprendizado de máquina para a predição de doenças cardiovasculares a partir de dados clínicos. Foram comparados três modelos: Árvore de Decisão, Random Forest e Rede Neural, utilizando métricas como Acurácia, Acurácia Balanceada, Recall, Precisão e F1 Score.

II. Introdução

As doenças cardiovasculares (DCV) são a principal causa de mortalidade global, resultando em cerca de 17,9 milhões de mortes anuais, o que representa 32% de todos os óbitos no mundo. A insuficiência cardíaca, muitas vezes resultante de várias DCVs, destaca-se como uma preocupação crítica de saúde pública. Compreender e prever as DCVs é vital para reduzir a mortalidade e melhorar a qualidade de vida dos pacientes.

A utilização de modelos de aprendizagem para prever doenças cardiovasculares é uma abordagem promissora. Esses modelos podem analisar um conjunto de dados, facilitando a detecção precoce e a gestão eficaz de indivíduos em risco. Pessoas com fatores de risco como hipertensão, diabetes e hiperlipidemia se beneficiam significativamente de intervenções antecipadas. Assim, o desenvolvimento de ferramentas preditivas não só aprimora a precisão dos diagnósticos, mas também alivia a carga sobre os sistemas de saúde, permitindo uma melhor alocação de recursos e um atendimento mais eficiente e personalizado, além de que abre portas para um tratamento mais eficaz e com melhores chances de sucesso.

III. Conceitos Fundamentais

Para melhor compreensão do projeto, é bom definir adequadamente os conceitos biológicos e técnicos que serão usados ao decorrer do relatório.

- **"Doenças cardiovasculares (DCV)"** é um termo que engloba doenças que afetam o coração e os vasos sanguíneos do corpo, podendo afetar uma ou várias partes do coração e/ou vasos sanguíneos, DCV está associada ao acúmulo de depósitos de gordura dentro das artérias e a um risco aumentado de coágulos sanguíneos, bem como a danos nas artérias de órgãos como cérebro, coração, rins e olhos. Existem muitos tipos diferentes de doenças cardiovasculares, incluindo, mas não se limitando a arritmia, doença arterial coronariana, insuficiência cardíaca, doença arterial periférica, doença da aorta, doença cerebrovascular e trombose venosa profunda.
- **"Machine Learning (ML)"**, ou aprendizado de máquina, é um subcampo da inteligência artificial (IA) que se concentra no desenvolvimento de algoritmos e técnicas que permitem que computadores "aprendam" a partir de dados. Em vez de serem explicitamente programados para realizar uma tarefa, os sistemas de machine learning usam dados de exemplo para treinar modelos que podem fazer previsões ou tomar decisões com base em novos dados.
- **"Data Frame"** é uma estrutura de dados bidimensional. Ele se assemelha a uma tabela de banco de dados ou a uma planilha do Excel, onde os dados são organizados em linhas e colunas. Cada coluna pode conter um tipo de dado diferente (numérico, string, booleano, etc.), e cada linha representa uma observação ou registro.
- **"Árvore de decisão"** A árvore de decisão toma uma amostra, começa no nó raiz, e faz uma série de decisões baseadas nas características da amostra até que chega a uma folha, onde faz a predição final. A simplicidade e interpretabilidade das árvores de decisão são suas principais vantagens, mas elas também podem ser propensas ao overfitting, especialmente em conjuntos de dados complexos.
- **"Random Forest"** é a combinação de múltiplas árvores de decisão treinadas em diferentes amostras e características, proporcionando um modelo robusto e com menor risco de overfitting.
- **"Redes Neurais"** são modelos computacionais inspirados no funcionamento do cérebro humano, projetados para reconhecer padrões complexos e aprender a partir de dados. Elas são compostas por unidades chamadas neurônios

artificiais organizados em camadas, que se comunicam entre si através de conexões com pesos ajustáveis.

- **"Matriz de confusão"** é uma tabela que mostra o desempenho de um modelo de classificação, comparando as previsões do modelo com os valores reais dos dados. Ela mostra o número de previsões corretas e incorretas para cada classe do problema de classificação.
- **"Acurácia"** é uma métrica que mede a precisão geral de um modelo de classificação, indicando a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. Em outras palavras, é a porcentagem de casos corretamente classificados em relação ao total de casos avaliados.
- **"Acurácia Balanceada"** é uma métrica de desempenho utilizada especialmente em problemas de classificação com classes desbalanceadas, ou seja, quando o número de exemplos em uma classe é significativamente maior do que em outras. A acurácia balanceada busca corrigir o viés que a acurácia tradicional pode ter em favor da classe majoritária.
- **"Recall"** é uma métrica de avaliação de desempenho utilizada principalmente em problemas de classificação, especialmente em situações onde o foco é minimizar falsos negativos. O recall mede a capacidade de um modelo de identificar corretamente todas as instâncias positivas em um conjunto de dados.
- **"Precision"** é uma métrica de avaliação de desempenho usada principalmente em problemas de classificação. A precisão mede a proporção de exemplos corretamente classificados como positivos em relação a todos os exemplos que o modelo previu como positivos. Em outras palavras, ela indica a exatidão das previsões positivas do modelo.
- **"F1-Score"** é a média harmônica entre a precisão e o recall, o que significa que ele leva em consideração tanto os falsos positivos (erros do modelo onde ele identifica incorretamente algo como positivo) quanto os falsos negativos (erros onde o modelo falha em identificar algo que é realmente positivo).

IV. Trabalhos Relacionados

Diante da busca por soluções que agilizem diagnósticos e aprimorem o desempenho dos profissionais da saúde, têm surgido diversas investigações voltadas para a implementação de técnicas de Machine Learning no contexto médico. Os estudos apresentados nesta seção estão alinhados com o presente trabalho, pois utilizam técnicas de aprendizado de máquina para prever eventos de saúde, especificamente doenças cardiovasculares (DCV).

Em um estudo conduzido por Al-Absi et al. (2021) desenvolveram um modelo de aprendizado de máquina para diferenciar indivíduos saudáveis de pessoas com DCV, com o objetivo de identificar potenciais fatores de risco associados à doença. O estudo utilizou uma coleção de medidas biomédicas, incluindo aspectos comportamentais, do grupo DCV do Qatar Biobank (QBB). Como resultado, os autores identificaram o CatBoost como o modelo mais eficaz, alcançando uma acurácia de 93%. Além disso, o estudo revelou uma nova lista de fatores de risco, além dos já conhecidos, como distúrbio renal, função hepática e aterosclerose.

Em um outro estudo, conduzido por Weng et al. (2017), foram avaliados modelos de aprendizado de máquina para a previsão de doença cardiovascular ao longo de um período de 10 anos, em conformidade com as diretrizes do Código Americano de Cardiologia. A pesquisa utilizou informações de uma grande população de pacientes que foram acompanhados ao longo do tempo para entender como certos fatores podem influenciar o desenvolvimento da doença. Os resultados indicaram que o modelo baseado em redes neurais apresentou a melhor performance, com 355 predições corretas adicionais de doença cardiovascular em comparação com o modelo de referência do Código Americano de Cardiologia.

Por fim, um estudo recente liderado por Alaa et al.(2019) investigou a eficácia de diferentes métodos para prever doenças cardiovasculares (DCV) usando dados do UK Biobank. Os modelos de previsão de risco atualmente recomendados têm limitações, pois se baseiam em um número limitado de fatores de risco. Os pesquisadores testaram se técnicas baseadas em aprendizado de máquina, especialmente uma chamada "AutoPrognosis", poderiam melhorar a previsão de risco de DCV e se considerar variáveis não tradicionais poderia aumentar a precisão das previsões. Eles desenvolveram um modelo baseado em aprendizado de máquina usando dados de mais de 420 mil participantes do UK Biobank e compararam seu desempenho com modelos convencionais. O modelo AutoPrognosis melhorou significativamente a previsão de risco em comparação com os modelos tradicionais, incluindo a identificação de mais casos de DCV. Além disso, descobriram que incluir mais informações sobre os participantes no modelo preditivo teve um impacto maior do que usar modelos mais complexos.

V. Objetivo

A causa exata da DCV não é clara, mas há muitos fatores que podem aumentar o risco de contraí-la, sendo chamados de fatores de risco. Quanto mais fatores de risco tiver, maiores serão as chances de desenvolver DCV. Os principais fatores de riscos são a pressão alta, uso de tabaco, colesterol alto, sedentarismo, obesidade, idade, gênero, álcool e dieta rica em açúcar. Neste contexto, aplicamos técnicas de aprendizado de máquina para desenvolver um modelo que prevê se uma pessoa tem doença cardiovascular (DCV) ou não. Nosso objetivo foi criar um sistema capaz de analisar dados de entrada, que serão os principais fatores de risco de cada paciente e prever resultados com precisão. Para isso, treinamos o modelo com um conjunto de dados rotulados, permitindo que ele aprenda os padrões relevantes. Em seguida, avaliamos a eficácia do modelo comparando suas previsões com dados reais, utilizando métricas de desempenho para medir a precisão e a acurácia das previsões.

VI. METODOLOGIA EXPERIMENTAL

Para alcançar nosso objetivo de desenvolver um modelo de aprendizado de máquina para prever doenças cardiovasculares (DCV) com base nos principais fatores de risco dos pacientes, seguimos a seguinte metodologia experimental:

A. Domínio da Aplicação

Primeiramente, utilizamos um conjunto de dados obtido na plataforma Kaggle (plataforma que permite aos usuários compartilhar e acessar uma ampla variedade de datasets) encontrado a partir da plataforma IEEEDataPort. Este conjunto específico contém informações de 70.000 pacientes, cada um descrito por 12 características diferentes, tais como:

- **id:** Identificação única do indivíduo. Tipo: inteiro (int).
- **age:** Idade do paciente. Tipo: inteiro (int).
- **height:** Altura do indivíduo em centímetros. Tipo: inteiro (int).
- **weight:** Peso do indivíduo em quilogramas. Tipo: ponto flutuante (float).
- **gender:** Gênero do indivíduo, representado por valores categóricos. Tipo: categórico. Possíveis valores: [1: Feminino, 2: Masculino].
- **ap_hi:** Pressão arterial sistólica medida em milímetros de mercúrio (mmHg). Tipo: inteiro (int).
- **ap_lo:** Pressão arterial diastólica medida em milímetros de mercúrio (mmHg). Tipo: inteiro (int).
- **cholesterol:** Nível de colesterol, classificado em três categorias. Tipo: categórico. Possíveis valores: [1: Normal, 2: Acima do normal, 3: Muito acima do normal].
- **gluc:** Nível de glicose, classificado em três categorias. Tipo: categórico. Possíveis valores: [1: Normal, 2: Acima do normal, 3: Muito acima do normal].
- **smoke:** Indica se o indivíduo é fumante. Tipo: categórico. Possíveis valores: [0: Não é fumante, 1: É fumante].
- **alco:** Indica se o indivíduo consome álcool. Tipo: categórico. Possíveis valores: [0: Não ingere, 1: Ingere].
- **active:** Indica se o indivíduo pratica atividade física regularmente. Tipo: categórico. Possíveis valores: [0: Pratica, 1: Não pratica].
- **cardio:** Indica a presença de doença cardiovascular. Tipo: categórico. Possíveis valores: [0: Não tem, 1: Tem].

Esses atributos foram escolhidos com base principais fatores de risco e em estudos anteriores sobre DCV.

Utilizamos a linguagem de programação Python juntamente com as bibliotecas pandas, para manipulação e análise dos dados, numpy para operações numéricas, scikit-learn e tensorflow para implementar nossos modelos de aprendizado de máquina. Para testar cada modelo, geramos uma matriz de confusão para calcular a acurácia daquele modelo. Fizemos isso para todos os modelos, a fim de encontrar o melhor.

B. Pré-Processamento

O primeiro passo foi realizar o pré-processamento dos dados, que incluiu a limpeza e preenchimento de dados ausentes, normalização de características e codificação de variáveis categóricas.

*1) **Limpeza dos dados**:* Durante a etapa de limpeza de dados, realizamos uma série de verificações e ajustes para garantir a qualidade e integridade do conjunto de dados utilizado. Inicialmente, verificamos a presença de dados duplicados e faltantes; como não havia registros duplicados ou valores ausentes, prosseguimos com a análise de valores nulos ou zero em colunas específicas, incluindo ”age”, ”height”, ”weight”, ”gender”, ”ap_hi” (pressão arterial sistólica), ”ap_lo” (pressão arterial diastólica) e ”cholesterol”.

Identificamos 21 casos em que a pressão arterial diastólica (ap_lo) foi registrada como 0 mmHg. Esses registros foram considerados inválidos e, portanto, excluídos do conjunto de dados. Em seguida, examinamos a consistência entre as medições de pressão arterial, verificando casos em que a pressão arterial diastólica (ap_lo) era maior que a pressão arterial sistólica (ap_hi), o que não é fisiologicamente possível. Foram identificadas 1.234 ocorrências desse tipo, que também foram removidas.

Além disso, avaliamos a presença de valores negativos nas colunas de pressão arterial. Encontramos um registro onde a pressão arterial diastólica (ap_lo) tinha um valor negativo. Após revisão e comparação com outros dados relacionados, concluímos que o valor negativo era provavelmente devido a um erro de digitação; assim, o valor foi corrigido para o correspondente positivo.

Por último, identificamos e tratamos outliers nas medições de pressão arterial sistólica (ap_hi) e diastólica (ap_lo). Considerando que outliers são valores discrepantes que se desviam significativamente da maioria dos outros dados, definimos limites para evitar influências indevidas nas análises subsequentes. Para a pressão arterial sistólica (ap_hi), estabelecemos um intervalo de 40 a 250 mmHg. Para a pressão arterial diastólica (ap_lo), o intervalo foi definido entre 30 e 160 mmHg. Registros que apresentavam valores fora desses limites foram removidos do conjunto de dados.

*2) **Transformação dos dados**:* Na etapa de transformação dos dados, realizamos diversas modificações para adequar o conjunto de dados às análises subsequentes. Primeiramente, a coluna de idade (”age”), originalmente expressa em dias, foi convertida para anos para facilitar a interpretação e análise dos dados. Em seguida, calculamos o Índice de Massa Corpórea (IMC) para cada indivíduo, utilizando as colunas de peso (”weight”) e altura (”height”). O IMC é uma medida adotada pela Organização Mundial da Saúde (OMS) para avaliar o peso ideal de uma pessoa com base em sua altura e peso.

Após o cálculo do IMC, criamos uma nova coluna específica para essa variável e, subsequente a isso, removemos as colunas de peso e altura para evitar redundâncias no conjunto de dados. A classificação do IMC seguiu os critérios estabelecidos pela OMS: valores menores que 18,5 foram classificados como magreza; entre 18,5 e 24,9 como normal; entre 25,0 e 29,9 como sobrepeso; entre 30,0 e 39,9 como obesidade; e valores maiores que 40,0 como obesidade grave.

Para assegurar a integridade e qualidade dos dados, examinamos possíveis outliers na distribuição do IMC. Com base em conhecimento médico, estabelecemos limites para o IMC entre

um mínimo de 10 e um máximo de 60. Registros com valores fora desse intervalo foram removidos do conjunto de dados para evitar distorções nas análises estatísticas posteriores.

Durante a preparação dos dados para análise, foi necessário transformar variáveis categóricas em um formato que pudesse ser utilizados nos modelos de aprendizado de máquina. Para isso, utilizamos uma técnica chamada One-Hot Encoding da biblioteca scikit-learn. Essa técnica transforma variáveis que contêm categorias (como ”gender”, ”cholesterol”, ”gluc”, ”smoke”, ”alco” e ”active”) em variáveis numéricas binárias. Essa transformação é importante porque muitos algoritmos de aprendizado de máquina funcionam melhor com dados numéricos.

O processo envolveu a codificação das variáveis categóricas convertendo cada categoria em uma coluna separada que indica a presença ou ausência dessa categoria para cada registro e a integração com o conjunto de dados, sendo as novas colunas resultantes dessa codificação adicionadas ao conjunto de dados, e as colunas originais que continham as categorias, foram removidas para evitar redundâncias.

Usamos a padronização com a função StandardScaler da biblioteca scikit-learn que é uma técnica de pré-processamento de dados usada para transformar os dados de modo que tenham média zero e desvio padrão igual a um, garantindo que todos os recursos estejam na mesma escala e, assim, melhorando o desempenho e a eficácia do modelo.

C. Reconhecimento de padrões

Para o treinamento do modelo, exploramos diferentes algoritmos de aprendizado de máquina. Analisamos e testamos 3 modelos que julgamos adequados para este problema de classificação, que não é linearmente separável: Árvore de Decisão (Decision Tree), Floresta Aleatória (Random Forest) e Redes Neurais Artificiais (Artificial Neural Networks, ANN).

*1) **Árvore de Decisão**:* Para otimizar o desempenho do modelo, adotamos uma abordagem sistemática, dividida em algumas etapas:

- **Seleção de Hiperparâmetros:** Os hiperparâmetros definem a estrutura e o funcionamento do modelo. Para encontrar a melhor combinação, utilizamos a técnica de Randomized Search com validação cruzada. Foram testadas diversas configurações, como a profundidade máxima da árvore, o número mínimo de amostras necessárias para dividir um nó, o critério de divisão, entre outros. A combinação de hiperparâmetros que apresentou o melhor desempenho foi selecionada para o treinamento final.
- **Validação Cruzada:** Para garantir que o modelo generalize bem para novos dados, utilizamos a validação cruzada com k-fold igual a 5. Nesse processo, o conjunto de dados de treinamento foi dividido em 5 partes, onde o modelo foi treinado em algumas delas e testado nas restantes. As métricas de desempenho, como acurácia, acurácia balanceada, recall e F1-Score, foram calculadas para cada repetição, assegurando que o modelo não esteja superajustado aos dados de treinamento.
- **Treinamento:**Após identificar a melhor configuração de hiperparâmetros e validar o modelo, realizamos o treinamento final utilizando todos os dados de treinamento disponíveis. Em seguida, o modelo foi testado em um conjunto de dados de teste, que não foi utilizado durante o treinamento, para avaliar seu desempenho em dados novos. As métricas finais, incluindo acurácia, acurácia balanceada, recall, precisão e F1-Score, foram calculadas para entender a eficácia do modelo em prever corretamente a presença de doenças cardiovasculares.

*2) **Floresta Aleatória (Random Forest)**:* Para otimizar o desempenho do modelo, seguimos uma abordagem sistemática envolvendo várias etapas:

- **Seleção de Hiperparâmetros:** Os hiperparâmetros são configurações que definem a estrutura do modelo e a forma como ele aprende. Para encontrar a melhor combinação desses hiperparâmetros (como o número de árvores na floresta, a profundidade máxima das árvores, e o critério de divisão das árvores), utilizamos uma técnica chamada Randomized Search com validação cruzada. Essa técnica testa diversas combinações de configurações de forma aleatória e avalia o desempenho do modelo para cada uma delas. Os hiperparâmetros que apresentaram o melhor desempenho foram selecionados para o treinamento final do modelo.
- **Validação Cruzada:** Para garantir que o modelo não apenas se ajuste bem aos dados de treinamento, mas também generalize bem para novos dados, utilizamos um método chamado validação cruzada. Este método divide os dados de treinamento em várias partes (folds), treina o modelo em algumas delas e o testa nas restantes, no nosso caso, usamos k-fold igual à 5. Esse processo é repetido várias vezes, e as métricas de desempenho (como acurácia e recall) são calculadas para cada repetição. Isso ajuda a garantir que o modelo não está ”memorizando” os dados de treinamento, mas sim aprendendo padrões úteis.
- **Treinamento:** Após encontrar a melhor configuração de hiperparâmetros e validar o modelo, realizamos o treinamento final utilizando todos os dados de treinamento disponíveis. Em seguida, o modelo foi testado em um conjunto de dados de teste, que não foi utilizado durante o treinamento, para avaliar seu desempenho em dados totalmente novos. As métricas finais, como acurácia, acurácia balanceada, recall e F1-Score, foram calculadas para entender a capacidade do modelo de identificar corretamente pacientes com e sem doenças cardíacas.

*3) **Rede Neural**:* A rede neural foi configurada com várias camadas, incluindo camadas de entrada, ocultas e de saída, e cada uma desempenha um papel específico no aprendizado dos padrões nos dados.

- **Preparação dos dados:** Antes de treinar a rede neural, foi necessário preparar os dados de entrada. Isso incluiu transformar a variável de saída ’cardio’ em uma forma adequada para a rede neural, conhecida como codificação ”One-Hot”. Este processo cria uma

representação binária para cada categoria de saída, essencial para a configuração da rede que utilizamos.

- **Estrutura da Rede Neural:** A rede neural em nosso estudo foi configurada com várias camadas. Primeiramente, uma camada de entrada que recebe todas as características dos pacientes. Seguida por duas camadas ocultas (com um número variável de neurônios, ajustado como hiperparâmetros) que utilizam uma função de ativação chamada "ReLU" para introduzir não-linearidades no modelo, permitindo que ele aprenda padrões mais complexos. Entre as camadas ocultas, foi adicionado um mecanismo de "Dropout", que desativa aleatoriamente uma fração dos neurônios durante o treinamento para evitar o sobreajuste (overfitting). Por fim, a camada de saída utiliza uma função de ativação "softmax" para produzir probabilidades para cada classe (presença ou ausência de doença cardíaca).
- **Otimização de Hiperparâmetros:** Para encontrar a configuração ideal da rede neural, utilizamos um processo chamado Randomized Search. Esse processo testa aleatoriamente diferentes combinações de hiperparâmetros (como número de neurônios em cada camada, a taxa de dropout, o número de épocas de treinamento, e o tamanho do lote de treinamento). O objetivo é identificar a combinação que proporciona o melhor desempenho do modelo em termos de precisão e outras métricas relevantes.
- **Validação Cruzada:** Para avaliar a eficácia do modelo e garantir que ele generalize bem para novos dados, aplicamos uma técnica chamada validação cruzada. Dividimos o conjunto de dados de treinamento em 5 partes (folds), treinando o modelo em algumas e validando em outras. Repetimos esse processo para todas as combinações possíveis, o que ajuda a avaliar o desempenho do modelo de forma mais robusta e a evitar que ele se ajuste demais aos dados de treinamento.
- **Treinamento:** Após encontrar os melhores hiperparâmetros, treinamos o modelo com o conjunto de dados completo. Em seguida, avaliamos seu desempenho em um conjunto de dados de teste que não foi utilizado durante o treinamento. As métricas de desempenho, como acurácia, acurácia balanceada, recall, precision e F1-Score, foram calculadas para fornecer uma visão abrangente da capacidade do modelo em prever corretamente os casos de doenças cardíacas.

VII. Análise dos resultados

A. *Árvore de decisão*

Os resultados obtidos com o modelo de Árvore de Decisão foram avaliados de forma detalhada, com foco em métricas que permitissem entender a eficácia do modelo. A análise incluiu a construção de uma matriz de confusão, que serviu para verificar a capacidade do modelo em distinguir corretamente entre as classes do problema. A acurácia balanceada foi a métrica principal utilizada para avaliar o desempenho do modelo. Essa métrica é particularmente importante em cenários onde os dados podem estar desbalanceados, como no caso de doenças cardiovasculares, onde a prevalência de uma classe pode ser maior que a outra. O valor da melhor acurácia balanceada obtida foi 0.7197, o que indica que o modelo conseguiu um equilíbrio razoável na classificação correta das duas classes. Os melhores hiperparâmetros encontrados para o modelo de Árvore de Decisão foram:

- "criterion": 'entropy'
- "max_depth": 30
- "max_features": None
- "mix_leaf_nodes": 40
- "min_impurity_decrease": 0.0
- "min_samples_leaf": 4
- "min_samples_split": 9

Esses hiperparâmetros foram otimizados para maximizar o desempenho do modelo, garantindo uma boa generalização sem superajuste aos dados de treinamento. O critério de divisão escolhido, "entropy", visa maximizar a informação adquirida a cada divisão do nó, enquanto a profundidade máxima de 30 foi selecionada para capturar complexidades nos dados sem levar o modelo a sobreajustar. O número limitado de nós folha e a necessidade de um número mínimo de amostras para divisão também foram fatores que contribuíram para um desempenho robusto.

B. *Random Forest*

Os resultados obtidos com o modelo Random Forest foram avaliados principalmente através de uma matriz de confusão, que demonstrou a capacidade do modelo de identificar corretamente as diferentes classes do problema. A métrica de acurácia balanceada foi utilizada como principal indicador de desempenho para assegurar que o modelo não favorecesse uma classe em detrimento da outra, sendo especialmente útil em casos de desbalanceamento dos dados. Os melhores hiperparâmetros para o modelo Random Forest foram:

- "bootstrap": True
- "criterion": 'gini'
- "max_depth": 10
- "max_leaf_nodes": None
- "min_impurity_decrease": 0.0
- "min_samples_leaf": 1
- "min_samples_split": 10
- "n_estimators": 165

Esses parâmetros foram otimizados para maximizar o desempenho do modelo em termos de generalização e precisão, levando em consideração a variabilidade inerente dos dados.

C. *Rede Neural*

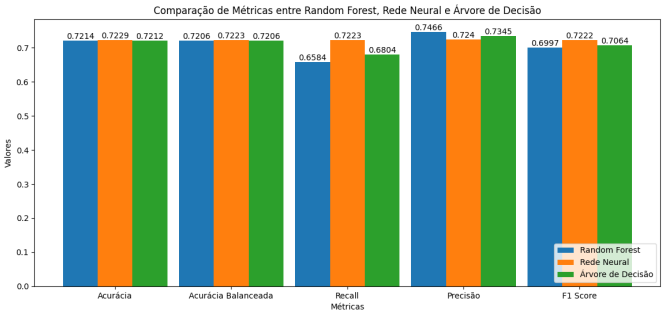
Para o modelo de Rede Neural, a análise dos resultados também envolveu a construção de uma matriz de confusão, que evidenciou a capacidade do modelo em classificar corretamente tanto os casos positivos quanto os negativos. Além disso, foram geradas curvas de aprendizado durante o treinamento e a validação em cada uma das divisões (folds) dos dados. Essas curvas foram fundamentais para visualizar a

melhoria do modelo ao longo do tempo, assim como para identificar potenciais problemas de overfitting (ajuste excessivo ao conjunto de treinamento) ou underfitting (ajuste insuficiente ao padrão dos dados). Os melhores hiperparâmetros encontrados para a Rede Neural foram:

- ”units1: 64
- ”units2”: 16
- ”optimizer”: ’adam’
- ”epochs”: 30
- ”dropout”: 0.3
- ”batch_size”: 64

Estes hiperparâmetros foram selecionados para equilibrar a complexidade do modelo e sua capacidade de generalização, minimizando o risco de sobreajuste e melhorando a robustez das previsões.

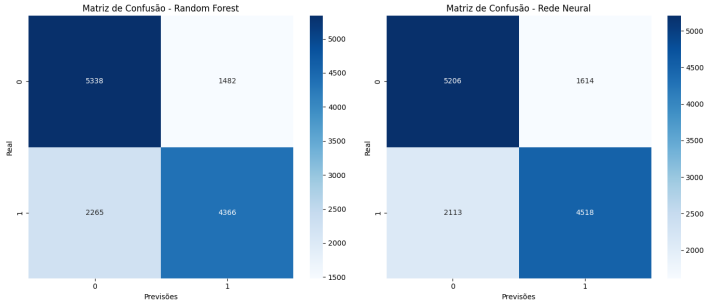
D. Comparação de desempenho

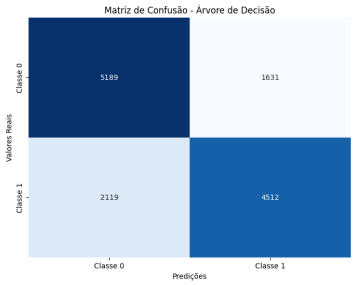


O gráfico apresentado ilustra a comparação de desempenho entre três modelos de aprendizado de máquina: Random Forest, Rede Neural e Árvore de Decisão. Esses modelos foram utilizados para a predição de doenças cardiovasculares com base em dados clínicos de pacientes. As métricas avaliadas incluem Acurácia, Acurácia Balanceada, Recall, Precisão e F1 Score.

- **Acurácia:** A acurácia mede a proporção de previsões corretas (tanto verdadeiros positivos quanto verdadeiros negativos) em relação ao total de previsões feitas. Observa-se que a Rede Neural obteve a maior acurácia (0.7229), seguida pela Árvore de Decisão (0.7212) e pelo Random Forest (0.7214). A pequena diferença entre os modelos indica que todos eles têm um desempenho semelhante em termos de acurácia global. No entanto, a ligeira superioridade da Rede Neural pode sugerir uma maior capacidade de generalização para este conjunto de dados específico.
- **Acurácia Balanceada:** A acurácia balanceada, que é relevante em cenários onde as classes estão balanceadas, indica a capacidade do modelo em tratar as classes de maneira equitativa. Neste caso, tanto a Rede Neural quanto a Árvore de Decisão alcançaram valores muito próximos (0.7223 e 0.7206, respectivamente), enquanto o Random Forest apresentou uma acurácia balanceada ligeiramente inferior (0.7206). Como os dados estão balanceados, esses resultados indicam que os três modelos têm um desempenho consistente, com a Rede Neural mostrando uma leve superioridade.
- **Recall:** O recall (ou sensibilidade) avalia a capacidade do modelo em identificar corretamente os verdadeiros positivos, ou seja, pacientes que de fato possuem a doença cardiovascular. Aqui, o modelo de Rede Neural se destacou com o maior valor de recall (0.7223), seguido pela Árvore de Decisão (0.6804). O Random Forest apresentou um recall significativamente menor (0.6584). Esse resultado sugere que a Rede Neural é mais eficaz na detecção de casos positivos, o que é crucial em aplicações médicas onde a detecção precoce é essencial para o tratamento eficaz.
- **Precisão:** A precisão mede a proporção de verdadeiros positivos em relação ao total de previsões positivas feitas pelo modelo. A Árvore de Decisão alcançou o maior valor de precisão (0.7345), indicando que, quando este modelo prevê um paciente como tendo uma doença cardiovascular, há uma maior probabilidade de estar correto em comparação com os outros modelos. A Rede Neural também teve um bom desempenho (0.7466), seguido pelo Random Forest (0.724). Estes resultados indicam que tanto a Rede Neural quanto a Árvore de Decisão são eficazes em evitar falsos positivos, mas a Árvore de Decisão tem uma leve vantagem.
- **F1 Score:** O F1 Score é a média harmônica entre precisão e recall, oferecendo uma única métrica que equilibra esses dois aspectos. Novamente, a Rede Neural obteve o maior F1 Score (0.7222), indicando que ela oferece o melhor equilíbrio entre precisão e recall. A Árvore de Decisão segue de perto (0.7064), enquanto o Random Forest fica atrás com um F1 Score de 0.6997. Isso reforça a robustez da Rede Neural como uma opção equilibrada para a predição de doenças cardiovasculares.

A análise do gráfico revela que, embora todos os modelos tenham desempenhos comparáveis, a Rede Neural se destaca como o modelo mais robusto e equilibrado para a predição de doenças cardiovasculares, superando os outros modelos em várias métricas importantes como recall e F1 Score. No entanto, a Árvore de Decisão demonstra uma precisão ligeiramente superior, o que pode ser útil em cenários onde a minimização de falsos positivos é crucial. Já o Random Forest, embora eficaz, mostrou-se menos eficiente em comparação com os outros dois modelos em métricas-chave como recall e F1 Score.





As matrizes de confusão apresentadas, juntamente com as métricas de desempenho discutidas anteriormente (Acurácia, Acurácia Balanceada, Recall, Precisão e F1 Score), fornecem uma visão abrangente sobre o desempenho dos modelos Random Forest, Rede Neural e Árvore de Decisão na predição de doenças cardiovasculares.

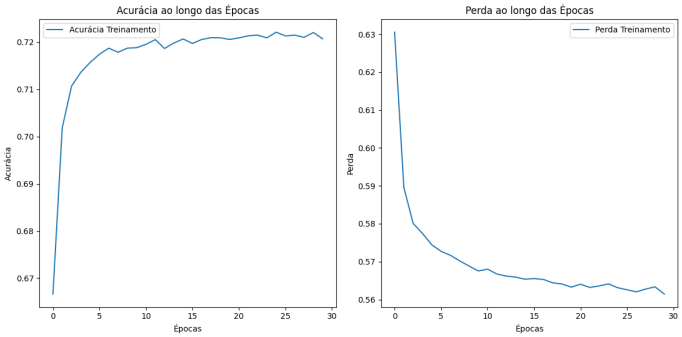
- **Random Forest:**
  - Verdadeiros Negativos: 5338
  - Falsos Positivos: 1482
  - Falsos Negativos: 2265
  - Verdadeiros Positivos: 4366

  O Random Forest apresentou a menor acurácia entre os três modelos, o que é corroborado pelo elevado número de falsos negativos (2265), que reduzem o recall para 0.6584, o menor entre os três modelos. Isso indica que o Random Forest teve dificuldades em identificar corretamente os casos positivos de doenças cardiovasculares. Apesar disso, ele apresentou um bom desempenho na identificação de pacientes sem a doença (elevado número de verdadeiros negativos), refletido em sua precisão de 0.724. O F1 Score relativamente baixo (0.6997) indica um equilíbrio desfavorável entre recall e precisão.
- **Rede Neural:**
  - Verdadeiros Negativos: 5206
  - Falsos Positivos: 1614
  - Falsos Negativos: 2113
  - Verdadeiros Positivos: 4518

  A Rede Neural apresentou o melhor equilíbrio entre as métricas, com o maior valor de recall (0.7223) e um alto F1 Score (0.7222). A matriz de confusão revela que este modelo conseguiu identificar corretamente a maioria dos casos positivos (4518 verdadeiros positivos), com um número relativamente baixo de falsos negativos (2113), o que justifica o recall elevado. A acurácia ligeiramente superior (0.7229) e a acurácia balanceada (0.7223) indicam um desempenho consistente mesmo com dados balanceados. O modelo, no entanto, apresentou o maior número de falsos positivos (1614), o que resulta em uma precisão ligeiramente inferior à da Árvore de Decisão.
- **Árvore de Decisão:**
  - Verdadeiros Negativos: 5189
  - Falsos Positivos: 1631
  - Falsos Negativos: 2119
  - Verdadeiros Positivos: 4512

  A Árvore de Decisão demonstrou um desempenho que se posiciona entre os modelos de Rede Neural e Random Forest. Embora tenha uma acurácia ligeiramente inferior à da Rede Neural (0.7212), a Árvore de Decisão apresentou uma precisão elevada (0.7345), o que é refletido na matriz de confusão, onde vemos um número moderado de falsos positivos (1631) e falsos negativos (2119). O recall mais baixo (0.6804) em relação à Rede Neural, juntamente com um F1 Score de 0.7064, sugere que, embora este modelo seja bom em evitar falsos positivos, ele sacrifica a sensibilidade, resultando em um maior número de casos positivos não identificados.

E. Curva de aprendizado da Rede Neural



As curvas de aprendizado apresentadas fornecem uma visão sobre o desempenho da rede neural ao longo do treinamento, com a evolução da Acurácia e da Perda monitoradas em 30 épocas. As duas curvas — uma para a acurácia e outra para a perda — são essenciais para entender como o modelo está aprendendo com os dados ao longo do tempo.

- **Curva de Acurácia:** A curva da esquerda mostra a evolução da acurácia ao longo das épocas de treinamento. Inicialmente, a acurácia cresce rapidamente, subindo de cerca de 0,67 na primeira época para aproximadamente 0,71 na quinta época. Esse crescimento inicial acentuado indica que o modelo está aprendendo rapidamente, ajustando seus parâmetros de maneira significativa nas primeiras interações com os dados.

  A partir da quinta época, o crescimento da acurácia começa a desacelerar, estabilizando em torno de 0,72 entre a 10ª e a 15ª época. Essa estabilização sugere que o modelo está se aproximando de seu potencial máximo de acurácia para o conjunto de dados utilizado, com apenas pequenas variações até o final das 30 épocas. A leve flutuação na acurácia ao final do treinamento pode ser atribuída a ajustes menores nos pesos da rede que não resultam em ganhos substanciais de acurácia, mas também não indicam sobreajuste (overfitting) significativo.
- **Curva de Perda:** A curva da direita mostra a perda ao longo das épocas. Inicialmente, a perda é alta, indicando que o modelo comete muitos erros no início do treinamento. Conforme as épocas avançam, a perda diminui drasticamente, especialmente nas primeiras cinco épocas, atingindo uma estabilização gradual em torno de 0,56 a partir da 20ª época.

  A redução contínua da perda, especialmente nas primeiras épocas, é um sinal de que o modelo está ajustando

bem seus parâmetros, melhorando sua capacidade de fazer previsões corretas. A estabilização da perda após a 20ª época, combinada com a estabilização da acurácia, sugere que o modelo atingiu um ponto onde as mudanças adicionais nos pesos não contribuem significativamente para a melhoria do desempenho.

VIII. Conclusões

Neste estudo, avaliamos o desempenho de três modelos de aprendizado de máquina — Árvore de Decisão, Random Forest e Rede Neural — na predição de doenças cardiovasculares utilizando dados clínicos de pacientes. A análise abrangeu diversas métricas, incluindo Acurácia, Acurácia Balanceada, Recall, Precisão e F1 Score, complementadas pela avaliação detalhada das matrizes de confusão de cada modelo.

Os resultados indicam que a Rede Neural se destacou como o modelo mais robusto e equilibrado, apresentando o melhor desempenho em métricas cruciais como Recall (0.7223) e F1 Score (0.7222). Isso sugere que a Rede Neural tem uma maior capacidade de identificar corretamente os casos positivos, o que é essencial na detecção de doenças cardiovasculares, onde a identificação precoce pode salvar vidas.

Por outro lado, a Árvore de Decisão apresentou a maior precisão (0.7345), tornando-se uma opção interessante em cenários onde a minimização de falsos positivos é crucial. No entanto, esse modelo demonstrou uma menor sensibilidade, refletida em um Recall inferior (0.6804), o que implica em uma quantidade maior de casos positivos não identificados.

O Random Forest, apesar de apresentar uma alta precisão na identificação de pacientes sem a doença, mostrou-se menos eficaz na detecção de casos positivos, com o menor valor de Recall (0.6584) entre os três modelos. Isso resultou em um desempenho inferior nas métricas de equilíbrio entre sensibilidade e precisão, como o F1 Score (0.6997).

Em síntese, a escolha do modelo ideal para a predição de doenças cardiovasculares deve considerar o equilíbrio entre a necessidade de minimizar falsos negativos (detecção correta de casos positivos) e falsos positivos (evitar alarmes falsos). A Rede Neural emerge como a melhor opção para maximizar a sensibilidade e o equilíbrio geral de desempenho, enquanto a Árvore de Decisão pode ser preferível em aplicações que demandam alta precisão. O Random Forest, apesar de ser uma ferramenta valiosa, pode necessitar de ajustes ou complementação com outras abordagens para otimizar seu desempenho na detecção de doenças cardiovasculares.

Estes achados contribuem para o campo de estudo das aplicações de aprendizado de máquina na medicina, evidenciando a importância da escolha adequada do modelo em função dos objetivos clínicos específicos e dos trade-offs inerentes a cada abordagem.

IX. Referências

[1] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one, 12(4), e0174944. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28376093/. Acesso em: 24 mai. 2024.

[2] Al-Absi, H. R. H., Refaee, M. A., Rehman, A. U., Islam, M. T., Belhaouari, S. B., & Alam, T. (2021). Risk factors and comorbidities associated to cardiovascular disease in Qatar: A machine learning based case-control study. IEEE Access, 9, 29929–29941. Disponível em: https://ieeexplore.ieee.org/document/9354689. Acesso em: 23 mai. 2024.

[3] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & van der Schaar, M. (2019). Can Machine Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? PLoS ONE, 14(5), e0213653. Disponível em: https://doi.org/10.1371/journal.pone.0213653. Acesso em: 25 mai. 2024.

[4] Organização Mundial da Saúde (OMS). "Cardiovascular diseases (CVDs)." Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 25 mai. 2024.

[5] Kaggle, 2021. Disponível em: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-datasetprediction/data. Acesso em: 23 maio 2024.

[6] FELIPE, Davi. Matriz de confusão: nunca mais se confunda utilizando esse exemplo. Medium, 30 set. 2019. Disponível em: https://medium.com/comunidadeds/matriz-de-confus%C3%A3o-nunca-mais-se-confunda-utilizando-esse-exemplo-35a9ac63b88a. Acesso em: 25 maio 2024.

[7] SCIKIT-LEARN: Machine Learning in Python. Disponível em: https://scikit-learn.org/stable/. Acesso em: 24 maio 2024.

[8] Tipos de aprendizado de máquina e algumas aplicações. TerraLAB - Laboratório de Computação Aplicada, Universidade Federal de Ouro Preto. Disponível em: https://www2.decom.ufop.br/terralab/tipos-de-aprendizado-de-maquina-e-algumas-aplicacoes/. Acesso em: 25 maio 2024.

[9] DOENÇAS cardiovasculares. Organização Pan-Americana da Saúde (OPAS). Disponível em: https://www.paho.org/pt/topicos/doencas-cardiovasculares. Acesso em: 25 maio 2024.

[10] Rodrigues, Vitor Borba. Métricas de Avaliação: Acurácia, Precisão, Recall — Quais as Diferenças?. Medium, 2020. Disponível em: https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c. Acesso em: 23 maio 2024.