Predicting Heart Disease:
A Comparative Analysis of Machine Learning Techniques on the Cleveland Heart Disease Dataset

Athul Sreejith | s3906868 Manasvi Kumar | s3938425 Yudie Zheng | s3817598

Universiteit Leiden
The Netherlands

Index

Data Set

- The dataset used in this analysis is the Heart Disease dataset obtained from the UCI Machine Learning Repository.
- It contains 303 instances with 13 attributes, which include demographic and clinical features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia.
- The target variable indicates the presence or absence of heart disease.
  - {0: 164, 1: 139}

Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age		years	no
sex	Feature	Categorical	Sex			no
cp	Feature	Categorical				no
trestbps	Feature	Integer		resting blood pressure (on admission to the hospital)	mm Hg	no
chol	Feature	Integer		serum cholestoral	mg/dl	no
fbs	Feature	Categorical		fasting blood sugar > 120 mg/dl		no
restecg	Feature	Categorical				no
thalach	Feature	Integer		maximum heart rate achieved		no
exang	Feature	Categorical		exercise induced angina		no
oldpeak	Feature	Integer		ST depression induced by exercise relative to rest		no
slope	Feature	Categorical				no
ca	Feature	Integer		number of major vessels (0-3) colored by flourosopy		yes
thal	Feature	Categorical				yes
num	Target	Integer		diagnosis of heart disease		no

Research Questions:

1. What factors are most strongly associated with heart disease?

2. Can we accurately predict heart disease based on the given features?

Methodology

Data Preprocessing :

1. Missing Values:

- The dataset had missing values in the 'ca' and 'thal' columns.
- We filled these missing values using the most frequent value (mode) in each column.

2. Normalization:

- We standardized continuous variables.
- This means we adjusted them to have a mean of zero and a standard deviation of one.
- Standardization ensures that each variable contributes equally to the analysis.

3. Encoding Categorical Variables:

•We converted categorical variables into numerical format.

•We used one-hot encoding for this process.

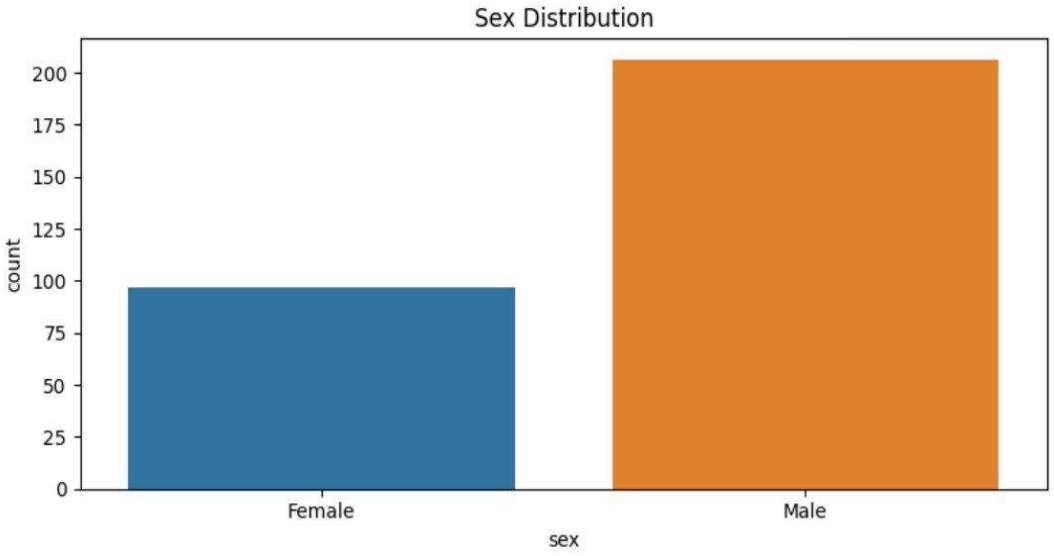
•This method creates new columns for each category with binary values (0 or 1).

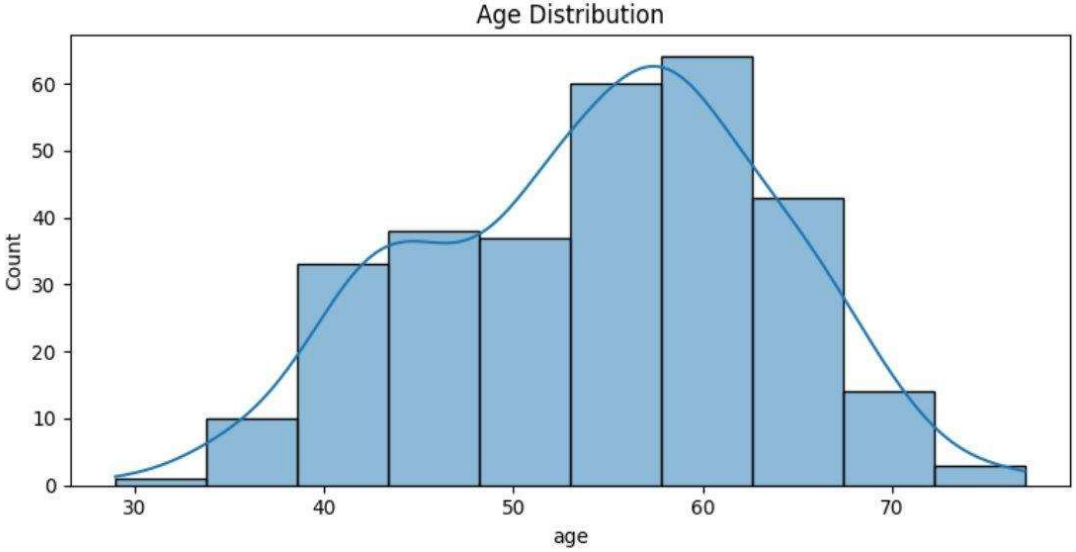
Exploratory Data Analysis (EDA)

- **Visualizations:** Various visualizations were created to understand the distribution and relationships between features.

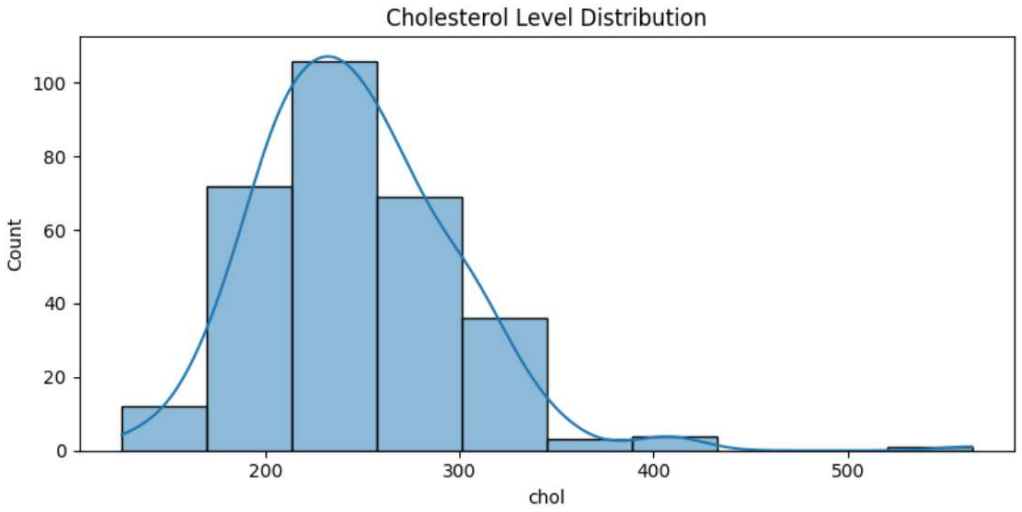
  - **Histograms:** To visualize the distribution of age, sex, and cholesterol levels.
  - **Scatter Plots:** To explore the relationship between age and cholesterol with the presence of heart disease.
  - **Correlation Matrix:** To identify the strength and direction of relationships between continuous variables.
  - **Box Plots:** To detect outliers and understand the distribution of continuous variables.
  - **Pie Charts:** To visualize the distribution of categorical variables like sex, chest pain type, exercise-induced angina, thalassemia, and their association with heart disease

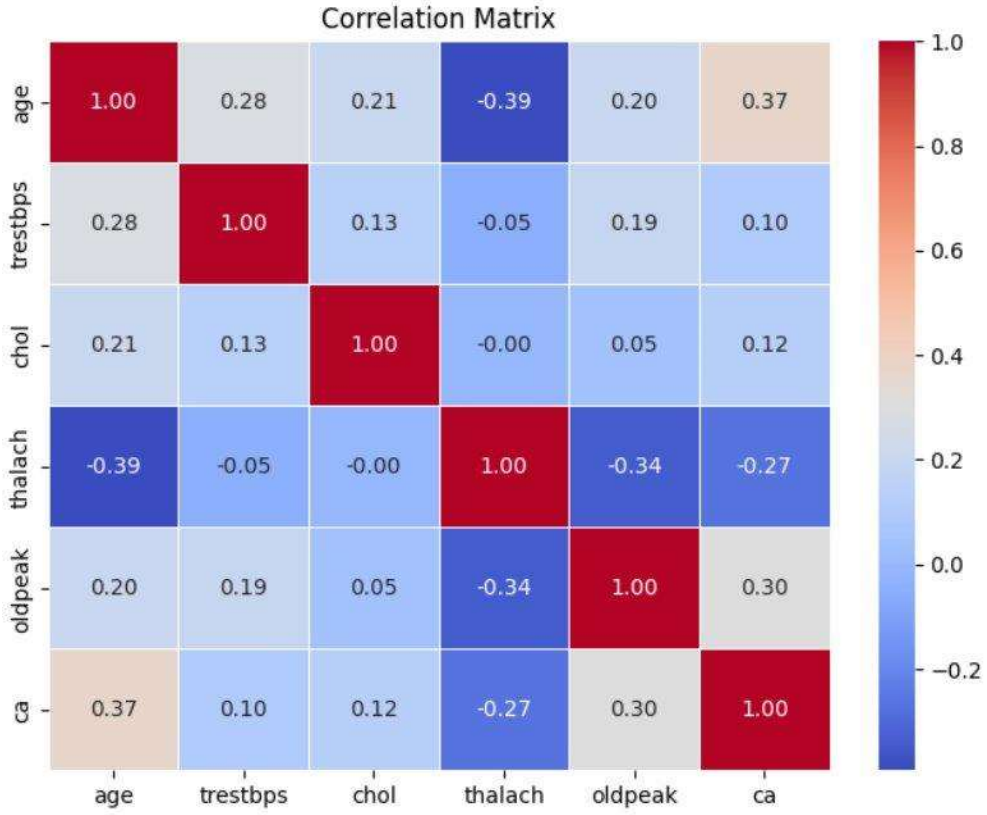
Age and Sex distribution





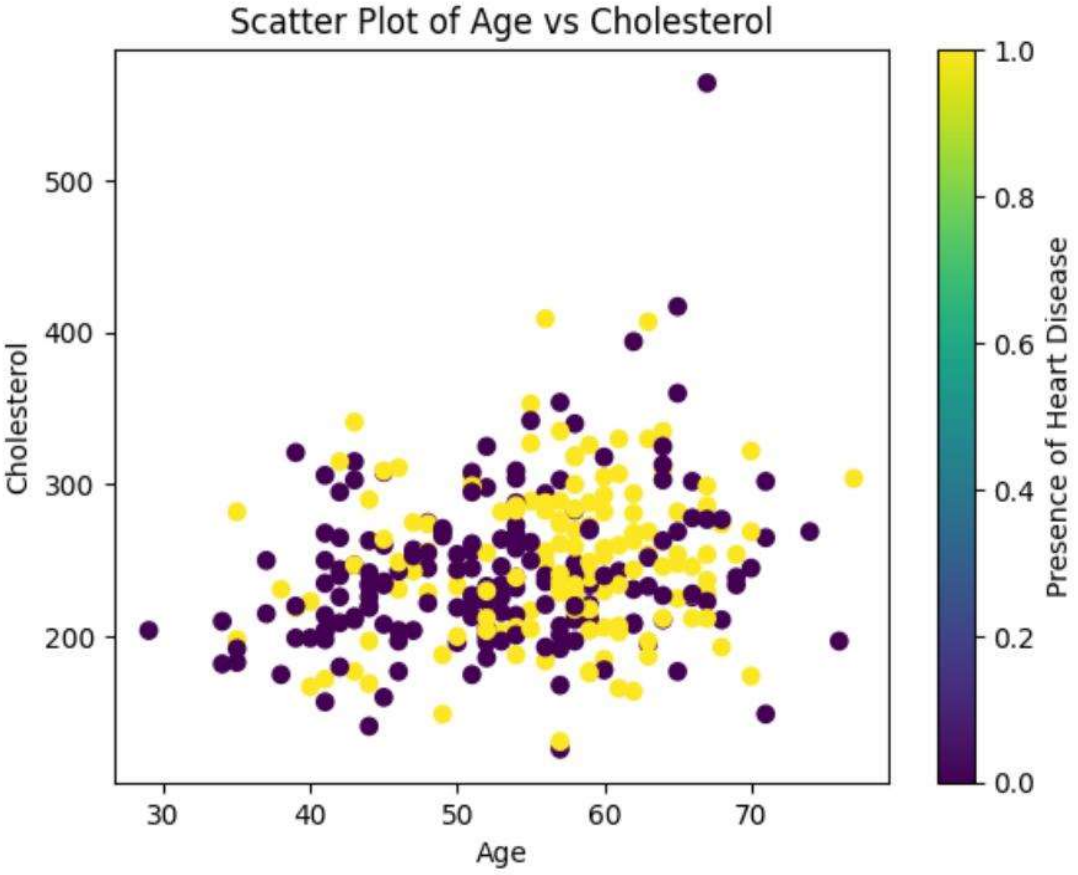
Cholesterol level Distribution



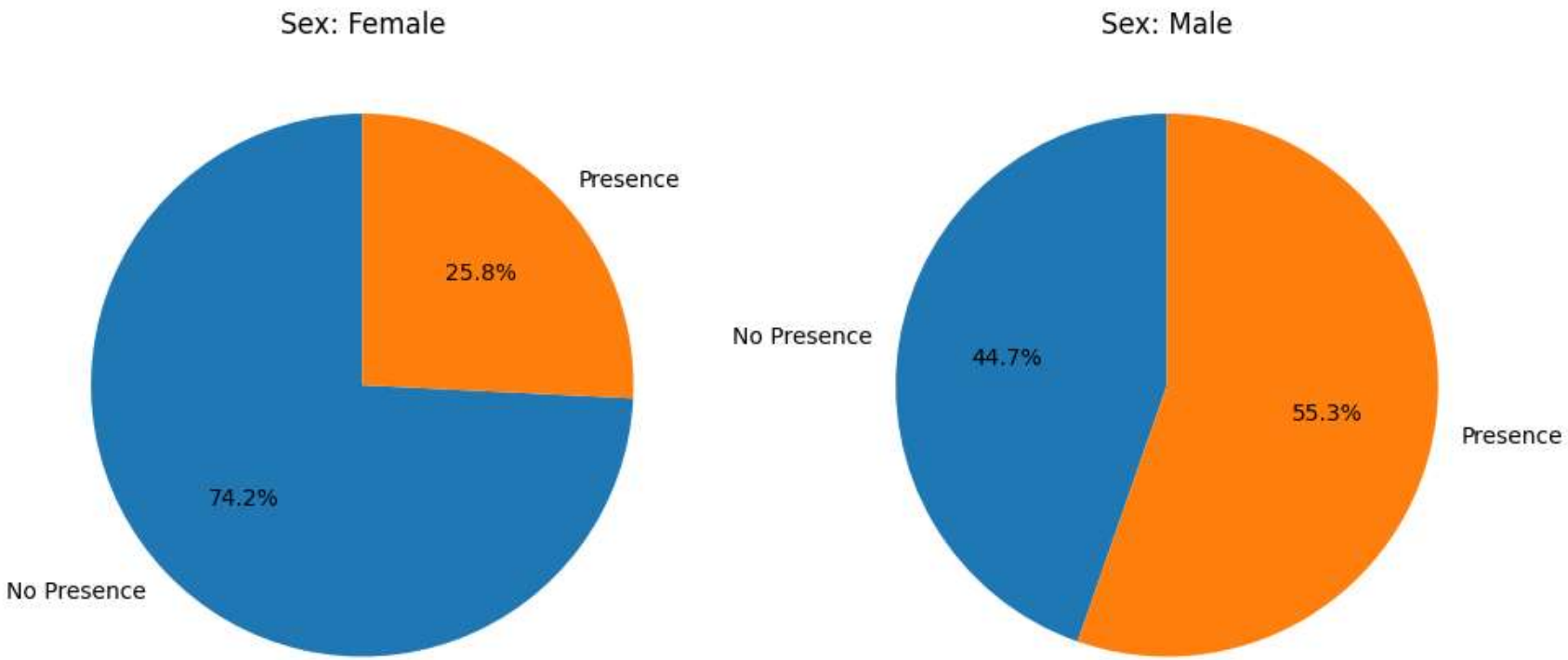


Ca: Number of major vessels (0-3) colored by flourosopy
Trestbps: Resting blood pressure (on admission to the hospital)
Chol: Cholesterol
Thalach: Maximum heart rate achieved
Oldpeak: ST depression induced by exercise relative to rest

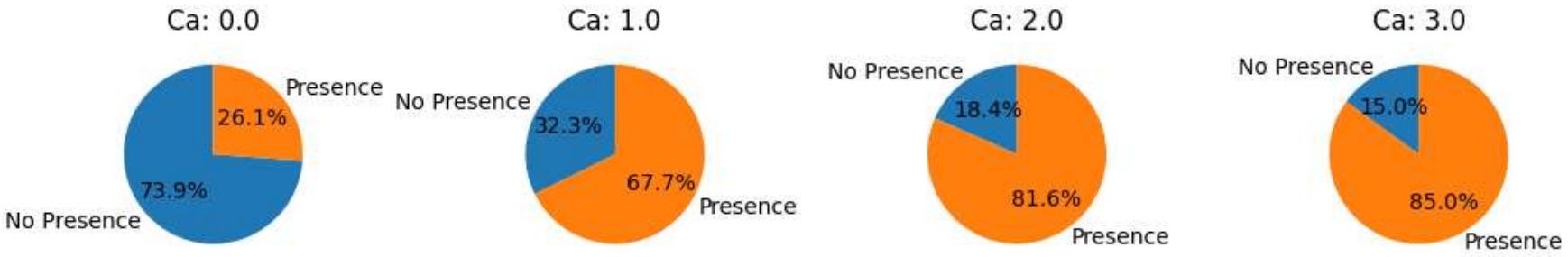
Heart disease relation with Cholesterol



Men More Prone to Heart Disease

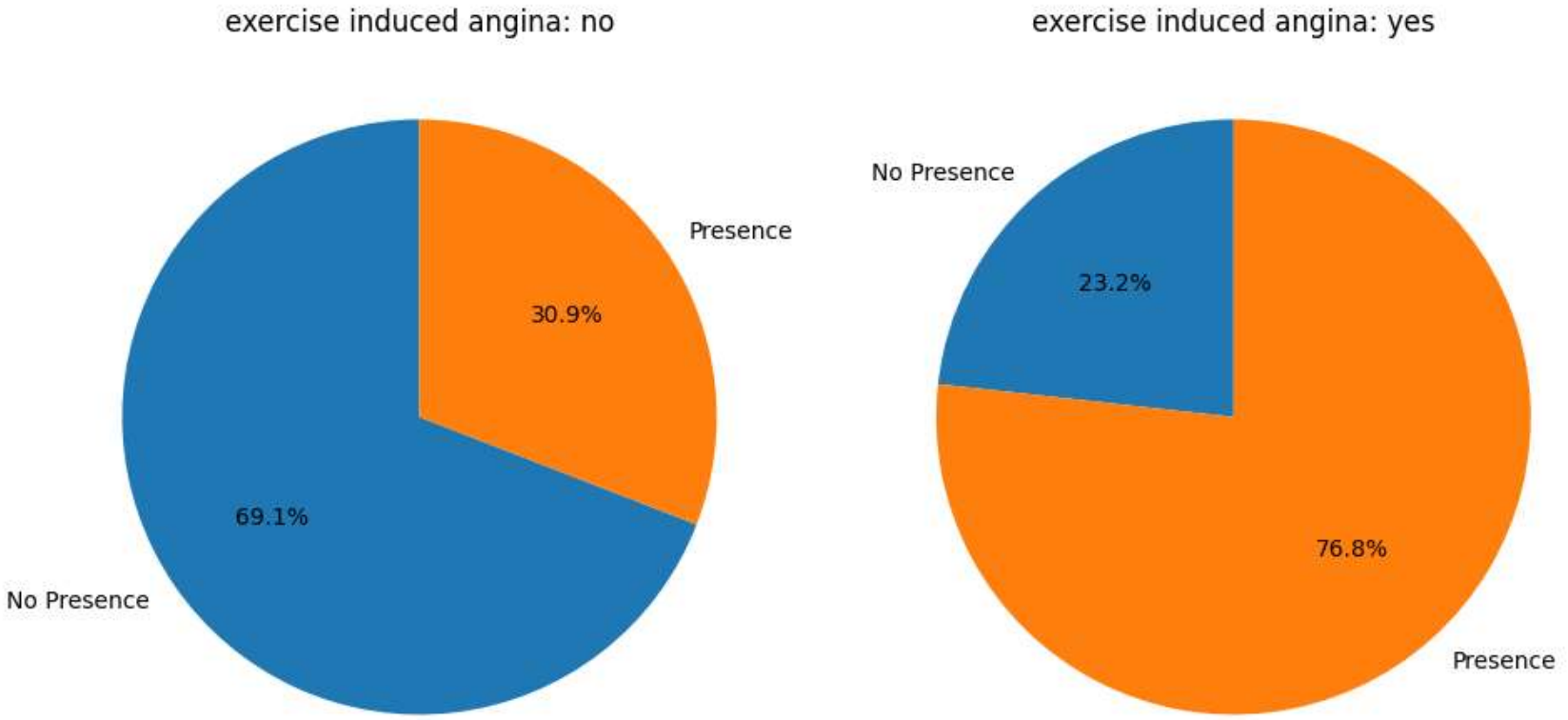


Severe Vessel Blockage or Narrowing Increases Heart Disease Risk

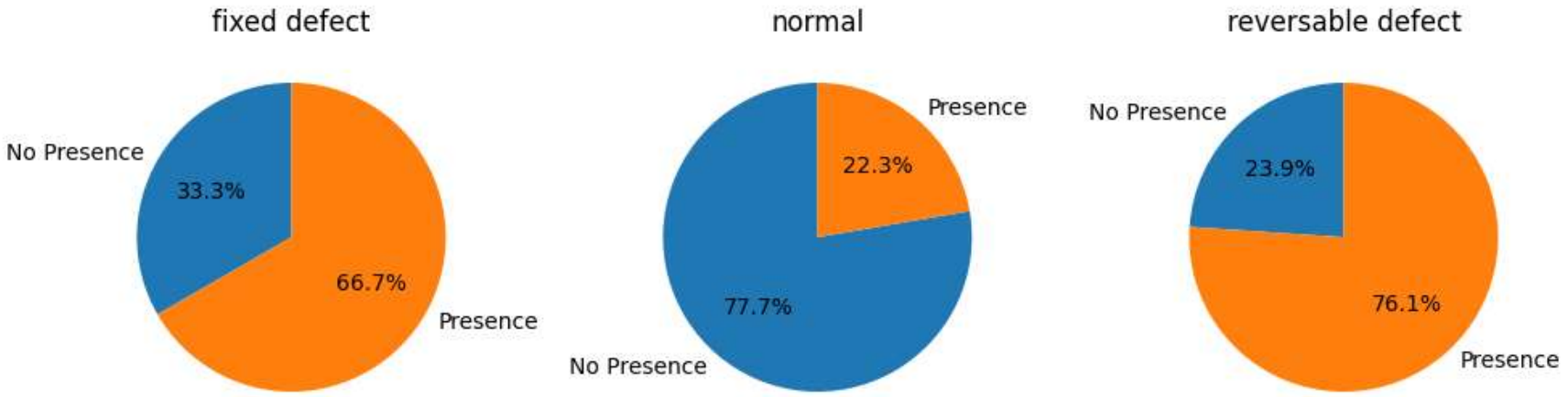


Ca: Number of major vessels (0-3) colored by fluoroscopy
Indicating the blockages or narrowing level of the vessels

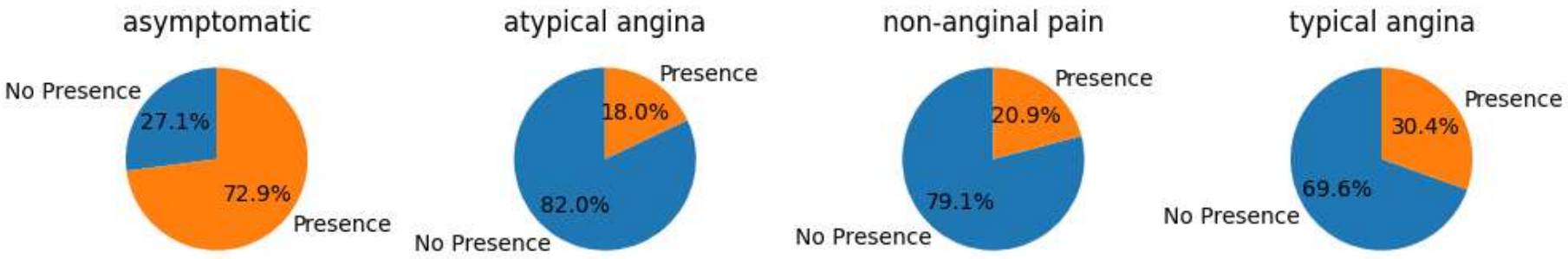
Exercise-Induced Angina: A Potential Indicator of Heart Disease



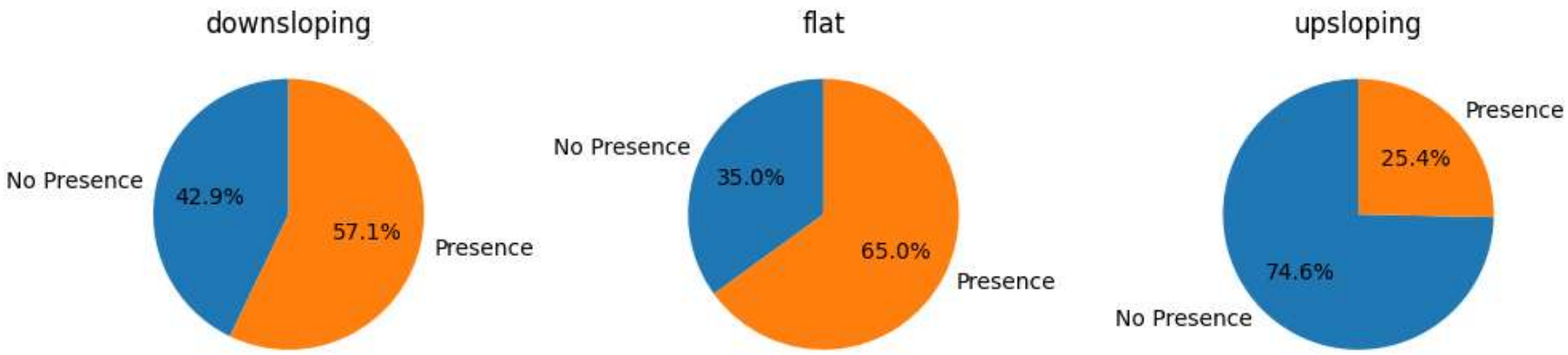
Thalassemia Defects Linked to Increased Heart Disease Risk



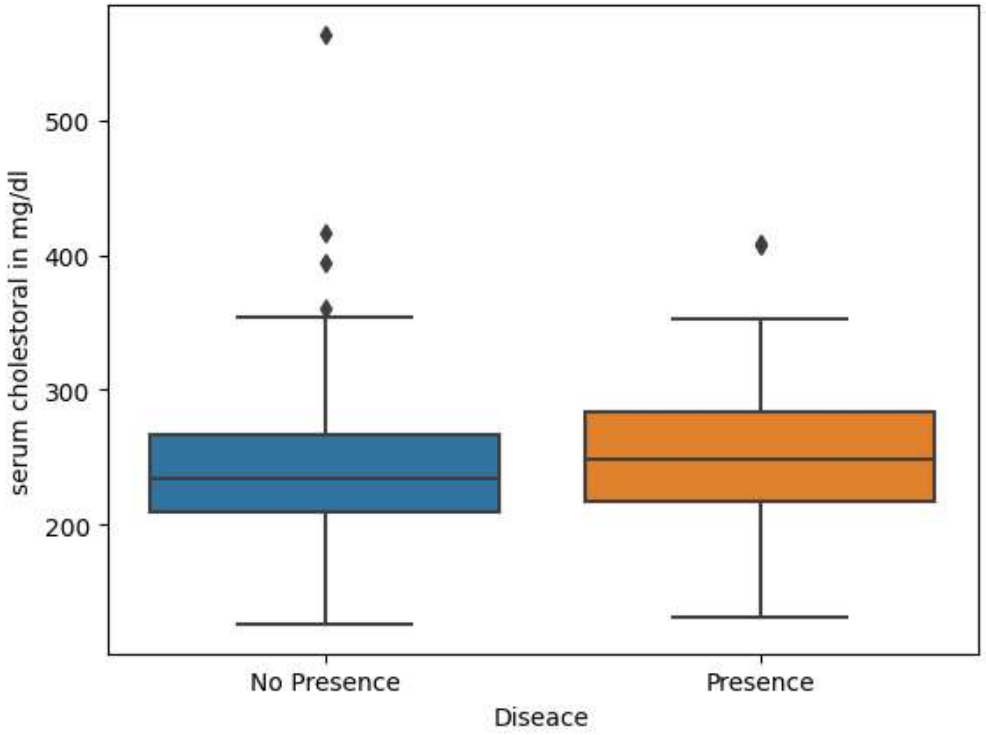
Asymptomatic Chest Pain More Likely to Indicate Heart Disease



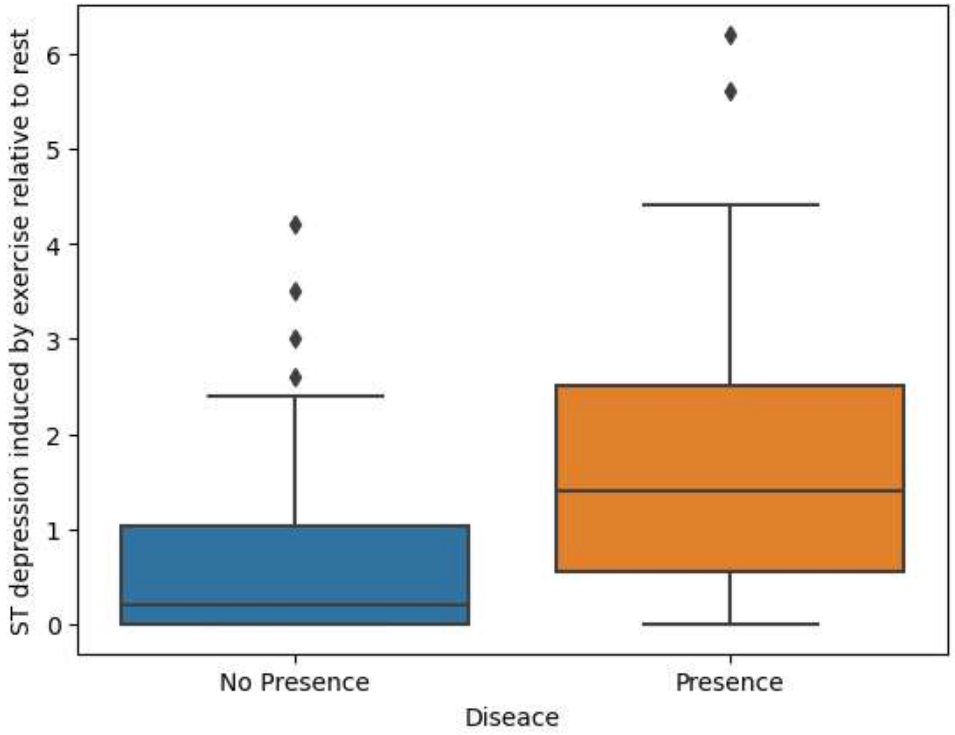
Flat and Downsloping ST Segments: Potential Indicators of Heart Disease



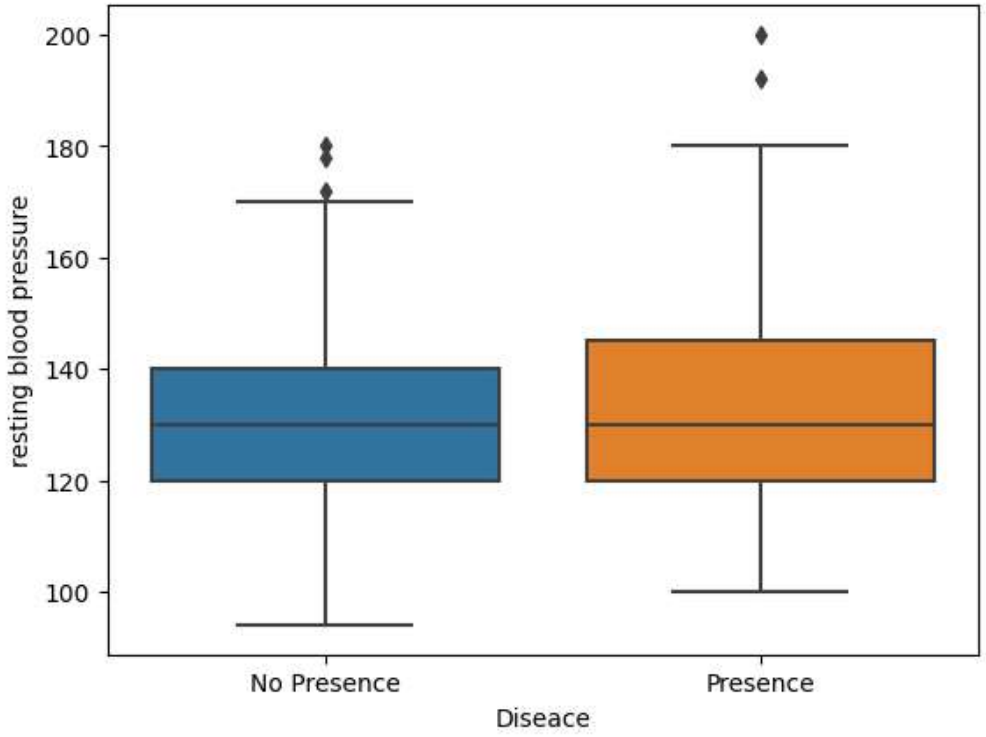
Serum Cholesterol: Potentially Unrelated to Heart Disease Risk



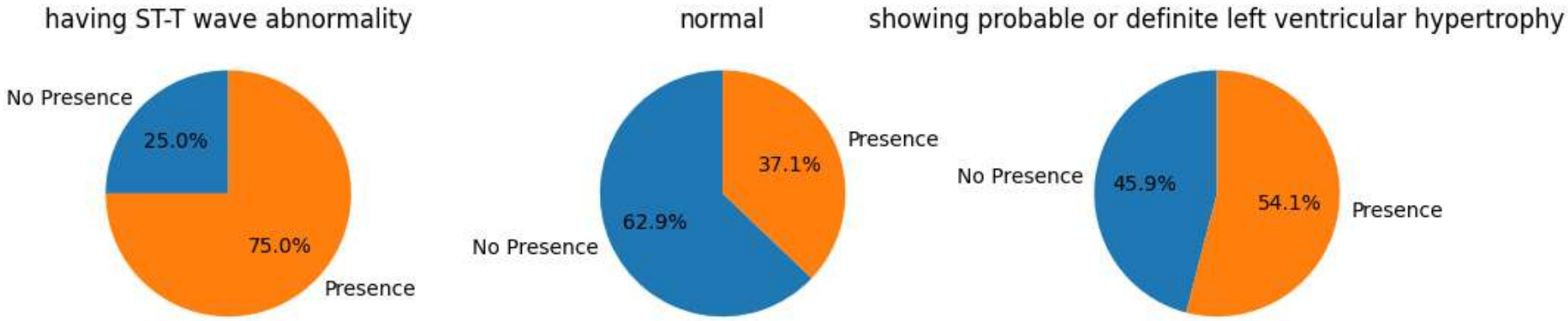
Increased Exercise-Induced ST Depression Signals Higher Heart Disease Risk



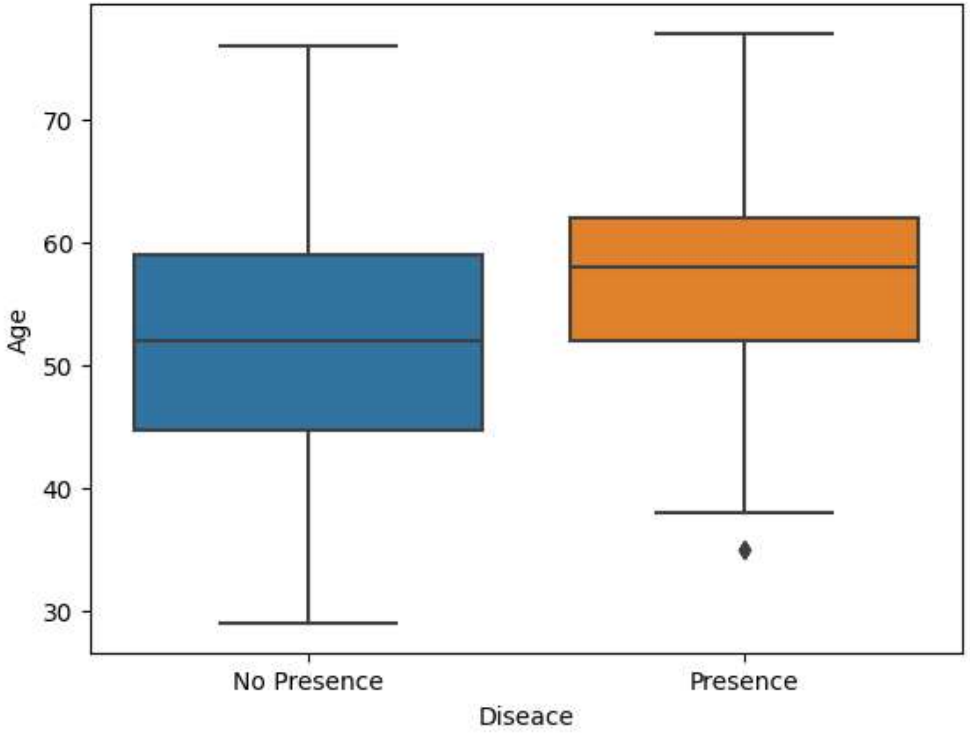
Resting Blood Pressure: Potentially Unrelated to Heart Disease Risk



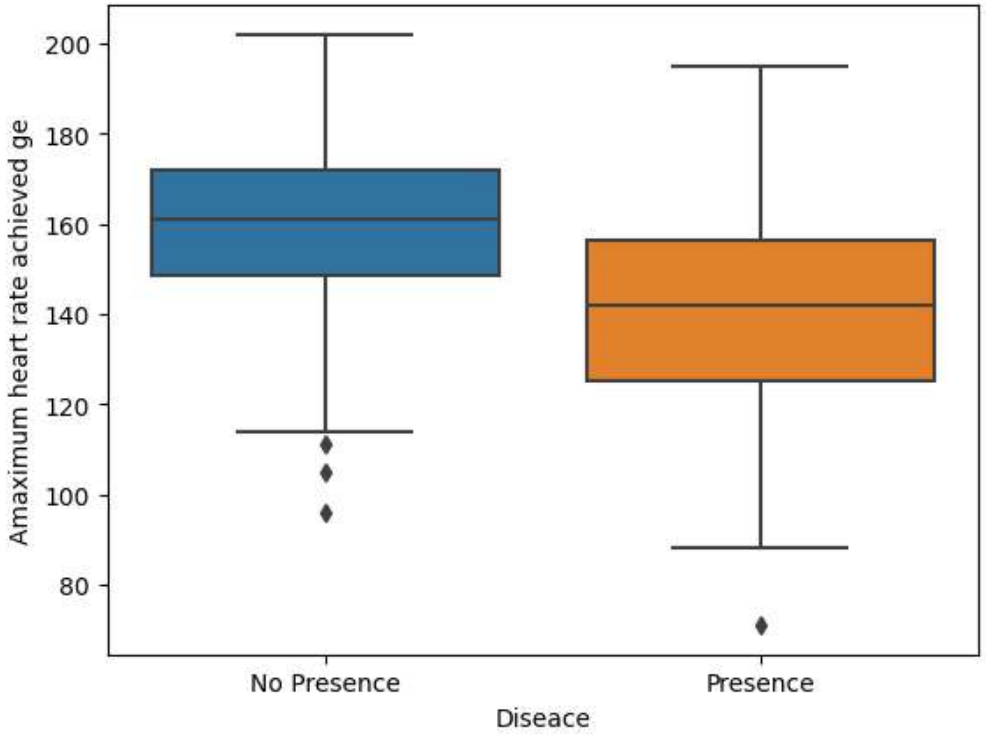
Abnormal Resting Electrocardiographic Results: Indicator of Heart Disease



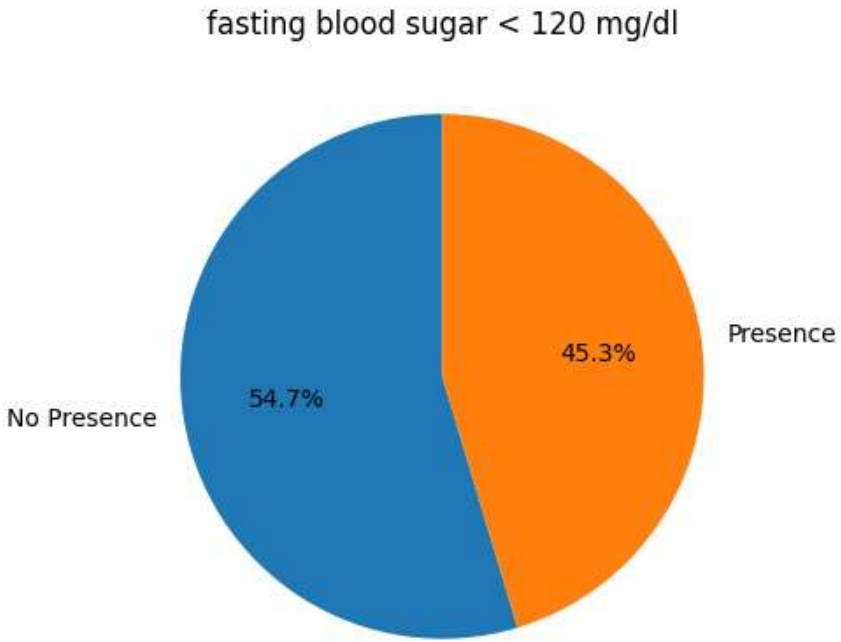
Higher Age Linked to Increased Risk of Heart Disease

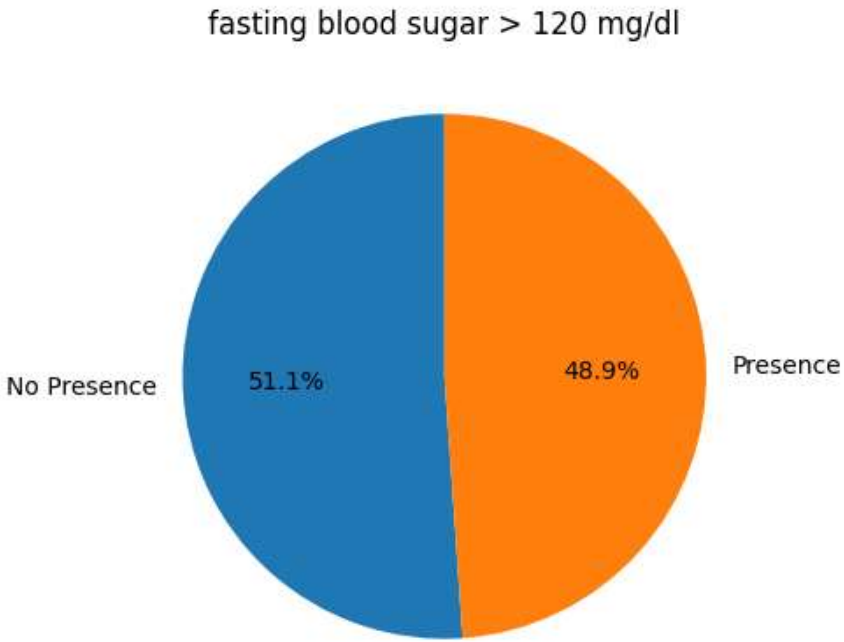


Higher Max Heart Rate Increases Heart Disease Risk



Fasting Blood Sugar: Not an Indicator of Heart Disease





Machine Learning Models

We chose 3 methods for the classification tasks which have been mentioned as follows:-

- **Logistic Regression:** Chosen for its simplicity and interpretability to predict the presence of heart disease.

- **Random Forest :** Chosen this classifier because it is a powerful and versatile machine learning algorithm which can handle both numerical and categorical data. Also It reduces overfitting by averaging multiple decision trees.
- Metrics used: Accuracy, precision, recall, and F1-score.

- **Support Vector Machine:** Chosen for its ability to use a subset of training points in the decision function (called support vectors), so it is also memory efficient. Also, it is very verastile, as different kernel functions can be specified for the decision function

- **Evaluation Metrics:** We used the following metrics ofor evaluating our methods and these metrics were applied to all our methods.:-
- Accuracy, precision, recall, and F1-score were used to evaluate the performance of the model.

- **HyperParameter Optimization:** We used GridSearch Cross Validation to find the best set of hyperparmeters for all our models

Logistic Regression

$$h_\theta(x) = g(\theta^T X)$$

$$h_\theta(x) = P(y = 1 | x; \theta)$$

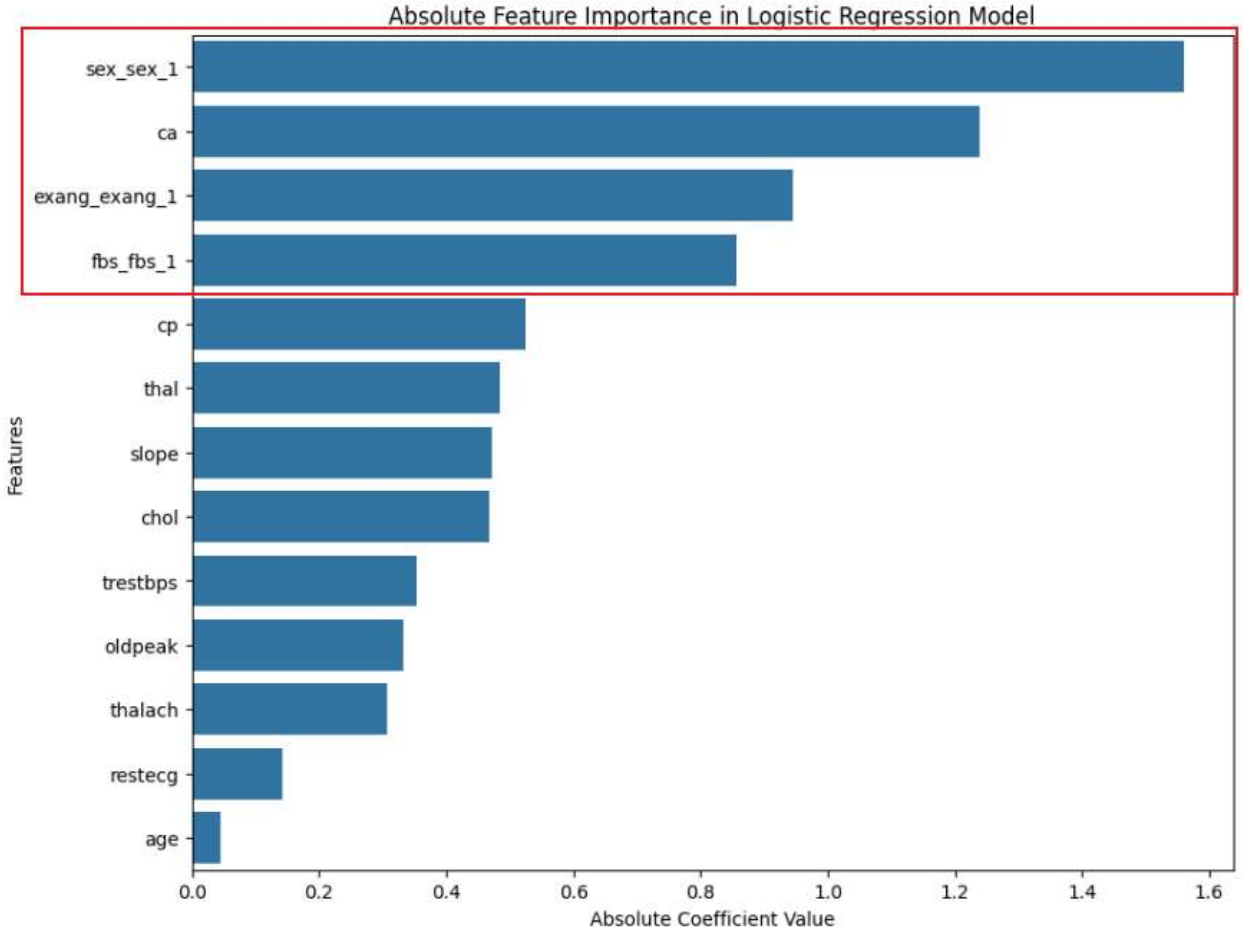
$$J(\theta) = -\frac{1}{m} \sum_{m}^{i=1} [y^i log(h_\theta(x^i)) + (1 - y^i) log(1 - h_\theta(x^i))] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$

Results

- CV (lambda = 0.01, 0.1, 1, 10)
- Lambda = 10

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

Feature Importance



- Sex
- blockages or narrowing level of vessels
- exercise induced angina
- fasting blood sugar

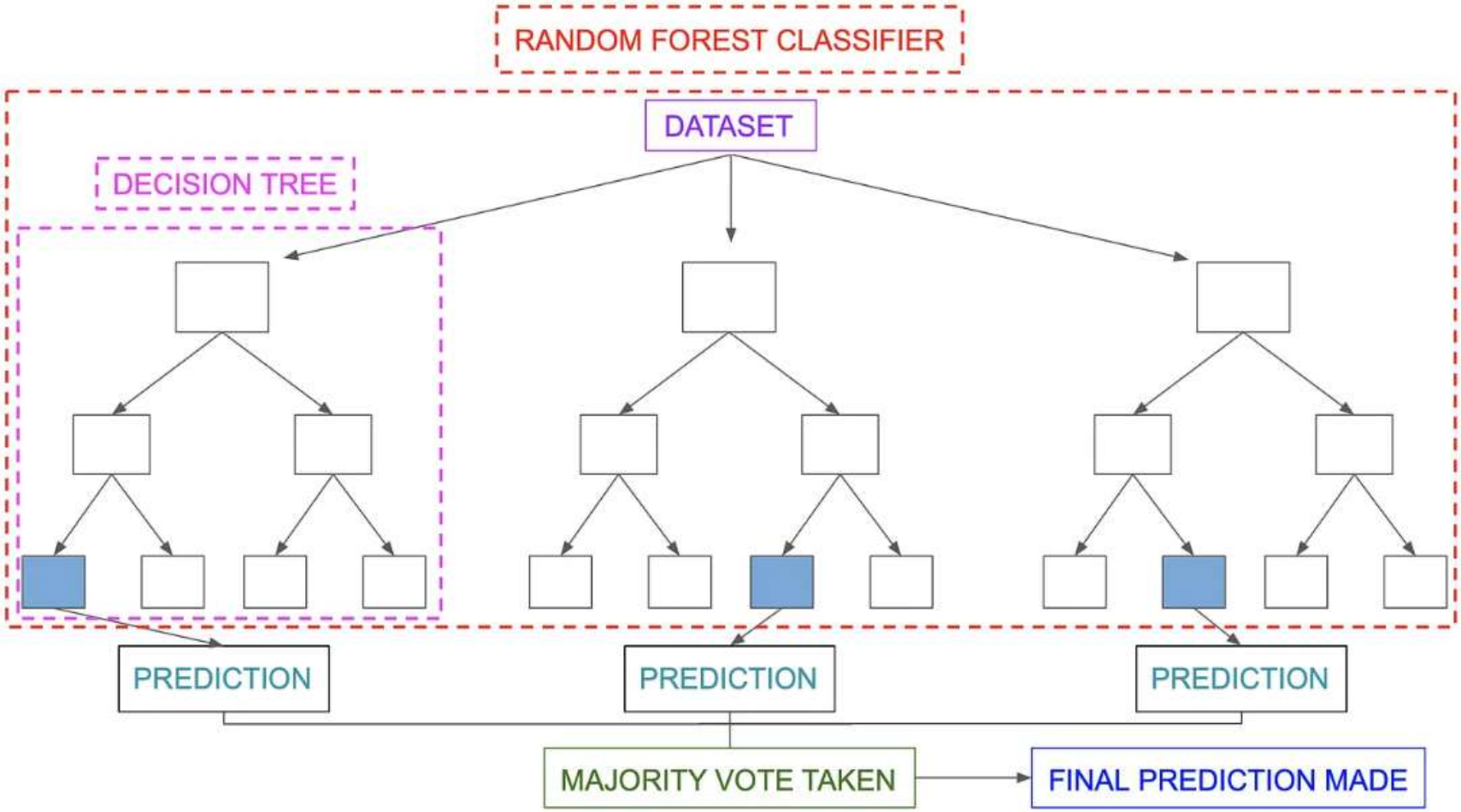
Random Forest Classifier

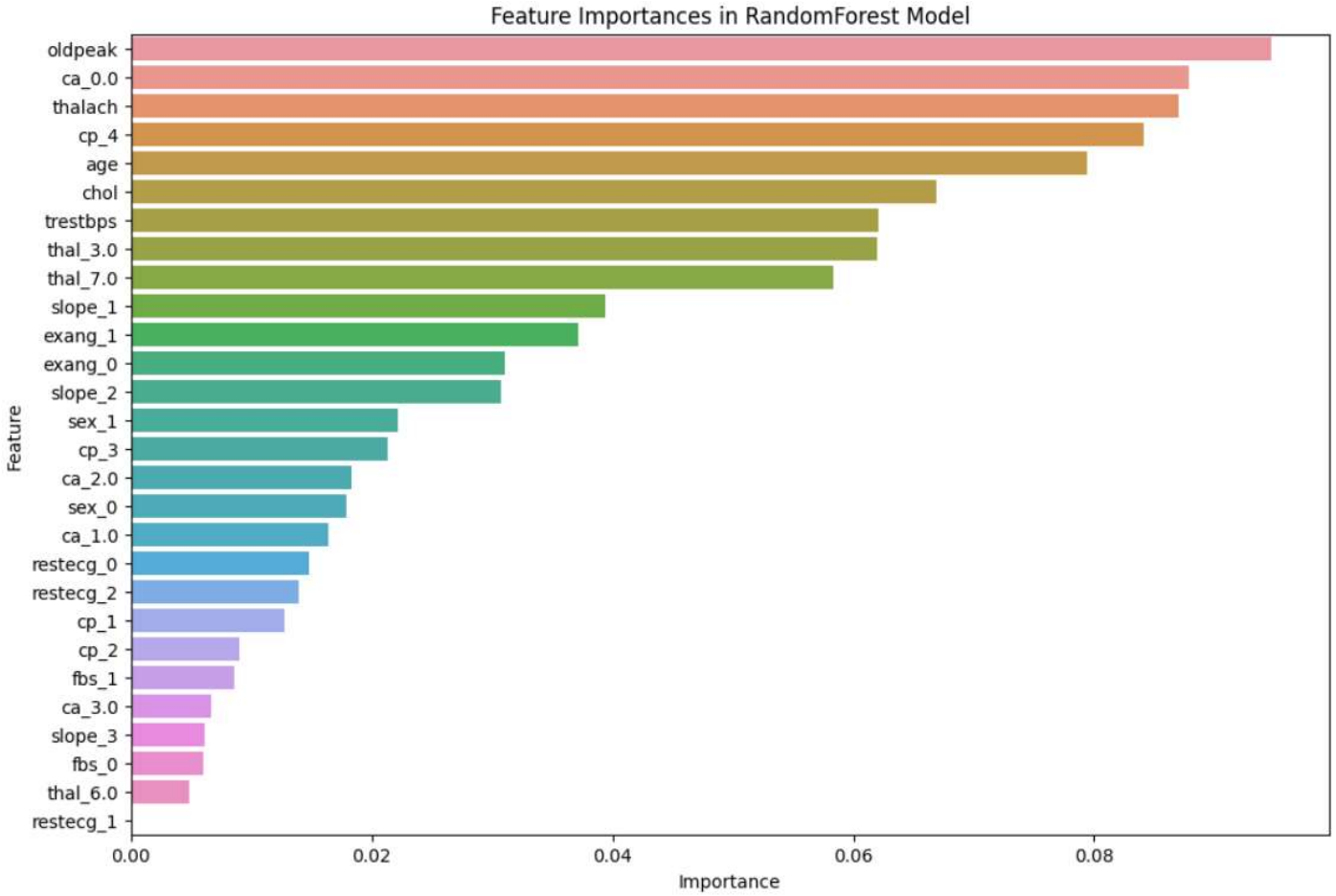
- The Random Forest model is a powerful ensemble learning technique that builds multiple decision trees.
- Each tree in the forest is trained on a random subset of the data and features, introducing diversity and preventing overfitting.
- The final prediction of the Random Forest is determined by majority voting among all the individual trees.
- In our study on heart disease prediction, Random Forest was chosen for its robustness and ability to handle both numerical and categorical data.

Features:

- It's more accurate than the decision tree algorithm.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of overfitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Cross validated using Gridsearch (CV=5)


RANDOM FOREST CLASSIFIER
DATASET
DECISION TREE
PREDICTION
PREDICTION
PREDICTION
MAJORITY VOTE TAKEN
FINAL PREDICTION MADE



Important Features:

1) ST Depression induced by exercise related to rest.
2) number of major vessels.
3) Maximum heart rate achieved.
4) Chest Pain
5) Age

Results:

Random Forest Accuracy: 0.8688524590163934

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	29
1	0.90	0.84	0.87	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

Support Vector Classifier

Allows classification errors: $\epsilon_i \geq 0$

$$\text{maximize: } M(\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n)$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

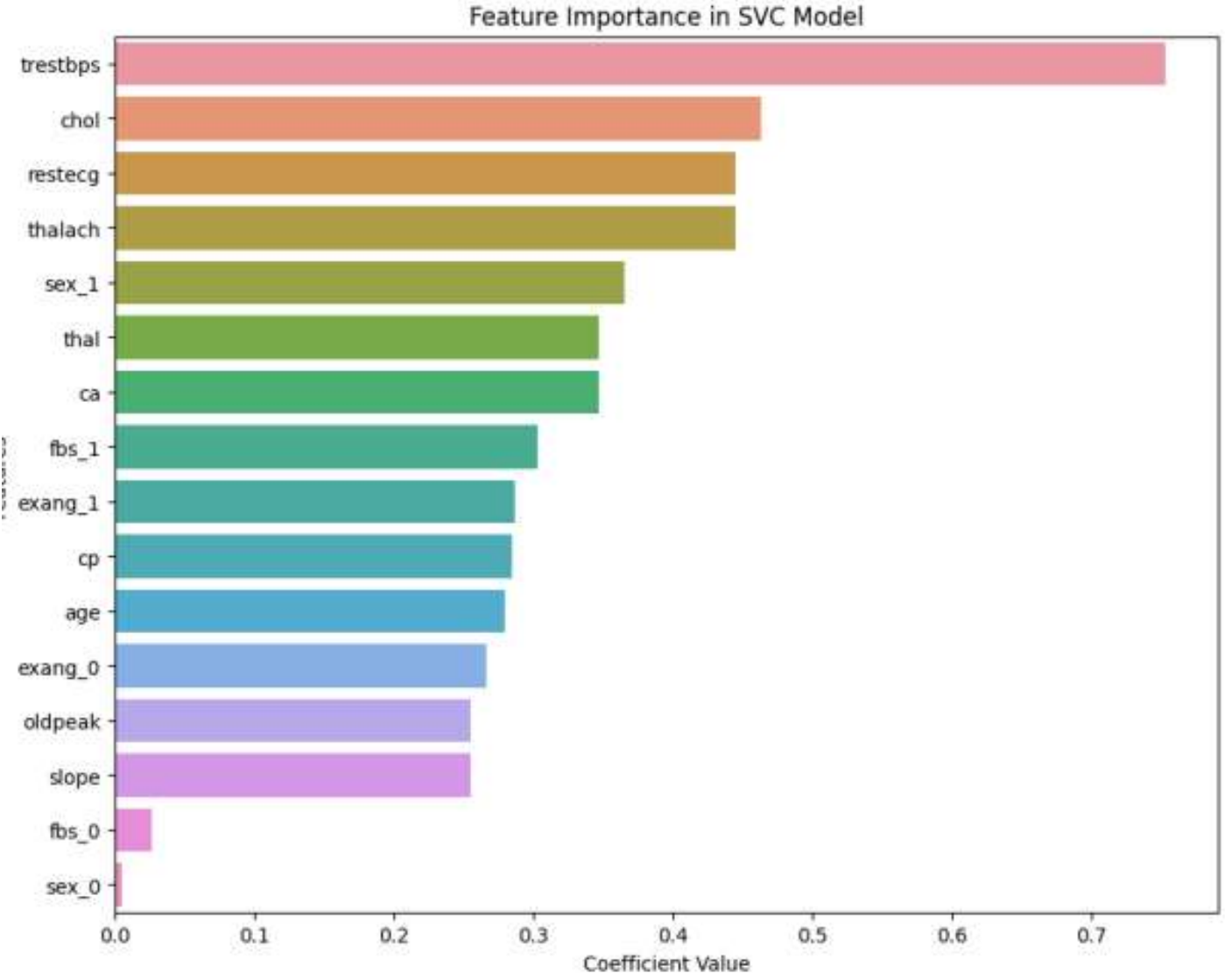
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \text{ and } \sum_{i=1}^{n} \epsilon_i \leq C$$

- $\epsilon_i = 0$: observation i is on the correct side of margin and hyperplane.
- $0 < \epsilon_i \leq 1$: observation i is on wrong side of margin, correct side of hyperplane.
- $\epsilon_i > 1$: observation i is on wrong side of the hyperplane.
- Observations with $\epsilon > 0$ are the *support vectors*.

```
Accuracy: 0.8688524590163934
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87        29
           1       0.90      0.84      0.87        32

    accuracy                           0.87        61
   macro avg       0.87      0.87      0.87        61
weighted avg       0.87      0.87      0.87        61
```



Important Features:

1) Resting Blood Pressure
2) Cholestoral
3) ECG at rest
4) Maximum Heart Rate Achieved
5) Gender

Conclusion:

- Based on the analysis of three models (Random Forest, Logistic Regression, and Support Vector Classifier), we aimed to identify the factors most strongly associated with heart disease and to accurately predict the presence of heart disease using the given features.

- The Random Forest model achieved an accuracy of 86.9%, with 'oldpeak', 'ca', and 'thalach' identified as the most important features.
- The Logistic Regression model showed a slightly higher accuracy of 88.5%, with 'sex', 'ca', and 'exang' being the most influential features.
- The Support Vector Classifier also had an accuracy of 86.9%, with 'trestbps', 'chol', and 'restecg' as the top features.
- While all three models performed similarly in terms of accuracy, the Logistic Regression model slightly outperformed the others, making it a robust choice for predicting heart disease in this dataset.

Oldpeak: A measure of ST depression induced by exercise relative to rest.
CA: The number of major vessels (0-3) colored by fluoroscopy.
Thalach: Maximum heart rate achieved.
Sex: Gender of the patient (with males having a higher association).
Chol: Serum cholesterol in mg/dl.
Restecg: Resting electrocardiographic results.
Exang: Exercise-induced angina.
Trestbps: Resting blood pressure (in mm Hg on admission to the hospital).

These factors consistently appeared as top predictors in our models. This analysis helped us answer our research questions by identifying key factors associated with heart disease and confirming that these models can effectively predict its presence.

Additionally, demographic factors like gender and clinical measurements such as serum cholesterol and resting ECG results were critical in determining heart disease risk. This comprehensive analysis shows the multifaceted nature of heart disease, which is influenced by a combination of physiological, demographic, and clinical factors.

Thank you for your attention!

Q&A

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University