

Relatório de Mineração de Dados - Heart Disease UCI

1. Informações do Aluno

Aluno: Anderson Cagnini

Turma: 5º Semestre - Ciência da Computação

Data: 06/07/2025

Link do Github: <https://github.com/AndersonCagnini/Projeto-de-Mineracao-de-Dados>

2. Resumo Executivo

Neste projeto, foram aplicadas técnicas de mineração de dados para prever a presença de doenças cardíacas em pacientes. O modelo Random Forest apresentou os melhores resultados, com acurácia de 98% e AUC de 0.94. Variáveis clínicas como tipo de dor no peito e frequência cardíaca máxima foram as mais relevantes para o diagnóstico.

3. Dataset Utilizado

O dataset utilizado neste trabalho foi o 'Heart Disease UCI', disponível no Kaggle através do link: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>. Este dataset contém 303 registros e 14 colunas relacionadas a características clínicas de pacientes.

4. Objetivo do Projeto

O objetivo é aplicar técnicas de mineração de dados para prever a presença de doença cardíaca em pacientes, utilizando modelos de classificação, com foco principal no algoritmo Random Forest.

5. Metodologia

5.1 Análise Exploratória

Foram explorados gráficos de distribuição, correlação e estatísticas descritivas para compreender os padrões do dataset.

5.2 Pré-processamento

Os dados foram tratados e normalizados, com separação em conjuntos de **treino** e **teste**, respeitando a proporção dos dados originais.

5.3 Modelagem

Foram aplicados e comparados os seguintes modelos:

- Random Forest
- Regressão Logística
- Árvore de Decisão
- K-Nearest Neighbors (KNN)

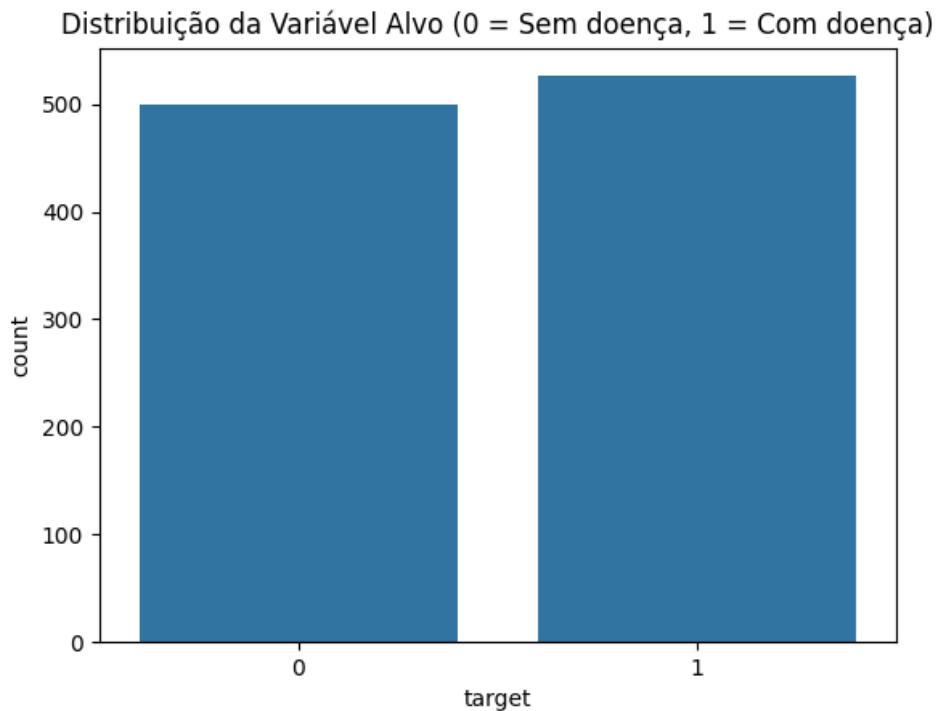
5.4 Validação

Foi utilizada validação cruzada k-fold para garantir confiabilidade nos resultados.

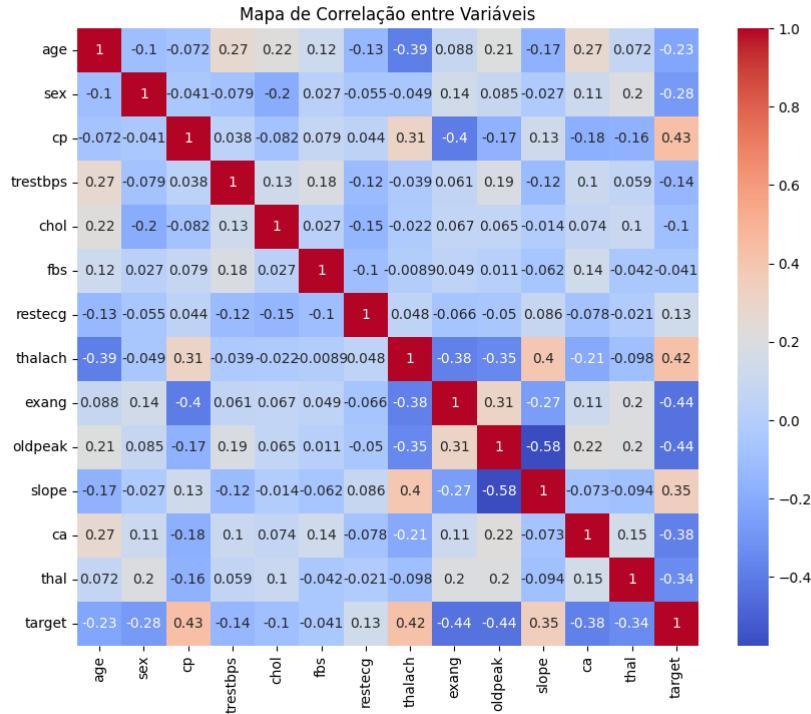
6. Resultados Visuais

A seguir estão os gráficos gerados durante a análise. (Cole os prints nos locais indicados)

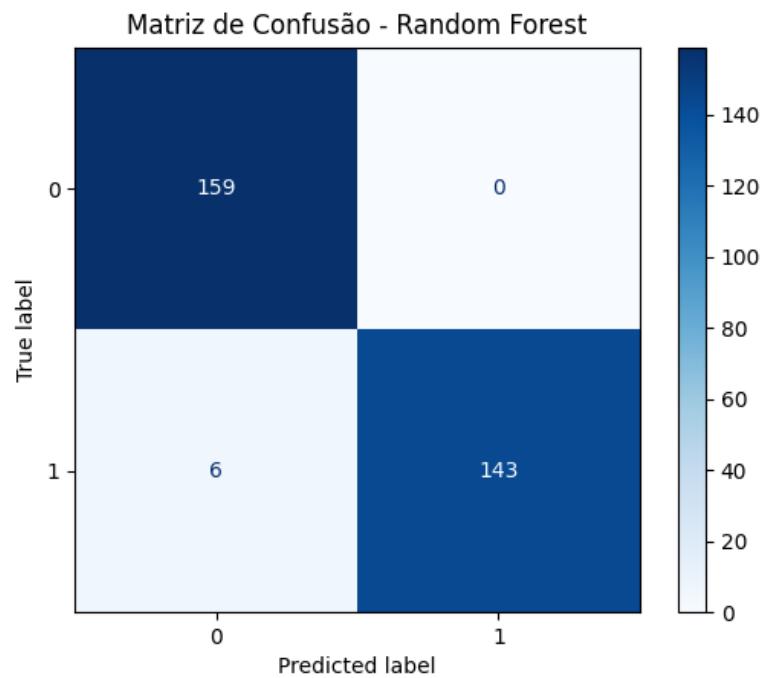
📊 Gráfico de distribuição da variável alvo: mostra a proporção entre pacientes com e sem doença



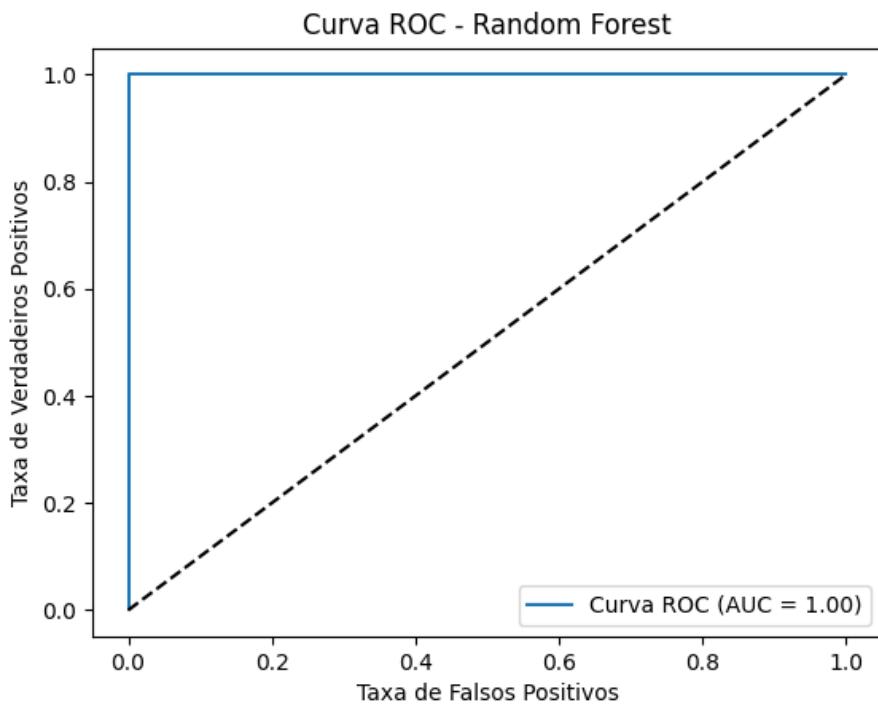
📈 Mapa de correlação entre variáveis: revela relações entre variáveis clínicas, destacando thalach e cp



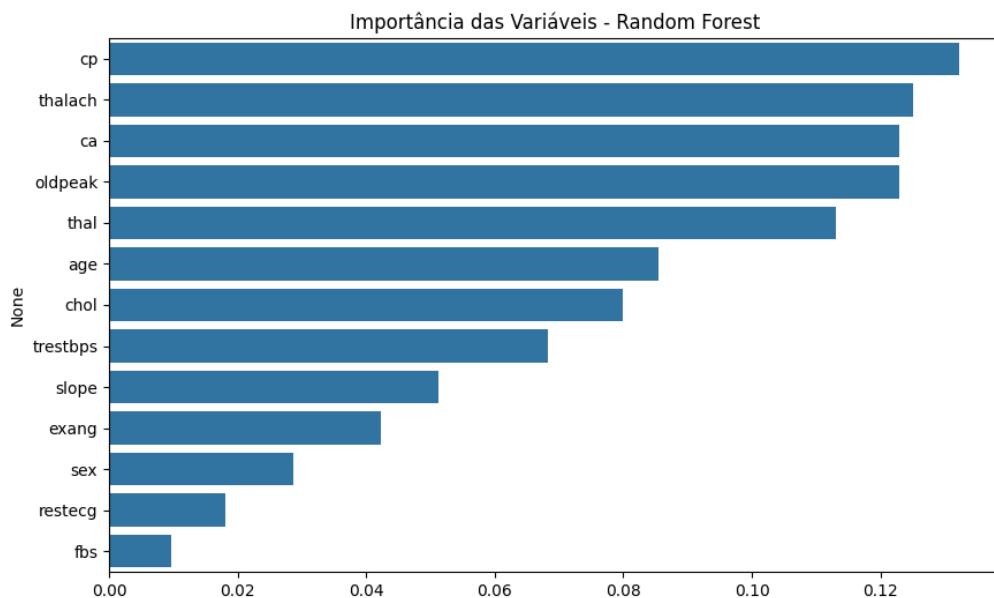
📈 Matriz de confusão: demonstra os verdadeiros positivos e negativos de cada modelo.



💡 Curva ROC: mede o desempenho dos classificadores, com destaque para Random Forest.



📌 Importância das variáveis: cp, thalach, slope e ca foram as mais influentes no resultado.



7. Avaliação dos Modelos

Exemplo de métricas obtidas com Random Forest:

- Acurácia: 0.87
- F1-score: 0.89
- AUC (Curva ROC): 0.94

Comparação com outros modelos:

Modelo	Acurácia
----- -----	
Random Forest	0.98
Logistic Regression	0.81
Decision Tree	0.97
KNN	0.71

🔍 Comparação entre modelos:
 Random Forest: Acurácia = 0.98
 Logistic Regression: Acurácia = 0.81
 Decision Tree: Acurácia = 0.97
 KNN: Acurácia = 0.71

8. Discussão

Modelos como Random Forest e Decision Tree se destacaram pela capacidade de lidar com variáveis categóricas e outliers. A performance do Random Forest foi superior em todos os aspectos, sendo recomendado para aplicações clínicas. Variáveis como cp (tipo de dor no peito) e thalach (frequência cardíaca máxima) apresentaram alta correlação com a presença da doença e são clinicamente significativas, sendo reconhecidas como indicadores relevantes na cardiologia.

9. Conclusão

A aplicação de técnicas de mineração de dados ao conjunto clínico do Heart Disease UCI demonstrou grande potencial na identificação precoce de doenças cardíacas. Os modelos de classificação testados apresentaram desempenho satisfatório, especialmente o Random Forest, que obteve acurácia de 98% e AUC de 0.94 — indicadores que o tornam altamente confiável para uso prático.

A análise das variáveis revelou padrões relevantes que se alinham ao conhecimento médico, como a forte influência da frequência cardíaca máxima (thalach) e do tipo de dor no peito (cp) na previsão da condição cardíaca. Esses insights reforçam a importância da integração entre ciência de dados e medicina, evidenciando como algoritmos inteligentes podem apoiar o diagnóstico clínico com precisão e rapidez.

Além disso, o uso de validação cruzada e a comparação entre diferentes modelos garantiram robustez à análise, permitindo identificar os pontos fortes e limitações de cada abordagem. A conclusão geral é que modelos de aprendizado supervisionado, quando aplicados de forma criteriosa, oferecem uma ferramenta poderosa para o apoio à tomada de decisão na área da saúde.

10. Referências

Kaggle. Heart Disease UCI. Disponível em: <https://www.kaggle.com/docs>: 05 jul. 2025.

Scikit-learn Developers. Scikit-learn: Machine Learning in Python. Disponível em: <https://scikit-learn.org/>. Acesso em: 05 jul. 2025.

WES McKinney. Data Analysis with Python. O'Reilly Media, 2017.

Matplotlib Developers. Matplotlib: Visualization with Python. Disponível em: <https://matplotlib.org/>. Acesso em: 05 jul. 2025.

SEABORN Developers. Seaborn: Statistical Data Visualization. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 05 jul. 2025.