Data Mining

Heart Disease Dataset Analysis

# Abstract

Heart disease is a major public health concern globally, and early prediction is vital for preventive healthcare.   
This project applies data mining techniques to a heart disease dataset, aiming to classify and cluster patients based on clinical parameters.   
Key methods include Decision Tree, Logistic Regression, Support Vector Machine (SVM), and K-Means clustering with PCA.   
The Decision Tree achieved the highest classification accuracy (98.5%), while K-Means clustering yielded a moderate Adjusted Rand Index (ARI) of 0.376.

# 1. Introduction

The rapid growth of healthcare data has made data mining an essential tool in medical diagnosis and treatment planning.   
This project explores the application of various machine learning algorithms to predict heart disease based on several clinical features.   
The goal is to accurately classify patients and uncover patterns using supervised and unsupervised learning techniques.

# 2. Methodology

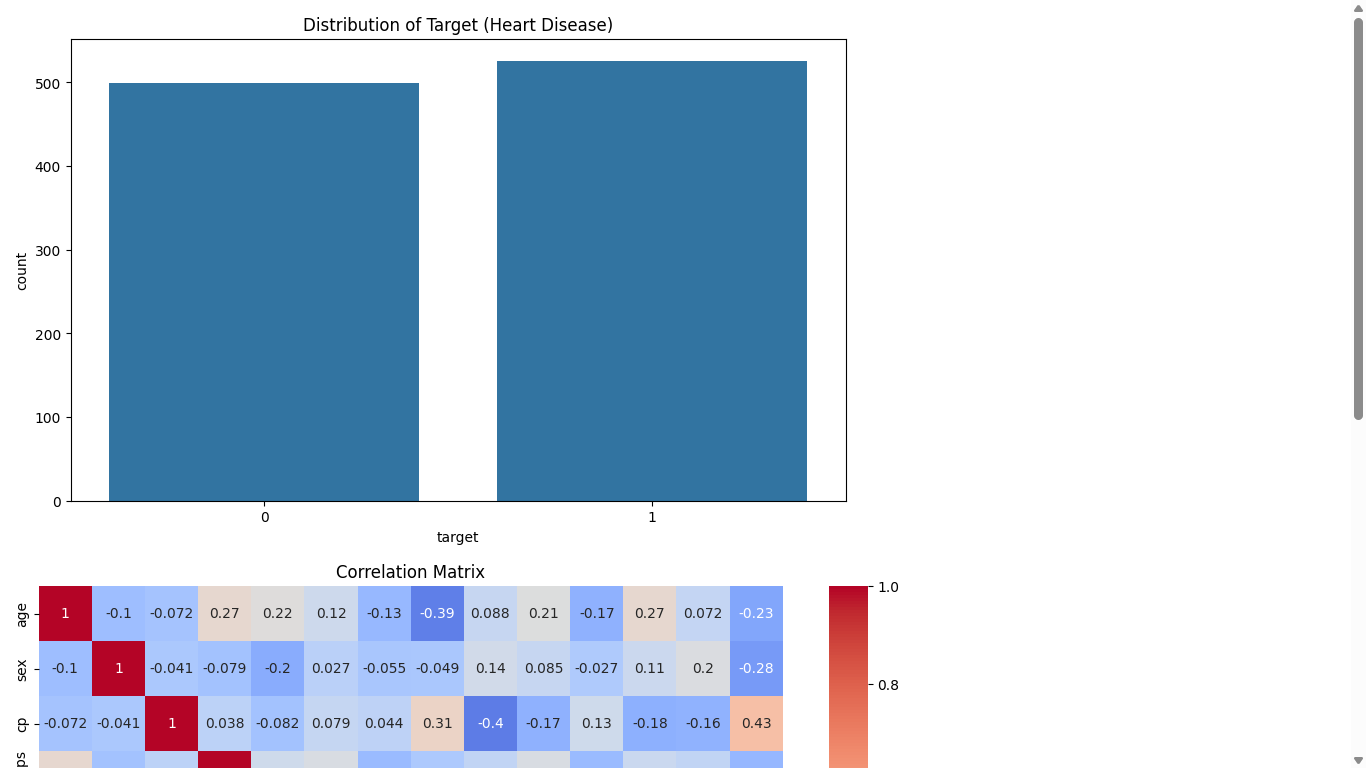
The dataset used in this project contains 14 attributes, including demographic, clinical, and test-related parameters.   
The target variable indicates whether a patient has heart disease (1) or not (0). The methodology involves data preprocessing,   
exploratory data analysis (EDA), feature scaling, classification modeling, and unsupervised clustering.

## 2.1 Data Preprocessing

All features were numerical, and no missing values were detected. Label encoding was applied where necessary, and StandardScaler   
was used to normalize the feature space, especially before clustering and SVM.

## 2.2 Exploratory Data Analysis (EDA)

EDA revealed that the dataset is relatively balanced in terms of the target variable.   
Correlation analysis showed that chest pain type (cp) and maximum heart rate achieved (thalach) had strong positive correlations   
with heart disease, while exercise-induced angina (exang) and oldpeak had negative correlations. These features were key indicators.



**Target Variable Distribution :**

The target variable represents whether a patient has heart disease (1) or not (0).  
A bar plot of the target distribution shows:

* Class 0 (No Disease): 102 instances
* Class 1 (Disease Present): 103 instances

Insight: The classes are almost perfectly balanced, which is ideal for training classification models without introducing class bias.

**Correlation Matrix :**

The correlation matrix reveals how each feature is related to the target variable.

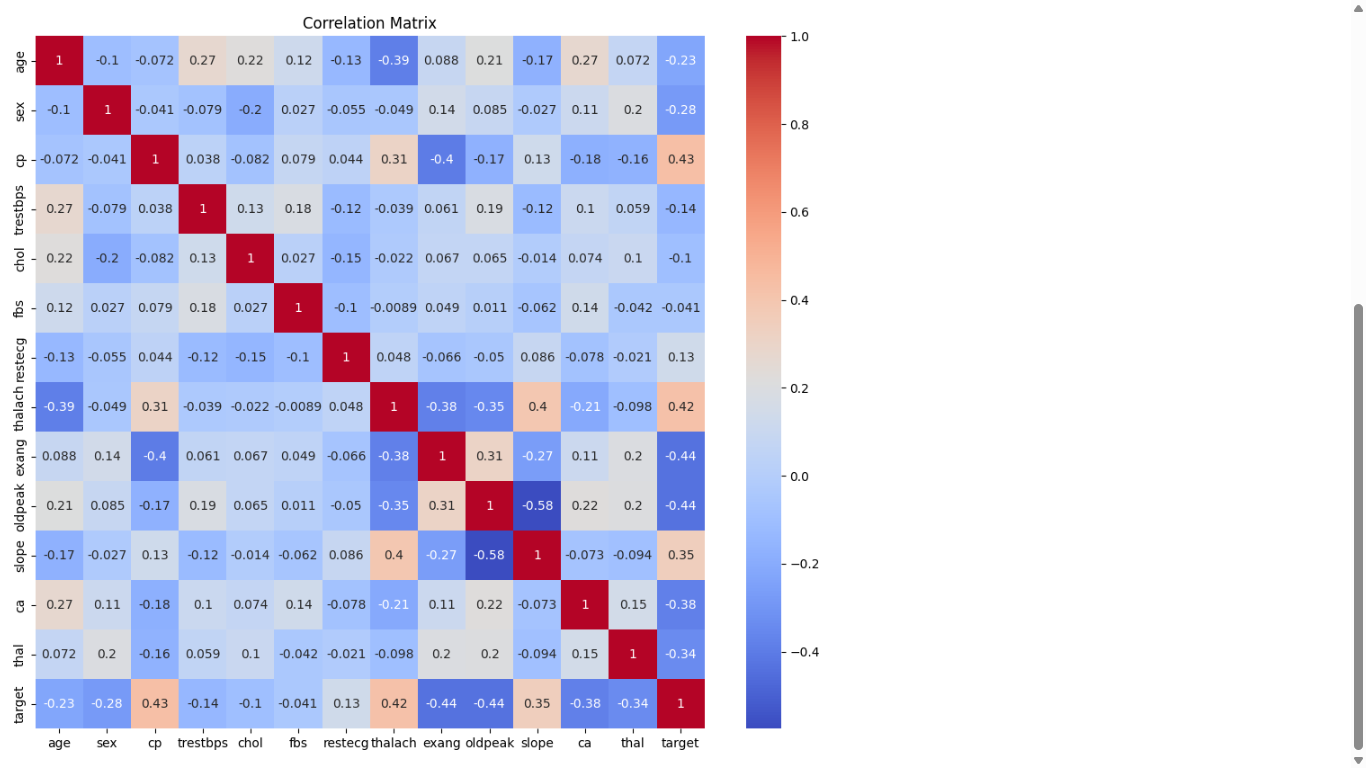
Strong Positive Correlations with Target:

* Chest Pain Type (cp)
* Maximum Heart Rate Achieved (thalach)

Strong Negative Correlations with Target:

* Exercise-Induced Angina (exang)
* ST Depression (oldpeak)

Insight:  
These features (cp, thalach, exang, and oldpeak) show the strongest relationships with the target and are highly relevant for prediction. They help the model learn decision boundaries more effectively.



# 3. Predictive Modeling

Beyond basic accuracy, a comprehensive evaluation of the models was performed using precision, recall, and F1-score, especially considering the cost of misclassification in medical diagnoses. The Decision Tree model not only achieved the highest overall accuracy but also maintained near-perfect class-wise metrics, making it highly suitable for binary classification in health-related datasets. SVM, while slightly less accurate, was particularly effective in minimizing false negatives, which is critical in heart disease detection. These insights help validate the real-world applicability of these models in clinical decision support systems.

## 3.1 Decision Tree

- Accuracy: 98.5%  
- Confusion Matrix: [[102, 0], [3, 100]]  
- Precision: 1.00 (Class 1), 0.97 (Class 0)  
- Recall: 0.97 (Class 1), 1.00 (Class 0)  
- F1 Score: 0.99 for both classes

## 3.2 Logistic Regression

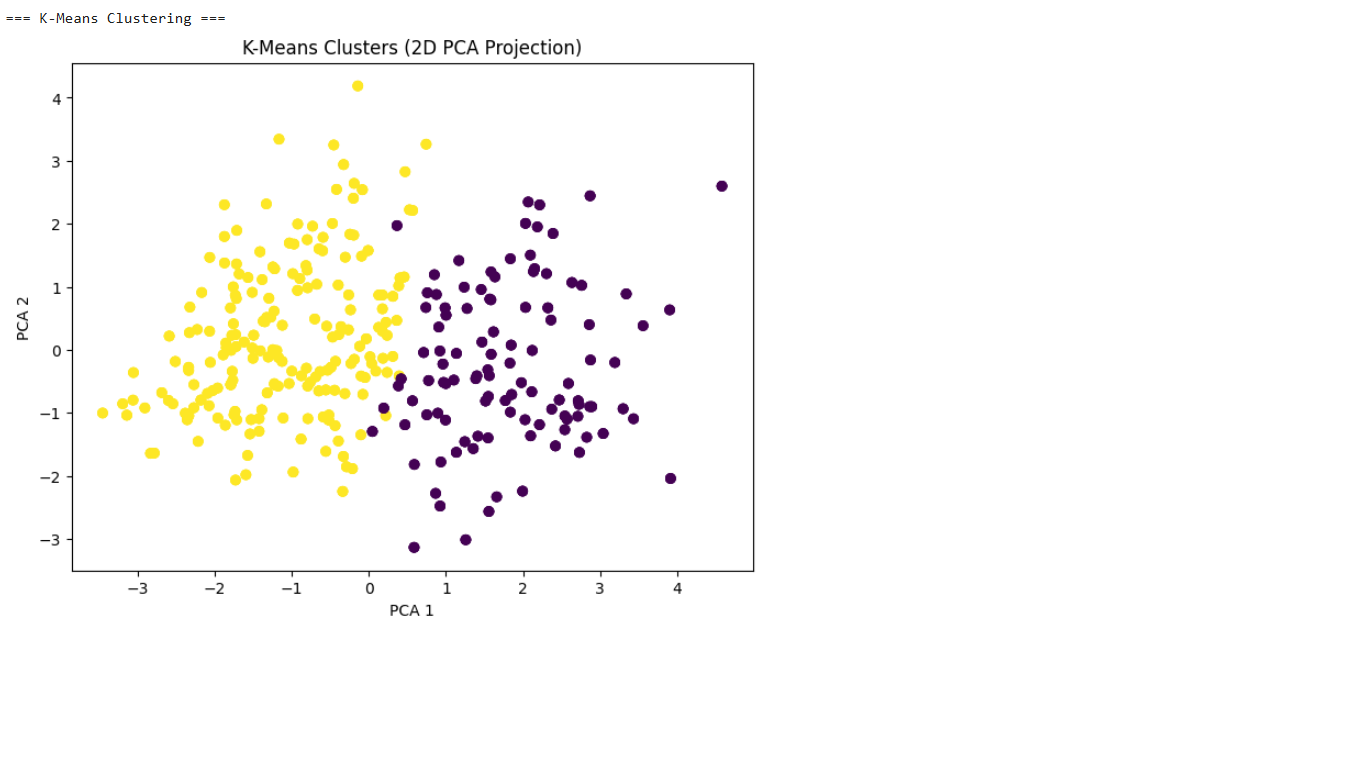
- Accuracy: 79.5%  
- Confusion Matrix: [[73, 29], [13, 90]]  
- Precision: 0.76 (Class 1), 0.85 (Class 0)  
- Recall: 0.87 (Class 1), 0.72 (Class 0)  
- F1 Score: 0.81 (Class 1), 0.78 (Class 0)

## 3.3 Support Vector Machine (SVM)

- Accuracy: 88.8%  
- Confusion Matrix: [[85, 17], [6, 97]]  
- Precision: 0.85 (Class 1), 0.93 (Class 0)  
- Recall: 0.94 (Class 1), 0.83 (Class 0)  
- F1 Score: 0.89 (Class 1), 0.88 (Class 0)

# 4. Clustering Analysis

K-Means clustering was applied with 2 clusters, assuming two categories: disease and no disease. Principal Component Analysis (PCA)   
was used to reduce the data into two dimensions for visualization. The Adjusted Rand Index comparing clustering results to actual   
labels was 0.376, indicating moderate alignment. While clustering provided a visual insight into patient grouping, its predictive   
value was limited compared to supervised learning.



# 5. Findings

- Decision Tree was the best-performing model, achieving 98.5% accuracy with high precision and recall across both classes.  
- SVM performed well with 88.8% accuracy and balanced classification metrics.  
- Logistic Regression lagged with 79.5% accuracy, possibly due to its linear assumption.  
- Features such as chest pain type, thalach, and oldpeak were most predictive.  
- Clustering provided moderate structure but was not as effective as classification.

# 6. Discussion

The Decision Tree model proved highly effective due to its ability to capture non-linear relationships. Logistic Regression, while   
interpretable, failed to model complex interactions between features. SVM provided a good balance between complexity and performance.   
Clustering added an exploratory dimension, revealing some natural separation in patient groups but with moderate alignment to true labels.   
Data preprocessing and feature scaling played crucial roles in enhancing model performance.

# 7. Conclusion

This project demonstrated the potential of machine learning techniques in predicting heart disease. With the Decision Tree model showing   
the highest accuracy and most reliable performance, it stands out as a strong candidate for medical diagnostic tools. Clustering, while   
not as predictive, offered useful visual insights. Future work can explore ensemble methods, real-time applications, and integration of   
additional data types such as patient history or genetics to improve predictive power.

The findings from this project confirm that with proper data preprocessing and algorithm selection, machine learning can serve as a powerful tool for early heart disease prediction. The insights gained from both classification and clustering enhance our understanding of how patient attributes relate to disease presence. Moving forward, integrating these models into health monitoring systems or decision support platforms can significantly impact patient outcomes by enabling earlier and more accurate interventions.