Previsão de Doença Cardíaca com Aprendizado de Máquina: Um Estudo com Dados Clínicos da UCI (University of California, Irvine)

1. Pré-processamento dos Dados

1.1 Importação de Bibliotecas

O código começa importando as bibliotecas necessárias. Essas bibliotecas são essenciais para manipulação de dados (pandas e numpy), visualização de gráficos (matplotlib e seaborn), e construção e avaliação de modelos de aprendizado de máquina (sklearn).

1.2 Carregamento dos Dados

Os dados sobre doenças cardíacas são carregados de um arquivo no repositório de Machine Learning da UCI. Cada coluna do dataset recebe um nome descritivo para facilitar a compreensão do conteúdo.

1.3 Tratamento de Valores Faltantes

Primeiramente, o código verifica se há valores faltantes. Valores representados por "?" são substituídos por NaN, e todas as colunas são convertidas para valores numéricos. Depois disso, linhas com valores faltantes são removidas.

1.4 Análise das Variáveis

A distribuição das variáveis numéricas e categóricas é visualizada usando gráficos de histograma e gráficos de barras, respectivamente. Isso ajuda a entender melhor como os dados estão distribuídos. Um mapa de calor é gerado para mostrar a correlação entre as variáveis numéricas.

2. Divisão do Dataset

2.1 Separação em Features e Target

Os dados são separados em duas partes: features (variáveis independentes) e alvo (variável dependente que queremos prever). A variável alvo é convertida em binária, onde 1 indica a presença de doença cardíaca e 0 indica a ausência.

2.2 Transformações nas Features

Para preparar os dados para o modelo de aprendizado de máquina, são aplicadas transformações: imputação de valores faltantes para variáveis numéricas, padronização (normalização) e codificação one-hot para variáveis categóricas.

2.3 Divisão em Conjuntos de Treinamento e Teste

Os dados são divididos em conjuntos de treinamento e teste, com 70% dos dados usados para treinar o modelo e 30% para testar sua eficácia.

3. Treinamento do Modelo

Um modelo de Árvore de Decisão é escolhido e treinado com os dados de treinamento. Este modelo é uma técnica de aprendizado supervisionado que é usada para classificação.

4. Avaliação do Modelo

4.1 Predição com Dados de Teste

O modelo treinado é utilizado para fazer predições com os dados de teste.

4.2 Cálculo de Métricas de Desempenho

As predições do modelo são avaliadas utilizando várias métricas de desempenho, como acurácia, matriz de confusão e relatório de classificação, que incluem precisão, recall e F1-score.

5. Interpretação dos Resultados

5.1 Importância das Variáveis

A importância das features é calculada para entender quais variáveis mais influenciam o modelo. As variáveis com maior importância são aquelas que mais contribuem para a previsão da doença cardíaca.

5.2 Visualização dos Resultados

Um mapa de calor da matriz de confusão é gerado para mostrar a quantidade de acertos e erros do modelo. Um scatterplot é criado para comparar as predições do modelo com os valores reais, permitindo visualizar a precisão das predições.

Conclusão

O projeto utiliza técnicas de aprendizado de máquina para prever a presença de doenças cardíacas com base em dados clínicos. A análise detalhada e as visualizações fornecem insights valiosos sobre as variáveis mais influentes e a precisão do modelo. Esta abordagem pode ser uma ferramenta poderosa na identificação precoce de pacientes em risco, contribuindo para intervenções mais direcionadas e eficazes na área da saúde.