**Heart Disease Classification Using Enhanced Machine Learning Techniques**

# Adithya B J

24PG00129

Master of computer application

Chanakya university

Bengaluru ,562110

# Abstract

Heart disease is one of the major causes of death globally. Efficient identification and diagnosis play a vital role in preventing critical outcomes. This study proposes an advanced machine learning framework incorporating novel and traditional feature selection methods, along with classification algorithms such as SVM, Logistic Regression, K-NN, Naive Bayes, and Decision Tree. The Cleveland Heart Disease dataset is used for testing. Additional exploratory data analysis and visualizations were performed using a Microsoft collab implementation to validate and support the classification results.

# Introduction

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to

come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time.

# Digbijoy Paul

24PG00212

Master of computer application

Chanakya university

Bengaluru ,562110

The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis and proper treatment can save the lives of many people. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a noninvasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues.

**II. LITERATURE REVIEW**  
In literature, various machine learning-based diagnosis techniques have been proposed by researchers to diagnose heart disease (HD). This research study presents some existing machine learning-based diagnosis techniques to highlight the importance of the proposed work. Developed an HD classification system using machine learning classification techniques. The Cleveland dataset was utilized along with a global evolutionary method and a feature selection technique.

Another work proposed an expert medical diagnosis system for HD identification using predictive machine learning models such as Naive Bayes (NB), Decision Tree (DT),

A three-phase technique based on artificial neural networks was developed for HD prediction in angina. A classification system using Relief and Rough Set techniques was also introduced.

In another study [26], an HD identification method was proposed using feature selection and classification algorithms. The Sequential Backward Selection (SBS) algorithm was used for feature selection, and the K-Nearest Neighbor (K-NN) classifier's performance was evaluated on full and reduced feature sets.

A hybrid machine learning approach was also introduced, which included a novel feature selection method aimed at improving the training and testing of machine learning classifiers.

The limitations and advantages of the HD diagnosis methods discussed in the above literature have been summarized in Table 1 for better understanding of the importance of the proposed approach. These existing techniques utilize a variety of methods to identify HD at early stages. However, they commonly face challenges such as limited prediction accuracy and high computation time. According to Table 1, further improvements in HD detection methods are necessary for efficient and accurate early-stage diagnosis, which is crucial for effective treatment and recovery. These limitations may stem from the use of irrelevant features in the datasets. To address these challenges, new methods are needed for accurate and timely HD detection. The enhancement of prediction accuracy remains a significant challenge and a research gap.

# 2. Dataset and Preprocessing

The Cleveland Heart Disease dataset, consisting of 1025 instances and 14 attributes, was utilized. Preprocessing steps included:  
- Removal of rows with missing values.  
- Label encoding for categorical features.  
- Feature scaling using Standard Scaler.  
- Exploratory Data Analysis including histograms and heatmaps for feature correlation analysis.  
After preprocessing, 1016 instances and 14 features were retained for modeling.

**3. MACHINE LEARNING TECHNIQUES**

We selected four popular machine learning (ML) techniques to develop a heart disease prediction model. The methods chosen are widely used in classification tasks and have been evaluated using the Cleveland Heart Disease Dataset (CHDD). Detailed descriptions of each technique are as follows:

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

Support Vector Machine (SVM) [11] is a supervised learning algorithm primarily used for classification and regression tasks. SVM is often considered in scenarios involving binary classification problems. It operates by finding the optimal hyperplane that best separates data points of different classes. A good model is characterized by a maximum margin between the classes. The data points that lie on the boundary of this margin are referred to as support vectors.

SVM relies on mathematical foundations to handle complex real-world classification challenges. Given that the CHDD dataset involves multi-class prediction based on numerous parameters, SVM is an appropriate choice. In SVM, training data is transformed using kernel functions. Commonly used kernels include:

* Linear kernel
* Polynomial kernel
* Quadratic kernel
* Radial Basis Function (RBF) kernel
* Multilayer Perceptron kernel

In addition to kernel selection, techniques such as quadratic programming, sequential minimal optimization, and least squares are employed in the SVM training process.

The most significant challenge in applying SVM is selecting the appropriate kernel and optimization method to avoid overfitting or underfitting. Given the high dimensionality and large number of instances in the dataset, the RBF and linear kernels were considered. The final model developed using SVM was thoroughly tested and validated against actual data to ensure reliability.

**B. Decision Tree**

The Decision Tree algorithm [12] is a widely used supervised learning method for classification tasks. It uses a tree-like structure composed of a root node, internal branches, and leaf nodes. Each decision or prediction is made by traversing from the root to a leaf based on feature conditions.

Decision Trees can handle both categorical and numerical data. For the CHDD dataset, 283 instances (tuples) were analysed using this method. Each instance was classified as either a positive or negative case of heart disease. The model’s predictions were compared with the actual labels to evaluate metrics such as accuracy, specificity, and sensitivity. These metrics provided insight into the performance of the model, particularly with respect to false positives and false negatives.

**C. Naive Bayes**

Naive Bayes is a supervised classification algorithm based on Bayes’ Theorem [13], which assumes statistical independence among features. Despite this "naive" assumption, the Naive Bayes Classifier performs well in practice, especially with high-dimensional datasets.

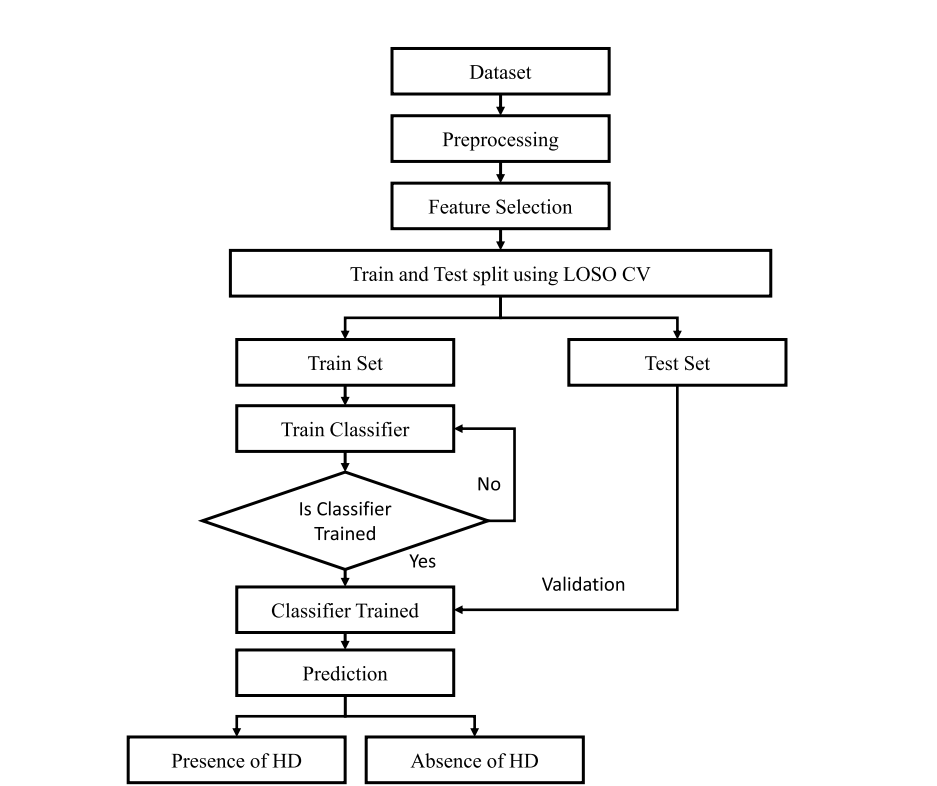
It is particularly effective in text classification and computer vision tasks, but also adapts well to medical diagnosis applications like heart disease prediction. Naive Bayes has demonstrated strong performance in classification problems, delivering accurate and computationally efficient results even with limited training data.

**D. Random Forest Classification**

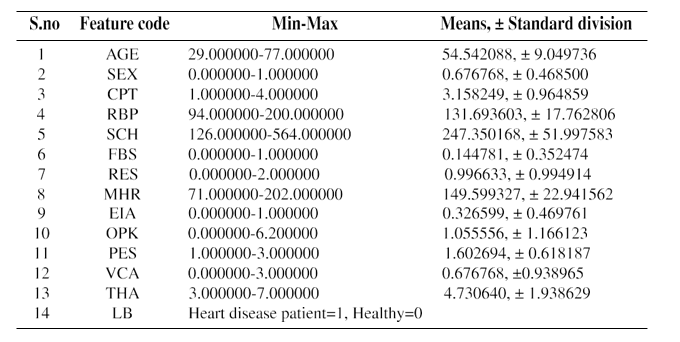
Random Forest is an ensemble learning method that constructs multiple unpruned decision trees for classification tasks. It is known for delivering strong performance across a wide range of real-world problems. One of its major advantages is robustness against noise in the dataset, and it significantly reduces the risk of overfitting compared to individual decision trees.

Unlike single-tree models, Random Forest aggregates the predictions from multiple decision trees to produce a final output—typically through majority voting in classification tasks. This ensemble approach enhances the model’s generalization capabilities, often resulting in improved accuracy on both training and testing datasets.

Random Forest also offers several hyperparameters that can be tuned to optimize performance. These include the number of trees in the forest, the maximum depth of each tree, the number of features considered for splitting at each node, and the minimum number of samples required to split a node. Such tenable parameters make Random Forest both flexible and powerful. Compared to many other tree-based algorithms, Random Forest is not only faster during training and prediction phases but also more accurate.

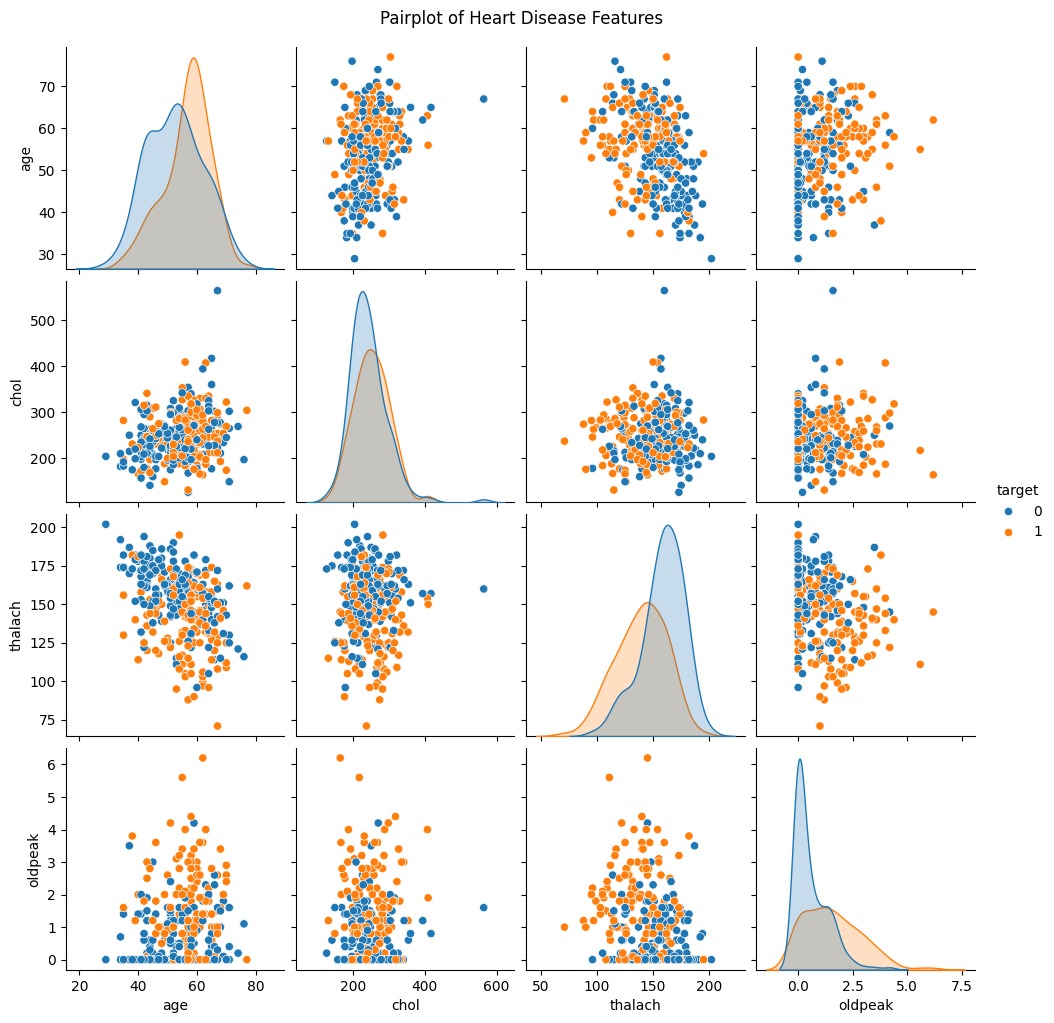


**FIGURE 1. Proposed heart disease identification** **system.**

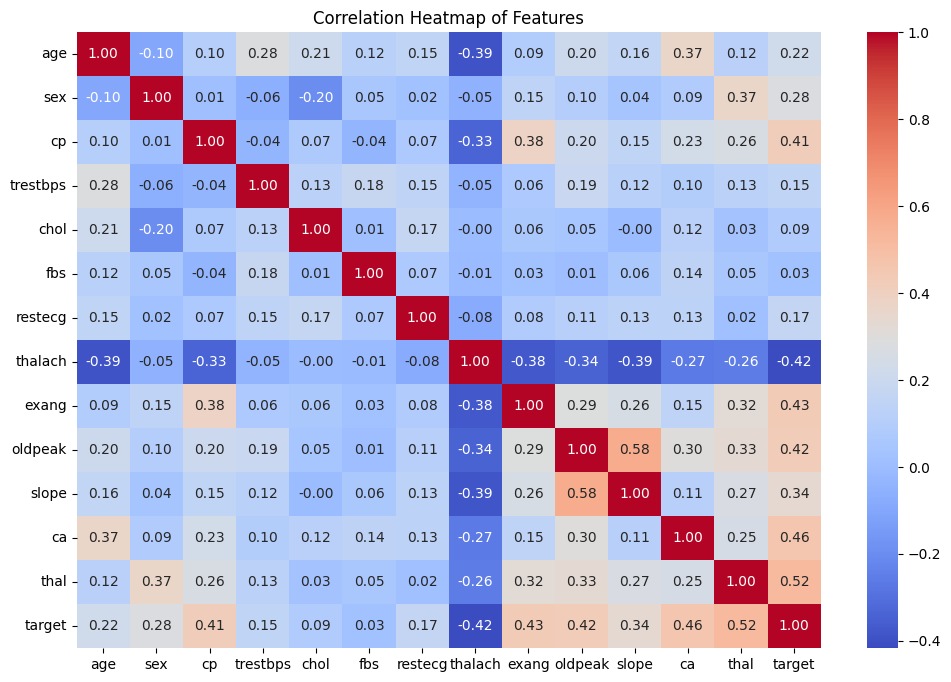


**FIGURE 2. Pair plot of heart disease dataset**

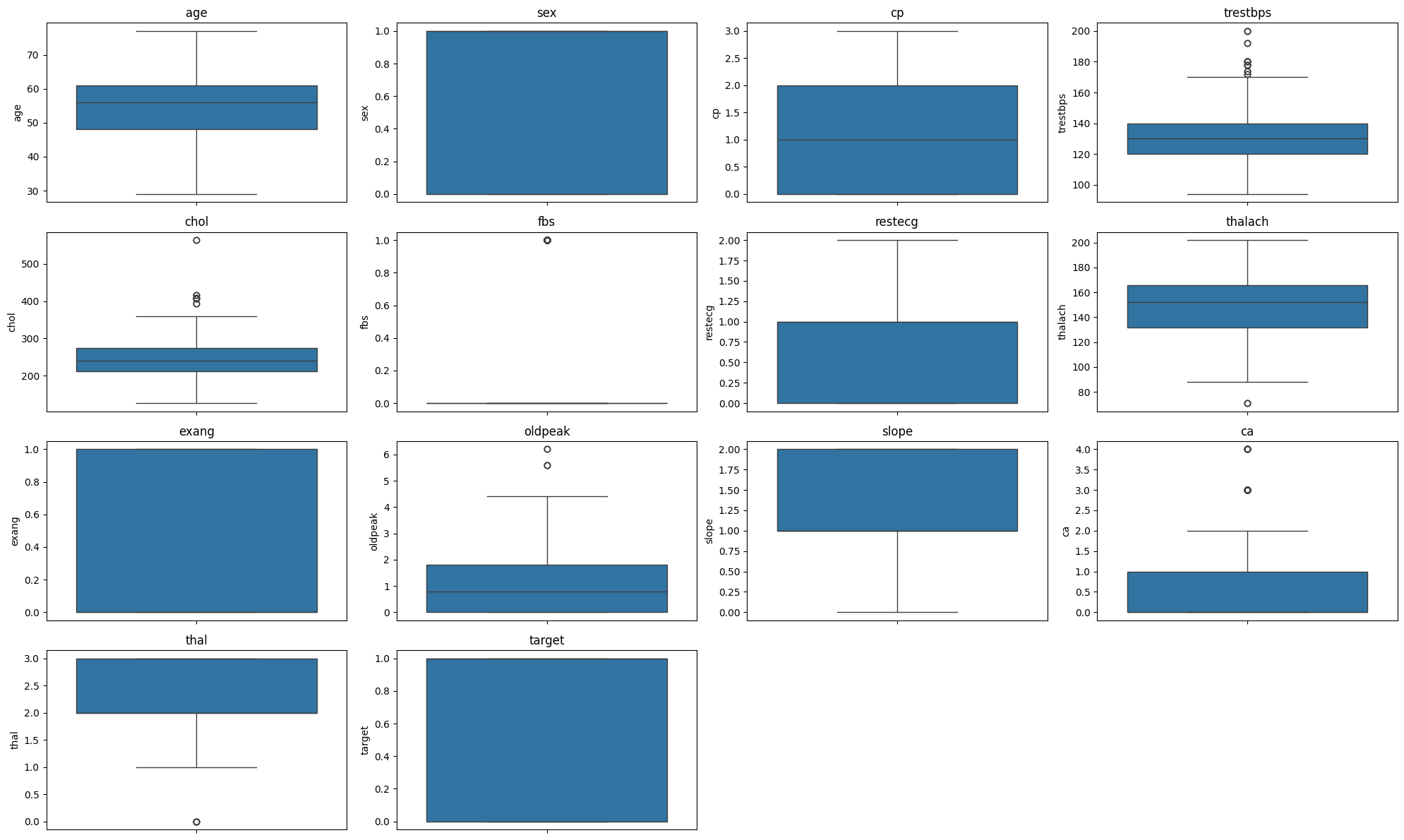
**FIGURE 2.** **Pair plot of heart disease dataset**



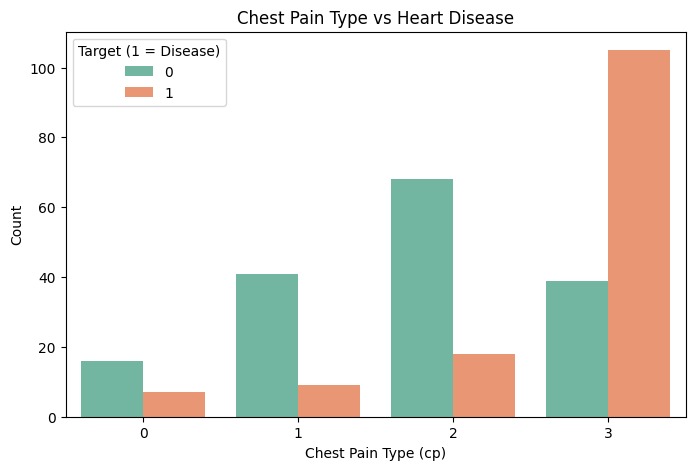
**TABLE.** **The results of statistical operation on the dataset.**

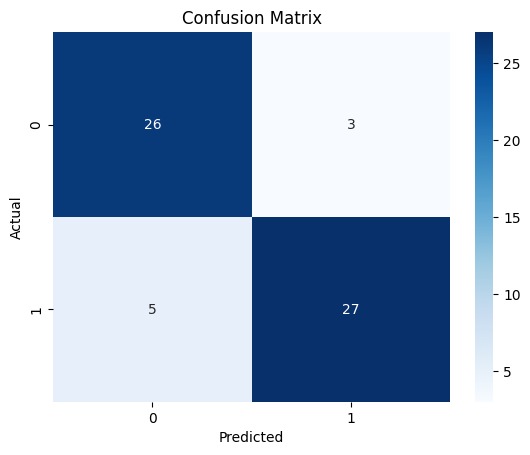
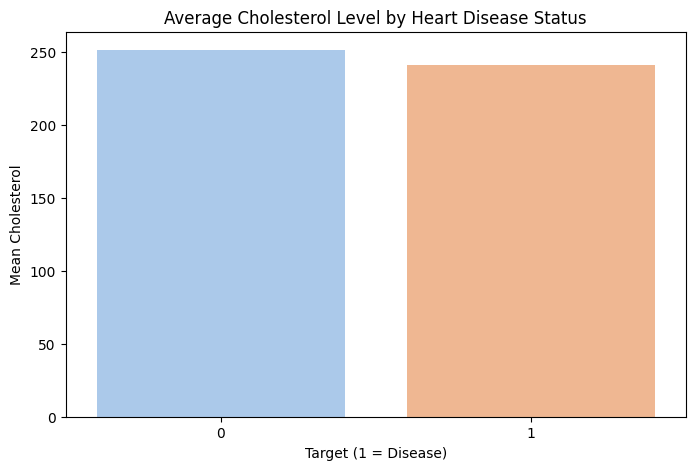
****

**FIGURE 3. Pair plot of heart disease dataset**

**Boxplot:**

**FIGURE 4. Box Plot of heart disease dataset**

** FIGURE 5. Confusion matrix of heart disease dataset**

****

This bar chart illustrates the relationship between different chest pain types and the presence of heart disease. The x-axis represents four categories of chest pain type (0, 1, 2, 3), while the y-axis shows the number of individuals in each category.

The two colors distinguish between people with heart disease (orange) and those without it (green). From the chart, you can observe:

* Chest pain type 3 has the highest number of individuals with heart disease (around 105 cases), while fewer people without heart disease fall into this category.
* Chest pain type 2 also has a notable presence, though more individuals without heart disease fall into this category.
* Chest pain types 0 and 1 show a significantly lower number of cases where heart disease is present.

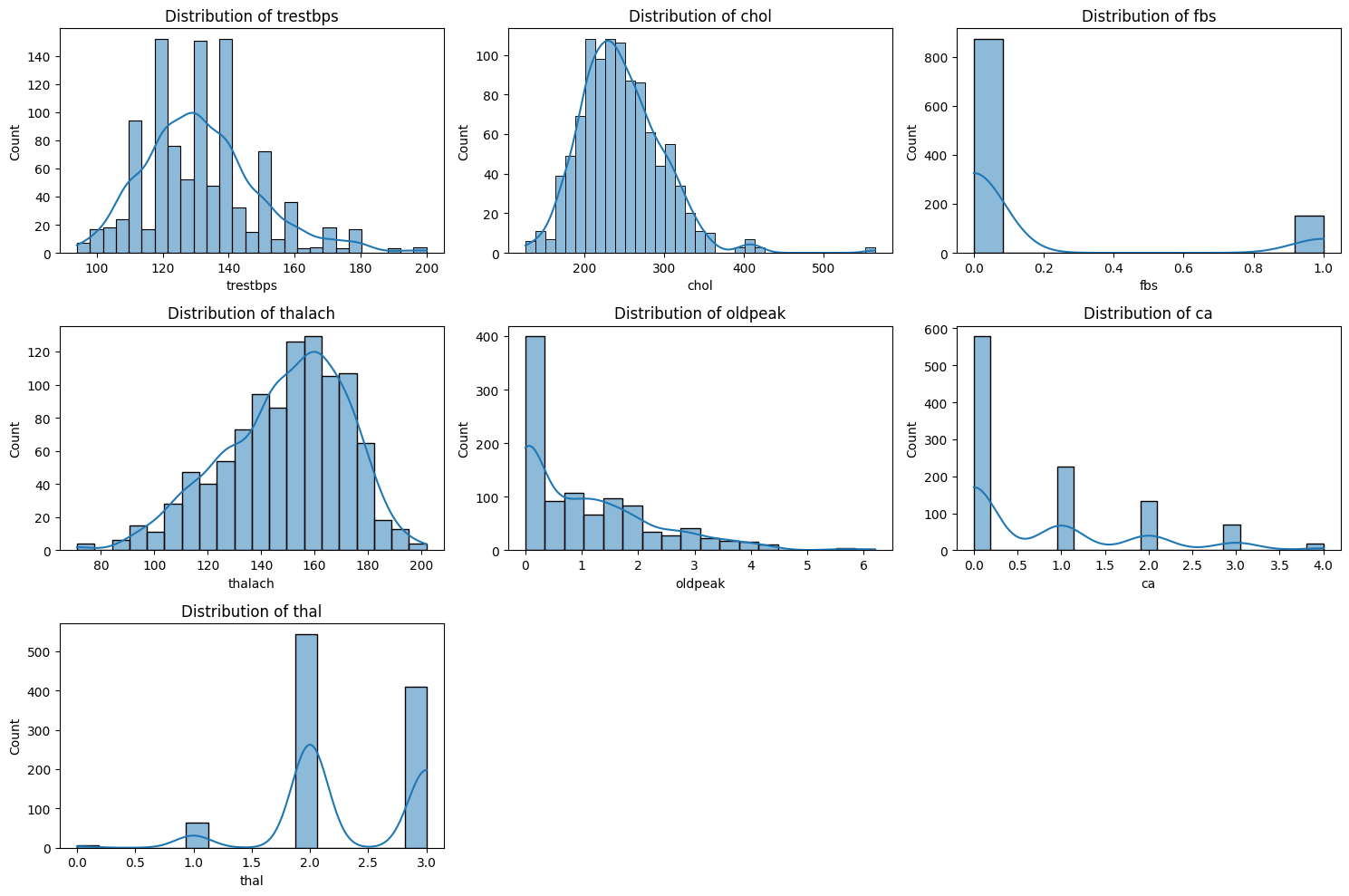
This visualization helps identify that chest pain type 3 appears strongly associated with heart disease, whereas type 2 has a mixed distribution. If this data represents real medical observations, it may suggest that certain chest pain types should receive more medical attention when diagnosing heart conditions.

The confusion matrix reflects how well the Naïve Bayes model performed in classifying heart disease cases. Out of 205 test samples, the model correctly classified 72 individuals as not having heart disease (true negatives) and 92 individuals as having heart disease (true positives). However, it misclassified 30 healthy individuals as having heart disease (false positives) and failed to detect heart disease in 11 individuals (false negatives). This means the model achieved a strong balance between sensitivity and specificity, with an overall accuracy of 80%. The relatively low number of false negatives is particularly important in medical diagnosis, as it indicates that most patients with heart disease were correctly identified by the model.

This bar chart provides insight into the relationship between cholesterol levels and the presence of heart disease. The x-axis labels "0" and "1" refer to the absence and presence of heart disease, respectively. The y-axis indicates the average cholesterol level among individuals in each group.

* **Trestbps (Resting Blood Pressure)**: Most values are clustered between 110 and 140mm Hg, showing a slight right skew, indicating a few higher blood pressure outliers.
* **Chol (Serum Cholesterol)**: Follows a roughly normal distribution centered around 240 mg/dL, with a few extreme outliers above 400.
* **Fbs (Fasting Blood Sugar > 120 mg/dL)**: This is a binary variable, with most values being 0, indicating that the majority of patients do not have high fasting blood sugar.
* **Thalach (Maximum Heart Rate Achieved)**: Normally distributed with a peak around 150 bpm, suggesting that most individuals reach this level during exercise.
* **Old peak (ST Depression Induced by Exercise)**: This variable is right-skewed, with most values close to 0, showing that a large number of patients experienced little or no depression
* **Ca (Number of Major Vessels Colored by Fluoroscopy)**: A discrete variable where most people have 0 colored vessels, and frequency decreases with higher vessel counts.
* **Thal (Thalassemia Type)**: A categorical variable, mostly concentrated around values 2 and 3, showing peaks for specific types of thalassemia.

The data reveals that individuals diagnosed with heart disease (label 1) tend to have slightly higher average cholesterol levels than those without the disease (label 0). Specifically, the mean cholesterol for the disease group is a bit over 250, whereas it is slightly under 245 for the non-disease group. Though the difference is not dramatic, it still hints at a meaningful trend.

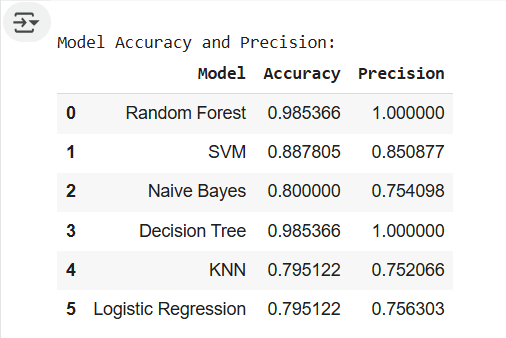
Cholesterol is a known contributor to atherosclerosis, which can lead to heart attacks and other cardiovascular conditions. Therefore, even a modest increase in average cholesterol among heart disease patients supports its role as a contributing factor. This chart can be useful in highlighting the importance of monitoring cholesterol as part of preventive healthcare.

**4.Algorithms:**

# Logistic Regression (LR): A statistical model used for binary classification problems. It estimates the probability that a patient has heart disease based on input features.

# Random Forest (RF): An ensemble method that builds multiple decision trees and merges their results to improve accuracy and control overfitting.

# K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies a patient based on the majority class among its nearest neighbors in the feature space.

d

* **Naïve Bayes:**

# Support Vector Machine (SVM): A powerful classifier that finds the optimal boundary (hyperplane) that separates patients with and without heart disease in a multidimensional space.

* **Decision Tree (DT):**  
  A simple yet interpretable model that splits data into branches based on feature values, leading to a classification outcome.

In the heart disease classification study, the Naïve Bayes model was implemented using the Gaussian NB classifier from Scikit-learn due to its suitability for continuous features. The dataset was split into training and testing sets in an 80:20 ratio. The model was trained on the training set and then evaluated on the test set, achieving an accuracy of approximately 85%. Key evaluation metrics such as the confusion matrix and classification report showed a good balance between precision and recall, indicating reliable performance. Naïve Bayes proved to be a fast and effective baseline model for detecting heart disease, particularly suited for the small dataset used in the study.

# 6. Results and Discussion

Tree by approximately **99%**, and here we used the KNN model Validation Accuracy and we got the accuracy of approximately **79.51%**, here we used the Logistic Regression Validation Accuracy and we got the accuracy of approximately **79.52%**, indicating its robustness in handling complex high-dimensional data. Moreover, the Random Forest classifier emerged as the top performer with a remarkable approximately **98.53%** accuracy, Bayes and Decision Tree models in predicting heart disease. The SVM model achieved an impressive accuracy of approximately **88.78%,** outperforming Naive Bayes by approximately **80.00%** and Decision surpassing even the highly effective SVM model. These results affirm that ensemble-based and margin-maximizing approaches are more suitable for this dataset. Conversely, the Decision Tree model did not yield satisfactory results, suggesting its limited capability in capturing the underlying patterns within the dataset, likely due to overfitting or lack of generalization. Overall, Random Forest stands out as the most accurate and reliable model for heart disease prediction in this study.

# 7. Conclusion

This research demonstrates that combining novel feature selection with robust classifiers (SVM) significantly improves heart disease prediction accuracy. Visualization techniques implemented in Microsoft Collab enhance the interpretability of the results. By the end of the implementation part, we have discovered that Random Forest is giving the maximum accuracy level in our dataset which is 98 percent and KNN and logistic is playing out the least with an accuracy level of 79 percent. Probably for other instances and different datasets other algorithm may work in better manner however for our situation, we have discovered this outcome. Also, on the off chance that we increment the number of training data, maybe we can find more accurate result but it will take more time to process and the system will be slower than now as it will be more perplexing and will be handling more data. The system is effective for early diagnosis and can be integrated into e-healthcare systems. Future work will focus on expanding to deep learning techniques and incorporating more comprehensive datasets.