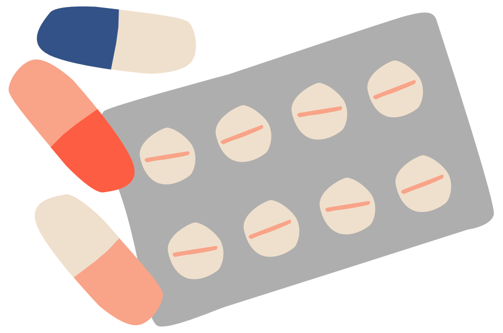


Heart Disease Prediction: A Data-Driven Investigation

Team Members: Aly Hassan, Alaa Shaban, Kenzy Ahmed, Mohamed Ehab
Supervisor: Dr. Mohamed Taher

01. Introduction

Heart disease is a leading cause of death worldwide. Detecting it early through data can save lives. This project investigates how clinical features contribute to disease risk using statistical and machine learning techniques—led by our data-savvy sleuth, Detective Data.

02. Objective

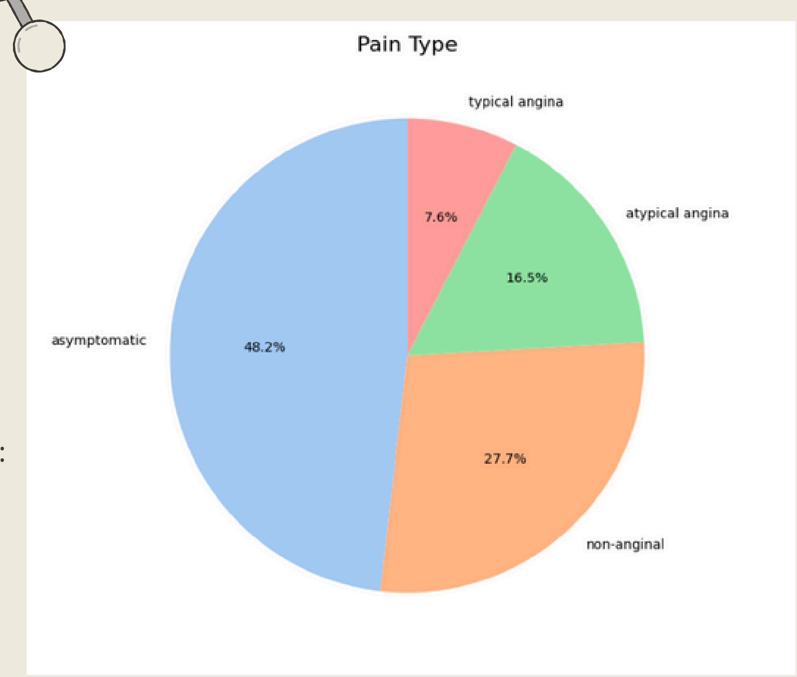
To identify the most predictive clinical indicators of heart disease severity using a combination of statistical testing and machine learning. We aim to filter out noise and focus on the variables with the strongest influence on patient outcomes.

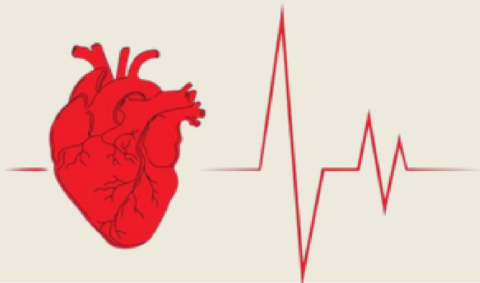
03. Hypothesis

- H₀ (Null Hypothesis): There is no significant association between the selected features and heart disease severity.
- H₁ (Alternative Hypothesis): At least one feature is significantly associated with heart disease severity.

04. Dataset

- We used the UCI Heart Disease Dataset (920 patient records).
- After data cleaning and variable selection, we analyzed 299 complete records with the most relevant features:
- Max Heart Rate
- Oldpeak (ST Depression)
- Exercise-Induced Angina
- Artery Block Count
- ST Slope
- Disease Severity (Target)





05. Methodology

Tools
Python, pandas, seaborn, sklearn

Statistical Analysis
T-tests, chi-square tests, correlation analysis, and ANOVA tests

Machine Learning Models:
- SVM (RBF Kernel) – Best performance (80% accuracy)
- Logistic Regression – Strong and consistent
- Random Forest – Robust, ensemble-based
- Decision Tree – Simple, but prone to overfitting

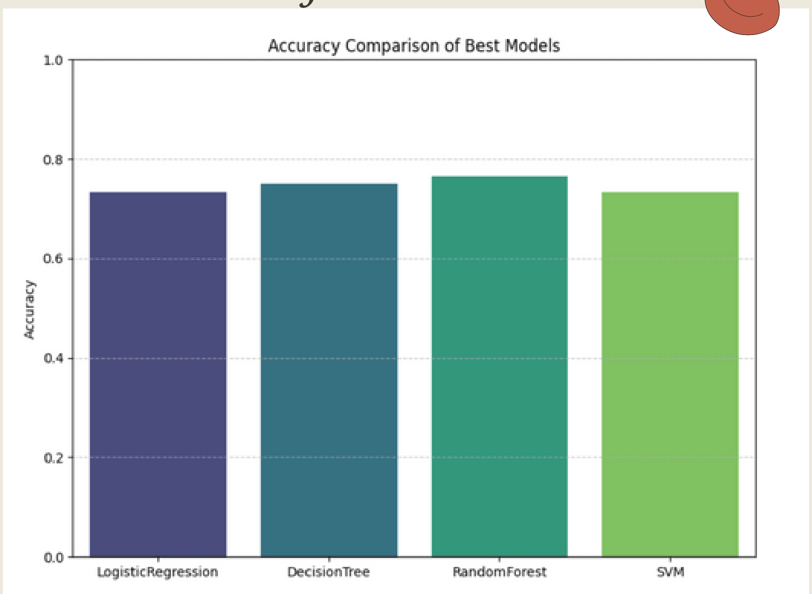
Data Preprocessing
- Removed missing/inconsistent entries
- Renamed columns for clarity
- Converted data types (e.g., blood sugar to Boolean)
- Dropped irrelevant features

Validation
5-fold cross-validation with GridSearchCV

07. Findings

- Artery Block Count: Strongest predictor (χ^2, $p < 0.00001$)
- Oldpeak (ST Depression): Highly significant ($t = 7.87$, $p < 0.001$)
- Max Heart Rate: Lower in diseased patients (−20 bpm on average)
- Exercise-Induced Angina: Strong positive correlation
- Slope: Significant per ANOVA ($p = 4.75e-12$)
- Cholesterol & Blood Pressure: Weak predictors, removed due to low correlation
- Note: Most data is from Cleveland → Possible location bias

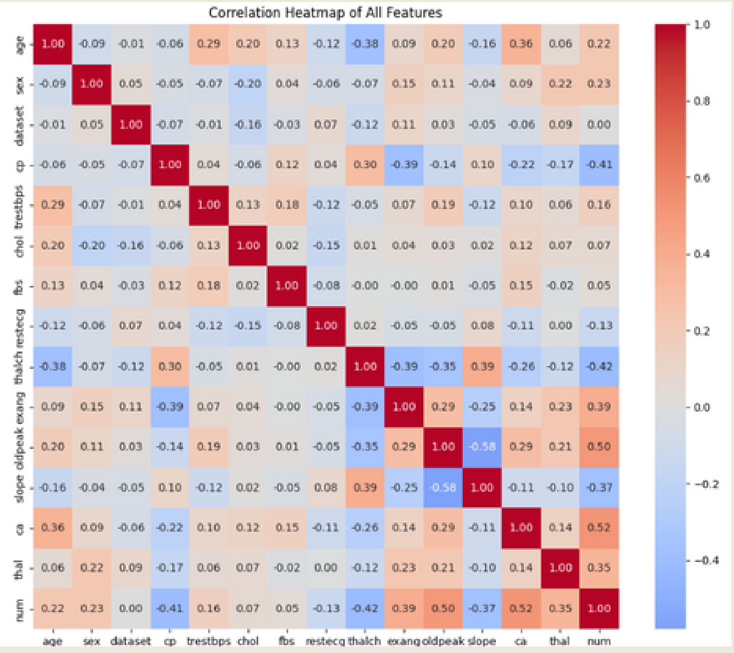
08. Model Performance



Compares the accuracy scores of four machine learning models:
- Logistic Regression
- Decision Tree
- Random Forest
- SVM (Support Vector Machine)

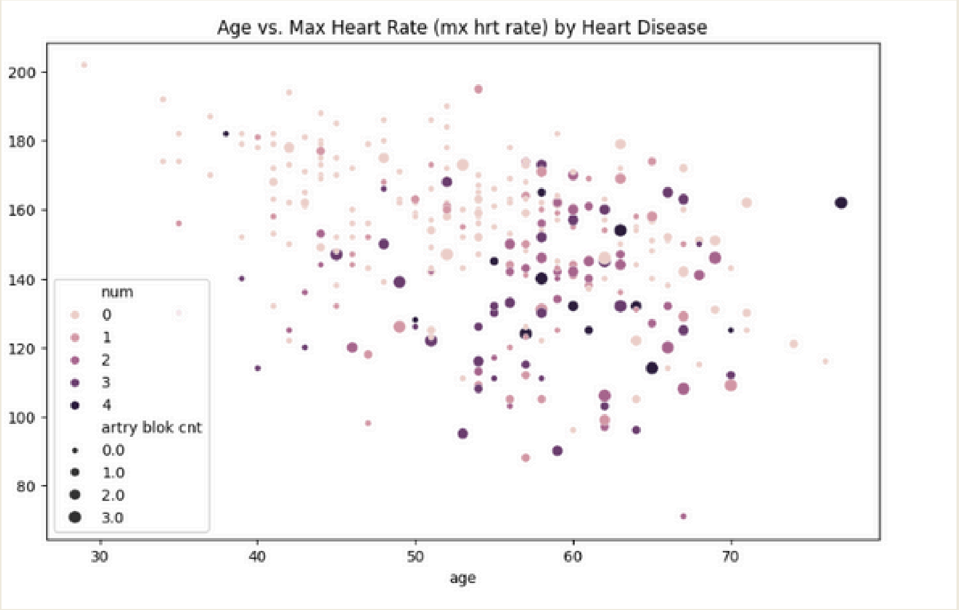
Insight:
- Random Forest has the highest accuracy (around 0.76).
- Decision Tree is slightly lower.
- Logistic Regression and SVM are tied or nearly tied for the lowest.

06. Analysis



This heatmap highlights the strongest correlations with heart disease severity.
🔴 Artery block count (+0.52), Oldpeak (+0.50), and Max heart rate (−0.42) were the most influential features.
Weak correlations from variables like cholesterol and blood sugar led to their removal.

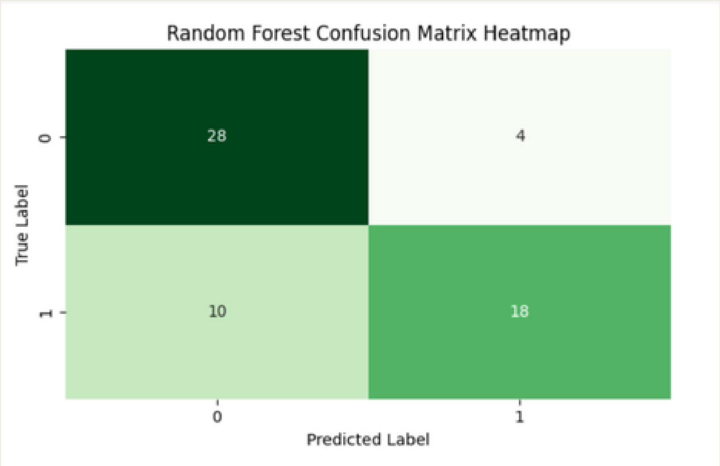
This scatter plot shows that patients with heart disease tend to have a significantly lower max heart rate.
The downward trend supports its role as a key negative predictor.



The graph shows the performance of the Random Forest classifier.

Matrix Values:
- True Negatives (TN): 28 (correctly predicted 0s)
- False Positives (FP): 4 (predicted 1 but was actually 0)
- False Negatives (FN): 10 (predicted 0 but was actually 1)
- True Positives (TP): 18 (correctly predicted 1s)

Random Forest did well in classifying class 0, but misclassified 10 samples from class 1.



09.Conclusion

The investigation confirms that artery block count, oldpeak, max heart rate, exercise-induced angina, and slope are significantly associated with heart disease severity.
Weak predictors like cholesterol, blood sugar, and resting blood pressure were dismissed.
✅ Hypothesis H₁ accepted: At least one attribute is significantly associated with heart disease.