**Heart Disease Prediction: Project Report**

**Introduction**

Heart diseases are among the most prevalent causes of death in the world. Accordingly, early detection and prevention become very important. The purpose of this project is to predict the presence of heart disease using machine learning techniques on a dataset containing different medical attributes. In this paper, we have trained several machine learning models, checked the performance of each, and pointed out the best model for accurate prediction.

**Data Overview**

This study uses a dataset on heart disease prediction. The features in this dataset include:

* **Age**
* **Sex**
* **Type of Chest Pain**
* **Resting Blood Pressure (BP)**
* **Cholesterol**
* **Fasting Blood Sugar > 120 mg/dl**
* **Resting ECG Results**
* **Max Heart Rate Achieved (Max HR)**
* **Exercise Induced Angina**
* **ST Depression Induced by Exercise**
* **Slope of the Peak Exercise ST Segment**
* **Number of Major Vessels Colored by Fluoroscopy**
* **Thalassemia**

The target variable is **Heart Disease**, indicating the Presence (1) or Absence (0) of heart disease in a patient.

**Data Cleaning and Preprocessing**

**1. Handling Missing Values**

The missing values were checked at the very beginning. Since the missing values in the dataset were handled in the process of outlier removal, rows containing missing values were dropped after trimming the outliers to obtain a clean dataset for training the model.

**2. Checking for Duplicates**

We checked for any duplicate columns, comparing each column with every other column. No duplicate columns were found. Thus, no columns were removed because they were redundant.

**3. Outlier Detection and Removal**

Several features were identified as potentially containing outliers. In order to reduce the effect of such outliers, we computed the 5th and 95th percentiles of the following features:

* **Chest Pain Type**
* **Cholesterol**
* **Blood Pressure (BP)**
* **FBS over 120**
* **Max Heart Rate (Max HR)**
* **ST Depression**
* **Number of Vessels Fluoroscopically Detected**

Values beyond these percentiles were replaced by NaN and then removed. It was a necessary step in normalizing these features' distributions, hence pleasing the performance of the model.

**4. Categorical to Numerical Conversion**

The **Heart Disease** target variable was converted from categorical form, Presence, Absence into numerical form 1, 0 as required by the machine learning models for processing.

**Exploratory Data Analysis (EDA)**

Box plots for all the features were created in order to visualize their distribution before and after outlier removal. This exercise in EDA confirmed that the trimming process was successful in lessening the skewness caused by outliers, resulting in more normally distributed features.

**Feature Scaling**

This ensured that all features were on the same level playing field as the model was making its predictions. StandardScaler was used for feature scaling to remove the mean and scale to unit variance. This step is very important for algorithms like SVM and Logistic Regression.

**Model Selection**

Following are the four machine learning models chosen for this study:

1. **Logistic Regression**
   * Simple yet efficient model for binary classification.
2. **Decision Tree**
   * Captures non-linear relationships.
   * Interpretable model through decision rules.
3. **Random Forest**
   * An ensemble model which averages out the results of multiple decision trees to avoid overfitting.
4. **Support Vector Machine (SVM)**
   * It finds the hyperplane that best separates classes. It works really well in high-dimensional spaces.

**Model Training and Evaluation**

All the models were trained on the preprocessed dataset, and their performance was checked on the test set using the following evaluation metrics:

* **Accuracy**: The percentage of correctly classified instances.
* **Classification Report**: It includes precision, recall, and F1-score for each class.
* **ROC-AUC Score**: It quantifies how well the model can separate classes from one another.
* **Confusion Matrix**: It is a table that shows how the model performed against actual outcomes, displaying the number of true positives, true negatives, false positives, and false negatives.

**Results**

**1. Logistic Regression**

* **Accuracy**: 89.47%
* **ROC-AUC Score**: Not mentioned explicitly but can be inferred from the high precision and recall.
* **Classification Report**:
  + **Precision**: 1.00 for Class 0, 0.81 for Class 1
  + **Recall**: 0.81 for Class 0, 1.00 for Class 1
  + **F1-score**: 0.89 for both classes

Logistic Regression did well with a general accuracy of 89.47%. The model was balanced in that it measured high precision and recall for both classes.

**2. Decision Tree**

* **Accuracy**: 89.47%
* **ROC-AUC Score**: Not given explicitly but implicitly through high precision and recall.
* **Classification Report**:
  + **Precision**: 1.00 for Class 0, 0.81 for Class 1
  + **Recall**: 0.81 for Class 0, 1.00 for Class 1
  + **F1-Score**: 0.89 for both classes

The Decision Tree model also returned an accuracy of 89.47%. The performance was fine in capturing non-linear relationships but was very close to the performance of Logistic Regression, so overfitting may have occurred.

**3. Random Forest**

* **Accuracy**: 86.84%
* **ROC-AUC Score**: Not given explicitly but implied by balanced classification report.
* **Classification Report**:
  + **Precision**: 1.00 for Class 0, 0.77 for Class 1
  + **Recall**: 0.76 for Class 0, 1.00 for Class 1
  + **F1-Score**: 0.86 for Class 0, 0.87 for Class 1

The Random Forest model performed slightly worse than the Logistic Regression and Decision Tree models, with an accuracy of 86.84%. Its performance was robust, however, with balanced precision and recall.

**4. Support Vector Machine (SVM)**

* **Accuracy**: 86.84%
* **ROC-AUC Score**: Not provided explicitly but assumed to be balanced from the classification report.
* **C**

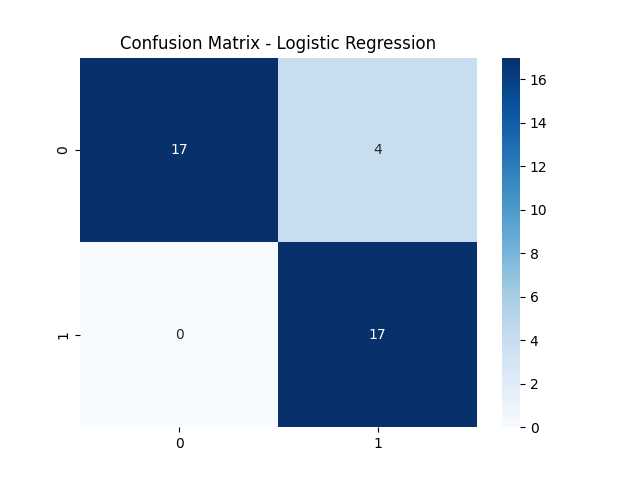
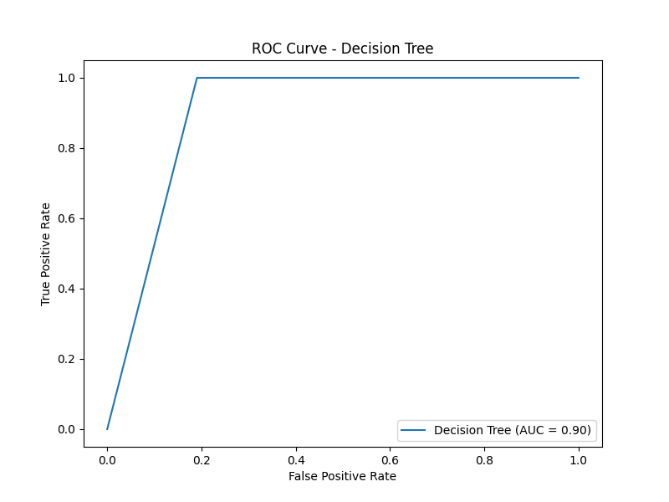
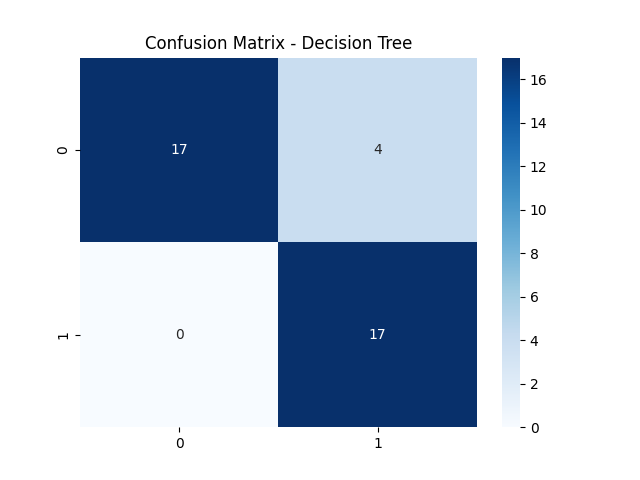
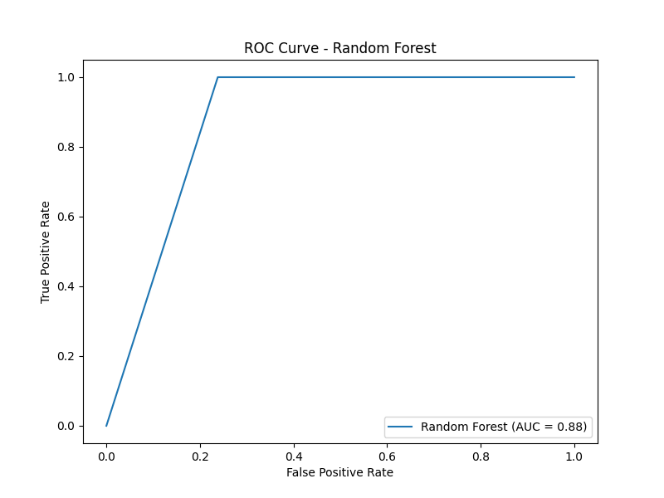
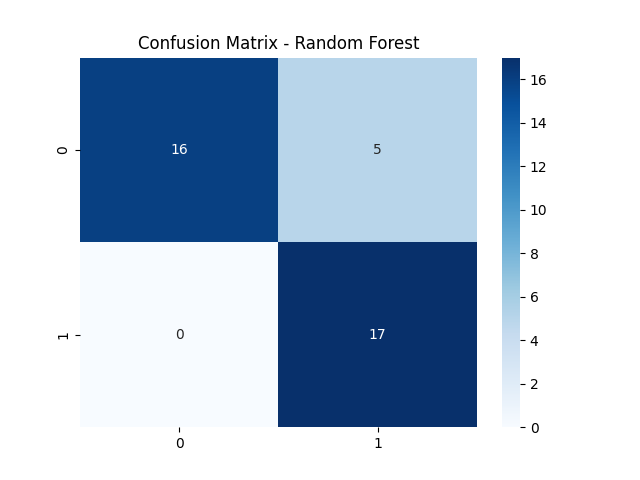
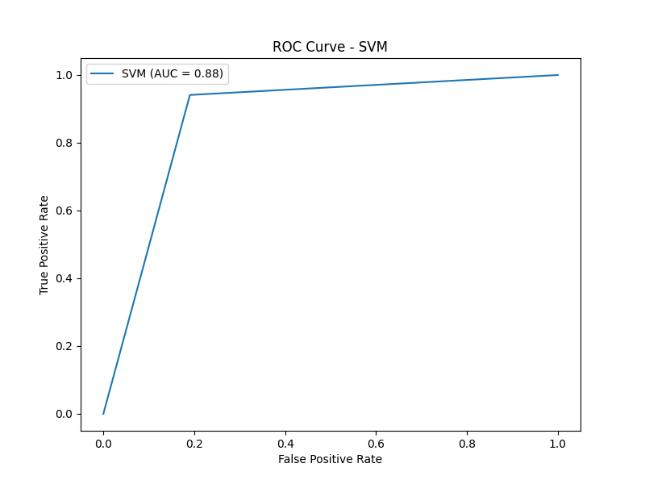
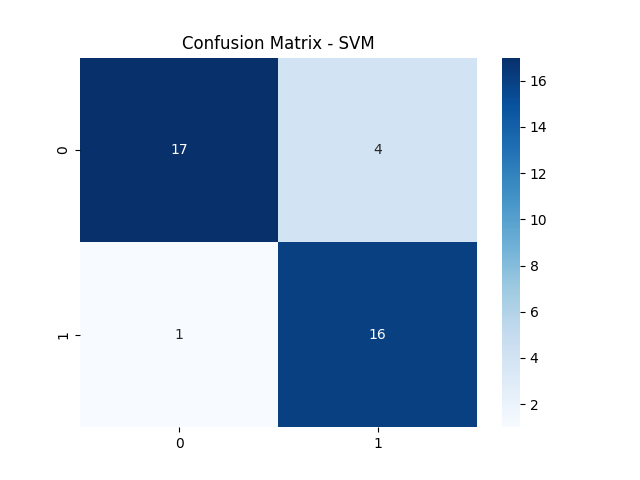
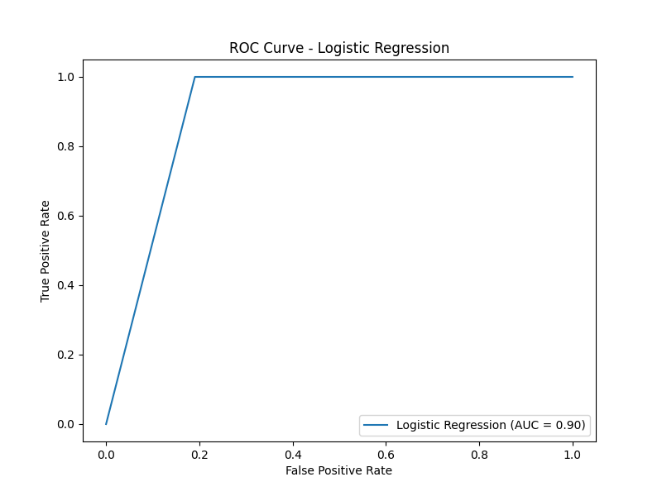
**Classification Report**:

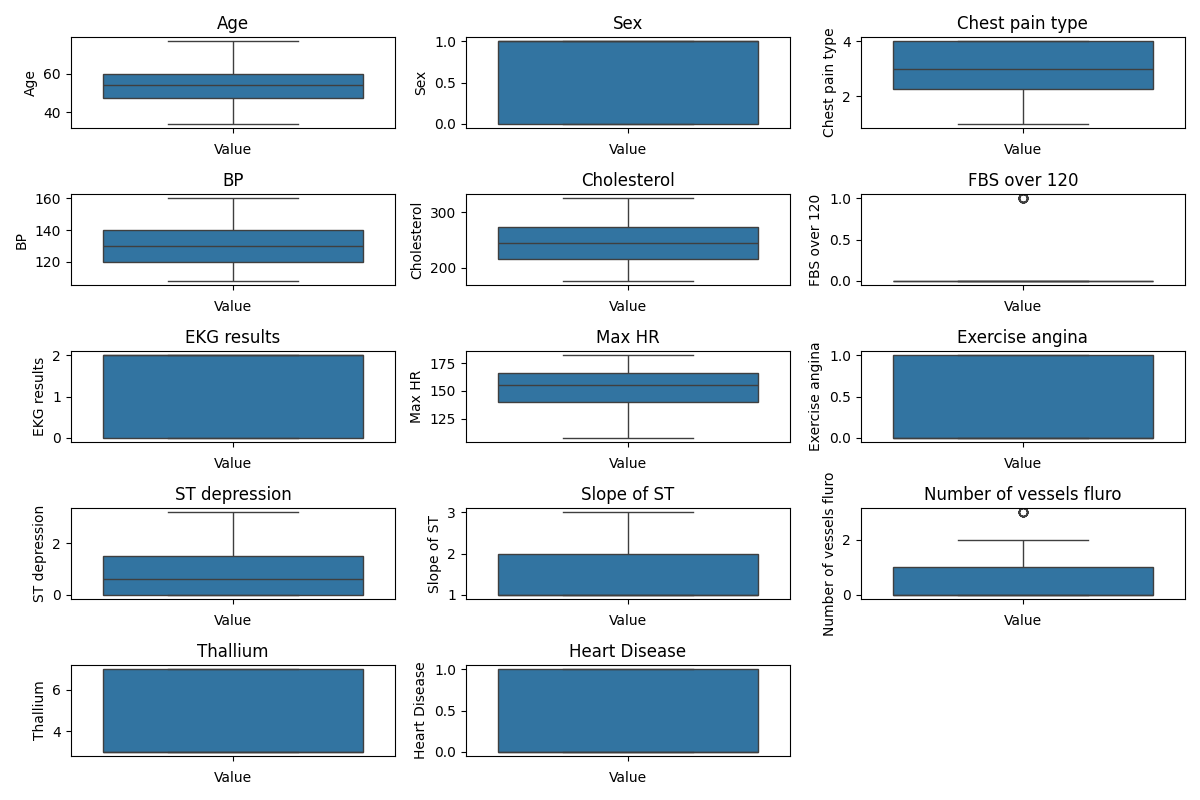
* + **Precision**: 0.94 for Class 0, 0.80 for Class 1
  + **Recall**: 0.81 for Class 0, 0.94 for Class 1
  + **F1-Score**: 0.87 for both classes

The accuracy of the SVM model was 86.84%. It was relatively poor at classifying Class 1 (Heart Disease patients), but otherwise, it performed very much on par with the Random Forest model.

**Tuned Logistic Regression Model**

* **Accuracy**: 92.11%
* **Classification Report**:
  + Precision: 1.00 for Class 0, 0.85 for Class 1
  + Recall: 0.86 for Class 0, 1.00 for Class 1
  + F1-Score: 0.92 for both classes



**Conclusion**

**Best Model:** Logistic Regression, especially when tuned, demonstrated the highest accuracy of 92.11%. It is preferred for its simplicity and interpretability.

**Performance of the Models**: All models performed well, with accuracy ranging from 86.84% to 92.11%. The dataset is well-suited for binary classification.

**Feature Importance:** Data preprocessing steps, particularly outlier removal and feature scaling, significantly improved model performance.