A Comparative Analysis of Classification Models

Name: Pratisha Bista

Group: L5CG1

University ID: 2408284

5CS037: Concepts and Technologies of AI

Module Leader and Tutor: Mr. Siman Giri

February 11, 2025

Abstract

The purpose of this report is to compare the performance of different classification models and analyze the impacts of hyperparameter optimization and feature selection in any model's performance. The primary goal is to detect the presence of heart disease based on various risk factors using a dataset from Kaggle, originally sourced from UCI Machine Learning Repository. The dataset contains patient attributes from multiple medical institutions. The research focuses on comparing three classification models including a Logistic Regression model using Sigmoid function in addition to Decision Tree Classifier and Random Forest Classifier. An evaluation of the impact that occurs when hyperparameters are tuned and features are selected upon model performance is performed.

Building a perfect predictive model remains secondary compared to understanding the dataset characteristics and model behavior while identifying opportunities for improvements. This research provides valuable insights for heart disease risk assessment while supporting UN Sustainable Development Goal 3: Good Health and Well-being through data-based healthcare practices.

Contents

Introduction

The proper condition of the heart represents the fundamental component required for maintaining a healthy existence. The heart maintains vital life by performing blood circulation to transport oxygen-rich fluid to all body regions. Many lifestyle aspects and genetic characteristics raise the chance of developing heart disease since it represents a major health condition worldwide.

The main purpose of this research explores how different medical traits can help determine heart disease presence. This approach prioritizes model performance interpretation along with identification of its constraints and determinants of success along with accuracy goals. Through a thorough analysis of the dataset combined with model evaluation I aim to identify which variables play the most significant role in predicting heart disease and how different models understand their relationship patterns.

The study follows an organized procedure beginning with exploratory data analysis (EDA) and data cleaning and preprocessing before developing quality input data. The study then explores model development, comparing a Logistic Regression model with Decision Tree and Random Forest classifiers. The study evaluates the influence of hyperparameter tuning and feature selection on model performance.

Exploratory Data Analysis (EDA)

The dataset used for this study contains 1025 rows and 14 columns, with the following attributes:

1. Age
2. Sex
3. Cp (Chest Pain type)
4. Trestbps (Resting Blood Pressure)
5. Chol (Serum Cholesterol)
6. Fbs (Fasting Blood Sugar > 120 mg/dl)
7. Restecg (Resting Electrocardiographic results)
8. Thalach (Maximum Heart Rate Achieved)
9. Exang (Exercise Induced Angina)
10. Oldpeak (Depression Induced by Exercise Relative to Rest)
11. Slope (Slope of the peak exercise ST segment)
12. Ca (Number of Major Vessels Colored by Fluoroscopy)
13. Thal (Thalassemia)
14. Target (Presence of Heart Disease)

The dataset consists of numerical values, with no string-based categorical features present. It is possible that some features had already been encoded, as evident in the dataset's structure. The data types primarily consist of integers for most columns, except for Oldpeak, which is a float type. Additionally, no missing values were found in the dataset.

During the EDA, it was noted that 723 identical rows were present within the dataset, which would need to be removed during the data cleaning process to improve model performance. Understanding the class imbalance of the target variable before proceeding further is important to mitigate its negative impact on model performance.

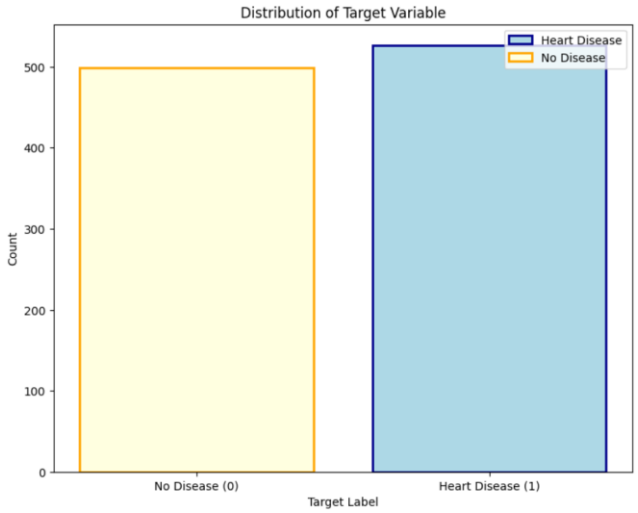


Figure 1: Distribution of Target Variable

Based on the above figure, more people have heart disease, but the difference is quite small, only about 2.64%. Therefore, no balance is needed for this case.

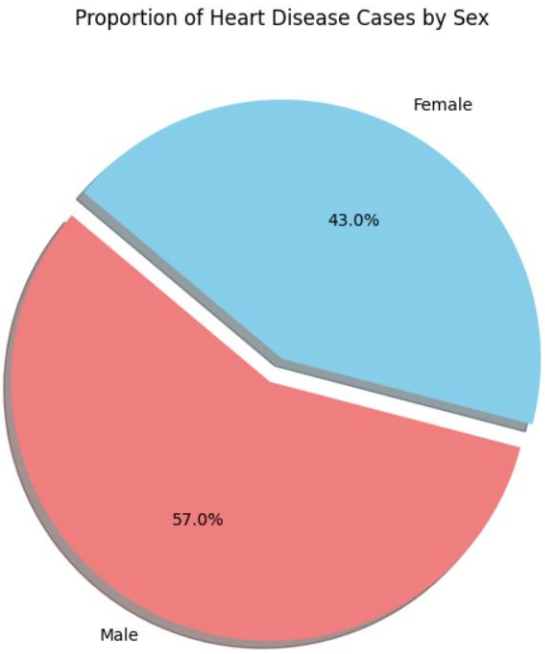


Figure 2: Proportion of heart disease by gender

Males have a 14% higher likelihood of developing heart disease compared to females. This disparity may be influenced by various factors, including genetic predisposition, hormonal differences, and lifestyle choices such as diet, exercise, and stress management.

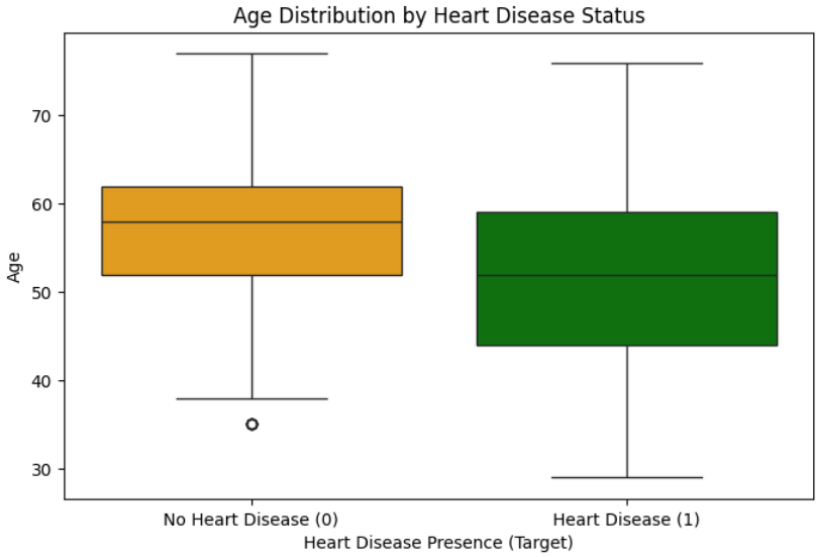


Figure 3: Age Distribution by Heart Disease Status

The average age of people who do not have heart disease exceeds the average age of people who do have the condition. People without heart disease demonstrate a wider spread of 50% data points in their age distribution compared to those with heart disease as shown by interquartile range. Every age group from 30s to 70s is equally present in both groups showing extensive age variation. The "No Heart Disease" group contains one exceptional case which shows an unusually young age. Heart disease exists within a wide range of ages but people with heart disease typically age younger than those without the condition because age serves as an insufficient determinant of heart disease risk.

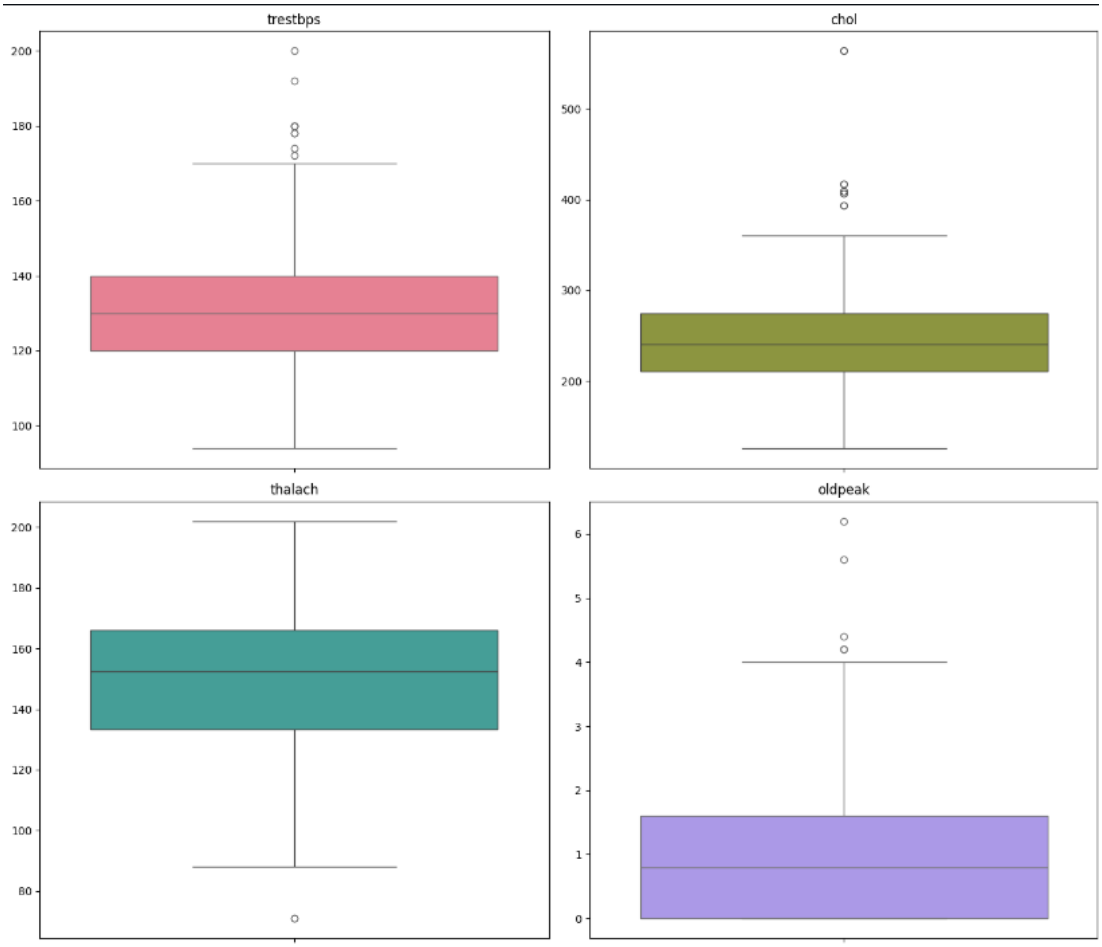


Figure 4: Box Plot of Continuous variables to detect outliers

The four boxplots reveal significant information about Resting Blood Pressure (trestbps) and Cholesterol (chol) and Maximum Heart Rate (thalach) and ST Depression (oldpeak). The center of the Resting Blood Pressure (trestbps) distribution sits at 125-130 mmHg and stretches from 115 to 140 mmHg. A couple of patients have elevated blood pressure readings over 160 mmHg as shown by outlier data points which indicates possible hypertension conditions that pose cardiovascular risks. Most subjects in the Cholesterol (chol) plot received measurements between 200 to 300 mg/dL yet there are multiple cases with marks exceeding 400 mg/dL up to an extreme reading above 500 mg/dL which indicates pronounced increased cardiovascular risk. The Maximum Heart Rate plot (thalach) shows a median value at 150 bpm and spanned most measurements from 130 to 170 bpm however various data points under 90 bpm may indicate heart conditions or poor cardiovascular health in participants. The ST Depression (oldpeak) chart has a median value of 0.5 and displays

most readings within the 0 to 1.5 range while two outliers rise above 4 and one value soars beyond 6 which might indicate exercise-related ischemic symptoms related to heart conditions. Several cases of extreme values in blood pressure measurements combined with cholesterol levels and ST depression status demonstrate why certain individuals face a higher cardiovascular risk thus requiring additional medical screening.

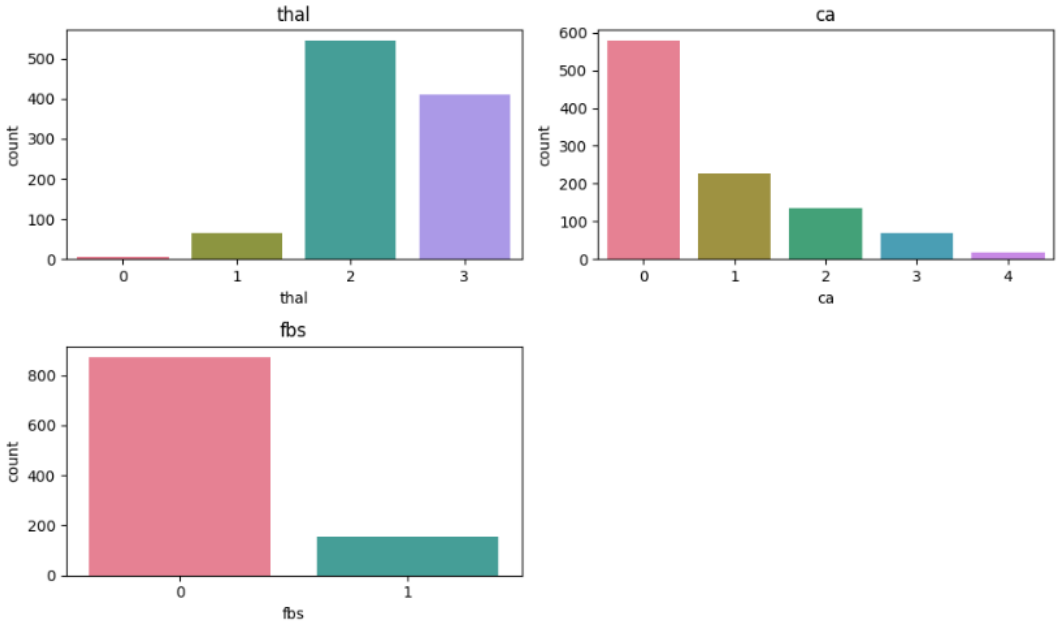


Figure 5: Class Imbalance of thal, ca, fbs

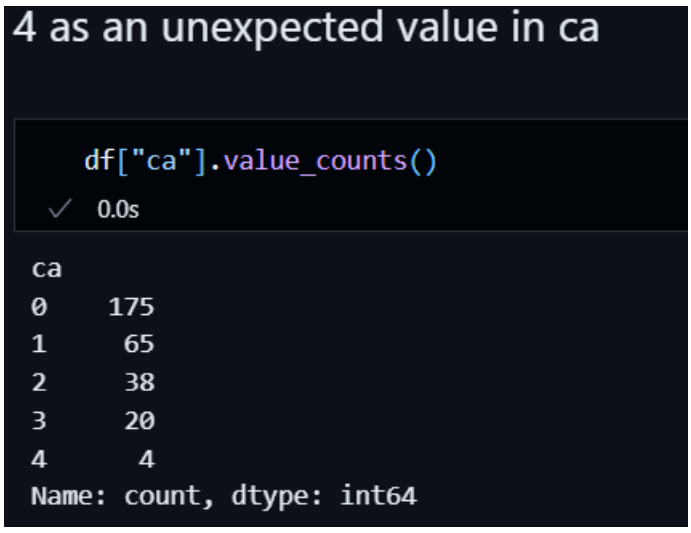


Figure 6: 4 as an unexpected value in Ca (Number of Major Vessels Colored by Fluoroscopy)

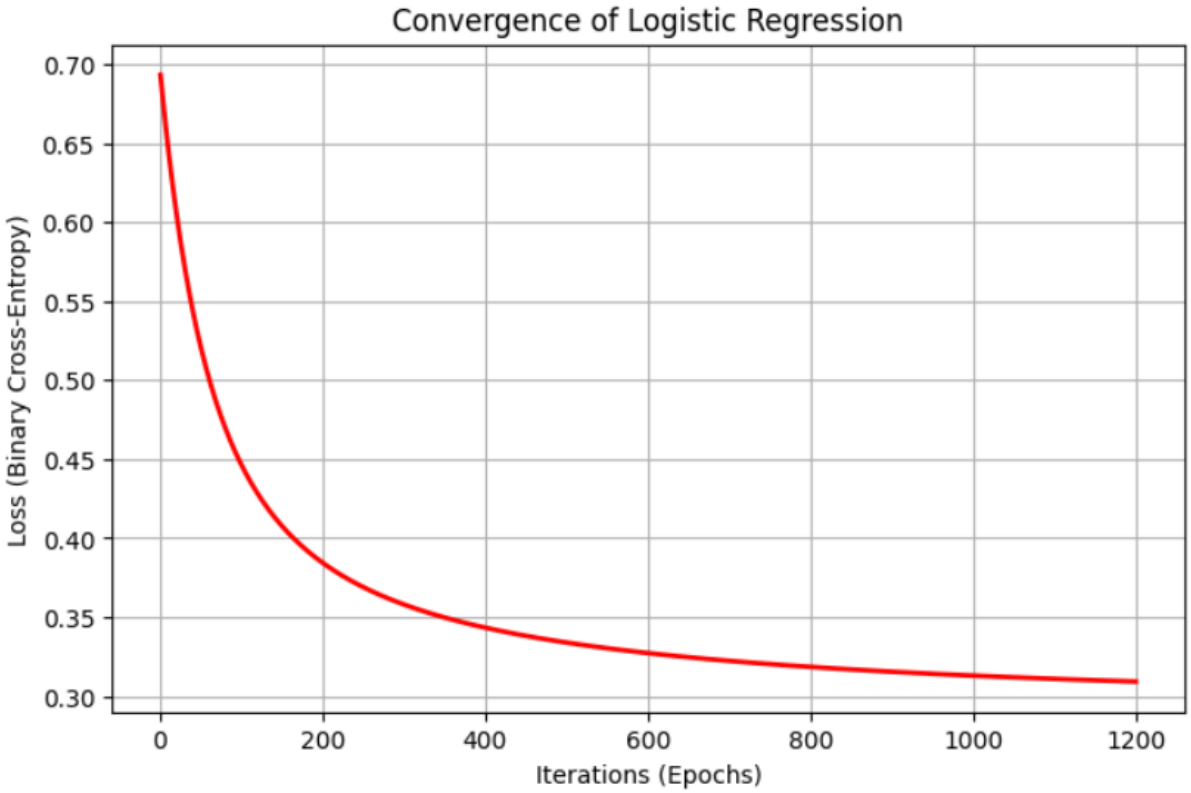
The number of main vessels (0–3) colored by fluoroscopy was represented by the 'ca' feature. However, value 4 appeared as an unexpected value in the dataset. It was found that

value 4 did not align with standard medical definitions. The occurrence of 4 was counted, and it was found to appear only four times in the newly cleaned dataframe, suggesting that it may have been an input error or misclassification. Since the 'ca' feature was ordinal (ordered categories), it was decided to avoid distorting the dataset by substituting the value 4 with the most common valid value (mode), which was 0.

Data Cleaning and Preprocessing

Duplicate rows were dropped to ensure data quality. The features for chest pain, electrocardiographic results, and thal types were all one-hot encoded. Outliers in the resting blood pressure, cholesterol, and oldpeak variables were capped at 170, 370, and 4, respectively, all set to the upper bound. This approach acknowledged the presence of extreme cases while ensuring that these outliers would not negatively impact model performance. The Logistic Regression model required feature set scaling before its application to ensure features maintained equal model contribution strength. Standardization techniques produced scale adjustments which normalized all features to maintain zero mean and one standard deviation. The feature scaling process utilized the StandardScaler tool from the sklearn.preprocessing module. A standardized training data set was used to fit the scaler before applying it to both the training dataset and testing data to achieve identical normalization. Logistic Regression and gradient-based models require feature scaling for optimal performance thus standardization played an essential role in creating a stable training process.

Logistic Regression Performance



The graph demonstrates how logistic regression approaches convergence throughout training while monitoring binary cross-entropy loss at each iteration. The y-axis represents binary cross-entropy loss while the x-axis presents iteration count. Logistic regression applies the sigmoid function to transform input data into class predictions that range between 0 and 1 as probability values. Evaluation of the model performance happens through the binary cross-entropy loss which determines the accuracy of predicted probabilities against actual labels. The loss value begins at 0.7 because the randomly initiated weights perform poorly. Training duration results in quick loss reduction until the model reaches a value of 0.3 when weight optimization completes. The reduced loss value signifies that the model now produces more precise predictions. The loss reduction gradually declines while the curve starts flattening which indicates the model is nearing its best possible solution.

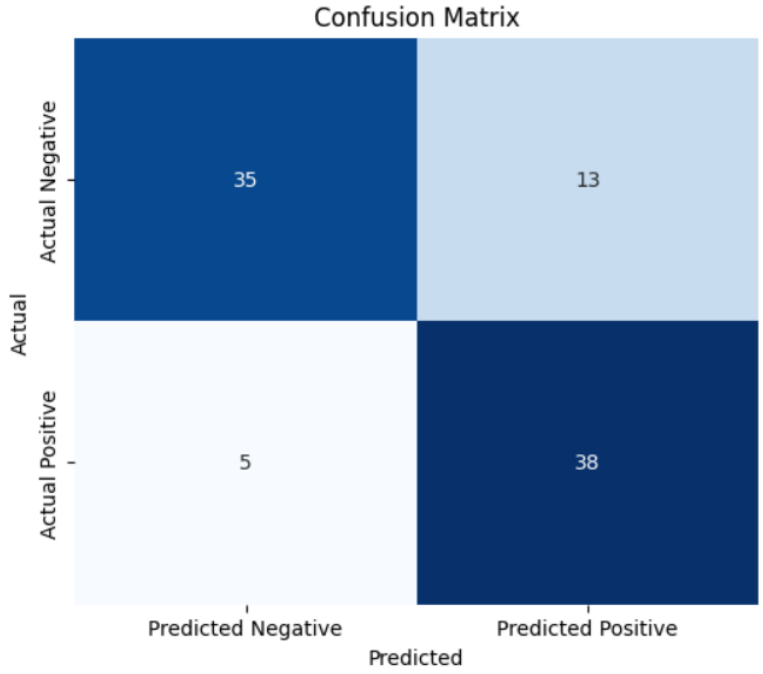


Figure 7: Confusion Matrix of Logistic Regression Model

The logistic regression model demonstrated 87.20% accuracy in training yet reached 80.22% accuracy in testing the previously unseen data. According to the confusion matrix the model properly recognized 38 TP cases along with 35 TN cases yet it misidentified 13 FP cases combined with 5 FN cases.

Similarly, precision revealed that the model distinguished 74.51% of actual positive outcomes correctly. The recall result amounted to 0.8837 which indicated the model correctly detected 88.37% of genuine positive cases. The combined score of precision and recall known as F1-score showed 0.8085 which indicates a productive blend between the two metrics.

The log loss value of 0.4172 supports that model prediction probabilities aligned well with actual labels yet requires additional improvement for more accurate probabilistic output.

The model exhibited high competency in detecting actual positive cases yet some bias emerged through its inaccurate identification of cases.

Primary Models: Decision Tree & Random Forest Classification

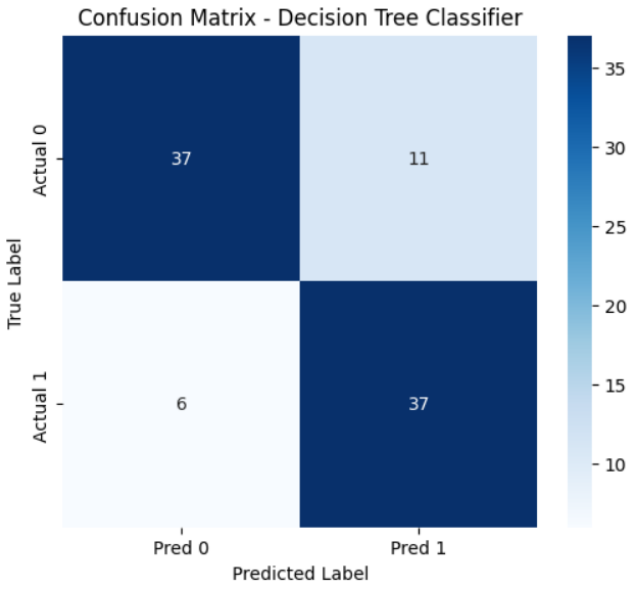


Figure 8: Confusion Matrix - Decision Tree Classifier

The Decision Tree Classifier achieved a perfect accuracy rating of 1.0000 on the training set while recording a lower accuracy level of 0.7363 on the test set and displaying signs of overfitting. The model experienced challenges in distinguishing between two classes as revealed through 37 TP and 37 TN together with 6 FN and 11 FP. The classification report displayed 0.83 precision along with 0.62 recall for class 0 yet 0.67 precision with 0.86 recall for class 1 which demonstrated that the model traded off precision values against recall rates. Overall the Decision Tree model achieved acceptable results particularly with regard to class 1 recall but its ability to generalize would likely improve through additional hyperparameter adjustments.

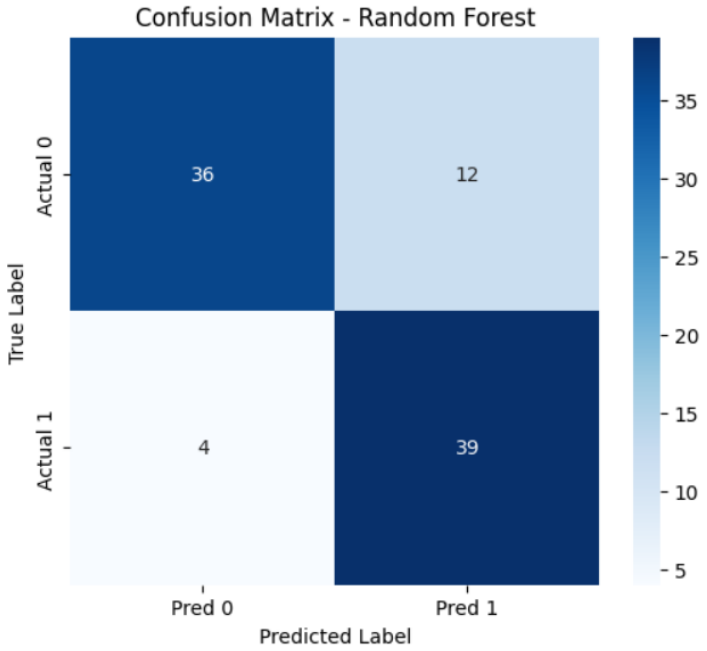


Figure 9: Confusion Matrix - Random Forest Classifier

Evaluation of the Random Forest Classifier on the same dataset yielded superior results than the Decision Tree model with 1.0000 training accuracy and 0.8242 test accuracy. The analysis through the confusion matrix showed 36 true positives (TP) and 39 true negatives (TN) and 12 false positives (FP) together with 4 false negatives (FN). This indicates the model demonstrated strong accuracy in its prediction results. The classification report showed class 0 achieved precision of 0.90 and recall of 0.75 while class 1 reached precision of 0.76 with recall of 0.91 presenting a desirable balance especially for class 1 with its strong recall performance. Both class F1-score results were excellent as the model achieved an overall accuracy of 0.82. The Random Forest model provided improved precision/recall balance compared to the Decision Tree because it achieved better generalization results on the test set. The Random Forest delivered better precision and recall results than the Decision Tree on class 1 thus demonstrating its superior capability when processing the dataset.

The Random Forest model exceeded the Decision Tree classifier because it achieved 0.8242 as its test accuracy measurement while the Decision Tree reached only 0.7363. Random Forest achieves better performance through its method of combining prediction outputs from different decision trees which decreases model overfitting while improving the

ability to project to new data. The model achieved superior performance through its high precision and recall metrics which resulted in reduced number of false positives and false negatives specifically for class 1 with precision of 0.90 and recall of 0.91 when compared to the Decision Tree. The F1-score values reached higher levels for both classes within the Random Forest predictor which shows balanced precision and recall capability. The Random Forest classifier employed better propensity against variance and bias characteristics to yield better prediction accuracy and dependability.

Hyperparameters optimization and Feature Selection

The hyperparameter optimization utilized GridSearchCV for Decision Trees and RandomizedSearchCV for Random Forests. The Decision Tree model received optimization through these selected set of optimal hyperparameters: 'gini' along with Max Depth value of 10 and Min Samples Leaf at 2 along with Min Samples Split set to 10. Criterion: 'gini', Max Depth: 10, Min Samples Leaf: 2, Min Samples Split: 10

The Random Forest model produced the following optimal set of parameters after RandomizedSearchCV optimization: N Estimators: 100 Min Samples Split: 10 Min Samples Leaf: 1 Max Features: 'log2' Max Depth: 12 Criterion: 'gini' Bootstrap: True

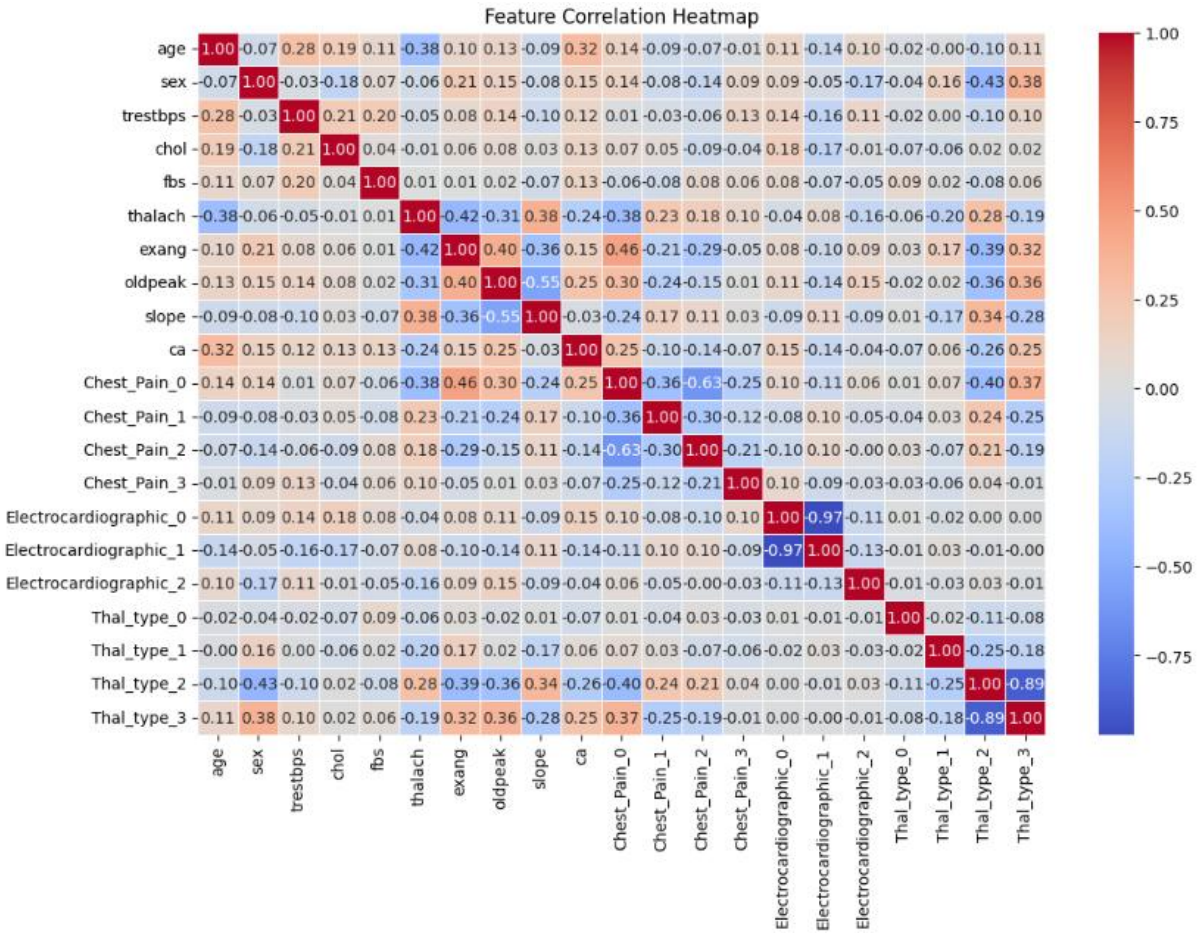


Figure 10: Feature Correlation Heatmap

The model received performance improvements along with overfitting reduction through a two-step feature selection process. The first step involved conducting a correlation analysis that permitted the removal of highly linked dataset features. The highly associated features Thal_type_3, Electrocardiographic_1, Electrocardiographic_0 and Thal_type_2 were eliminated from the dataset through data selection. The performed procedure eliminated duplicate data which enabled the model to avoid depending on repetitive information.

A feature importance evaluation was conducted for both Decision Tree and Random Forest models. The Random Forest model recognized thalach, ca, Chest_Pain_0, age, chol, oldpeak, trestbps, slope, exang and sex as the ten most significant features based on their influence on modeling accuracy. The Decision Tree model identified Chest_Pain_0 followed by oldpeak then trestbps, age, thalach, chol, ca and sex, Thal_type_1 and exang as its important features.

Rebuilding the Best Model with Optimized Features & Hyperparameters

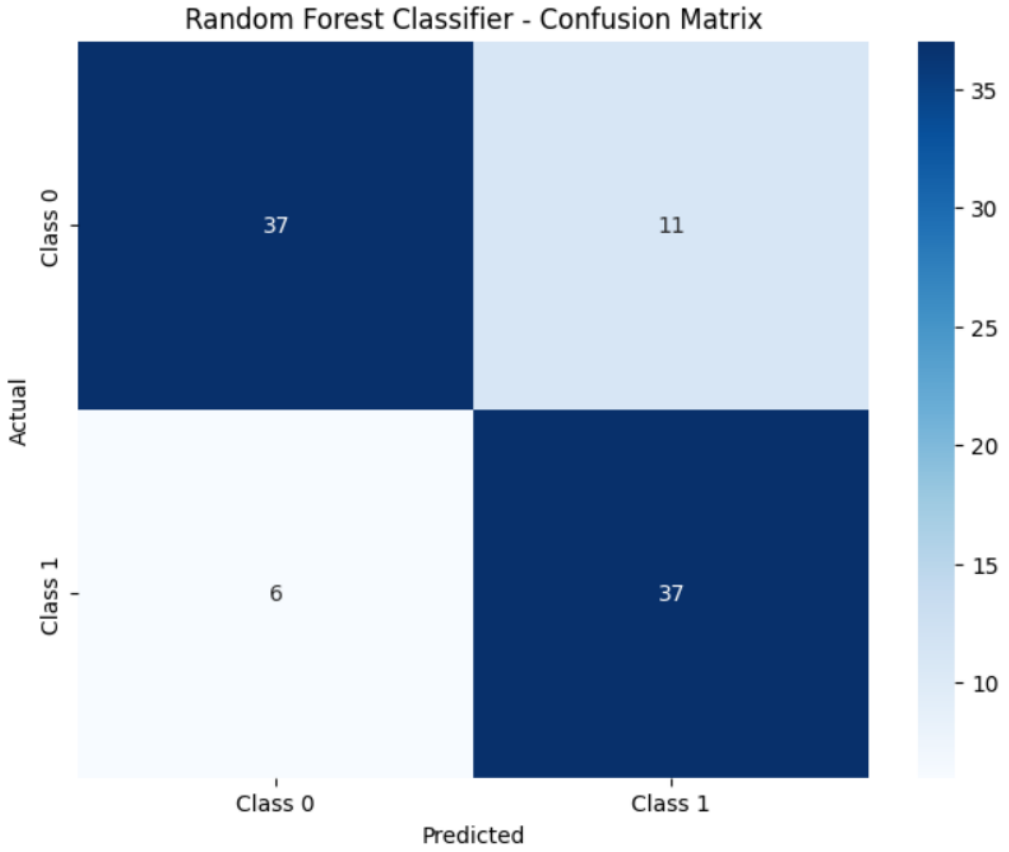


Figure 11: Final Model

The Random Forest classifier excelled over all other models so the model's rebuilding phase implemented a new set of selected hyperparameters and features.

The retrained Random Forest classifier displayed 94.31% accuracy when learning from the training data and achieved 81.32% accuracy when predicting from the test data. The model demonstrated good predictive power on fresh datasets without deteriorating its performance on previously used data.

A test set classification report delivered extra information that revealed the model's operating characteristics. Both classes achieved balanced F1-score performance from the Random Forest classifier based on precision and recall metrics which measured at 0.86 and 0.77 for class 0 and 0.77 and 0.86 for class 1. The model delivered equal detection skills across both classes yet demonstrated superior precision towards class 0.

The complete dataset achieved average performance levels through macro average and weighted average F1-scores of 0.81. The model showed a balanced performance since

it monitored precision and recall consistently while simultaneously maintaining optimal values between them. Such behavior is beneficial for handling classification tasks with equal classes.

Performance enhancements demonstrate the effectiveness of the applied hyperparameter optimization and feature selection methods which showcase Random Forest classifier potential to generate accurate predictions.

Conclusion

The Random Forest model achieved better generalization after optimizing its hyperparameters and choosing relevant features from the dataset which reduced the overfitting issue without compromising its predictive capabilities. The training accuracy dropped to 94.31% which showed the model had stopped memorizing training data as it adopted more versatile patterns for generalization. The transition into an enhanced model state resulted in a foreseeable decrease of test accuracy from 82.42% to 81.32%. The predictive performance of the model maintained a steady equilibrium as class 0 demonstrated 86% precision with 77% recall while class 1 showed 77% precision with 86% recall to minimize misclassification mistakes. The model became more interpretable through feature selection because it identified the most important attributes while maintaining adequate performance outcomes. Regardless of a slight decrease in test accuracy, the optimized Random Forest model created a dependable predictive structure that exhibited robust generalization performance.

Given that the dataset was quite small, with only 302 rows, and there were limitations in both domain knowledge and model expertise, the classification task was completed to the best possible extent. Despite these challenges, the models managed to generalize well, showing promising results. Future efforts could focus on trying out different models like Gradient Boosting, fine-tuning hyperparameters with more advanced techniques, or even experimenting with ensemble methods to boost performance. Expanding the dataset, whether through additional data collection or synthetic data generation, could also help make predictions more reliable. Moreover, incorporating deeper domain insights could refine feature

selection and preprocessing, leading to even better results. While the models performed well given the constraints, there’s still plenty of room for improvement.