A Comparative Analysis of Regression 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 regression models and analyze the impacts of hyperparameter optimization and feature selection on model performance. The primary goal is to predict resting blood pressure based on various patient attributes using a dataset from Kaggle, originally sourced from the UCI Machine Learning Repository. The dataset contains patient information collected from multiple medical institutions. This research focuses on comparing three regression models, including a Linear Regression model (implemented from scratch), Decision Tree Regressor, and Random Forest Regressor. An evaluation of the impact of hyperparameter tuning and feature selection on 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 into heart disease risk assessment while supporting UN Sustainable Development Goal 3: Good Health and Well-being through data-focused healthcare practices.

Contents

Introduction

Heart health represents the fundamental basis of overall well-being since it ensures proper functioning of the body. The heart distributes oxygen-rich blood throughout the entire body therefore its smooth operation stands vital for sustaining biological existence. Multiple lifestyle and genetic elements increase a person's heart disease susceptibility which stands as the main reason behind worldwide health difficulties.

The main purpose of this research project involves developing an approach to forecast resting blood pressure measurements from existing medical traits. The approach targets the analysis of predictive model capabilities together with performance evaluation alongside identification of operational restrictions and determining influential determining elements. Analysis of the database and utilization of diverse regression models will provide understanding about which factors affect resting blood pressure the most along with how various models understand these connections.

The research uses a sequential method involving exploratory data analysis (EDA) then data cleaning and preprocessing steps to achieve high-quality input information. Fundamental analysis of model development involves comparison between the baseline Linear Regression and Decision Tree and Random Forest regression models. A thorough evaluation of how model performance changes when hyperparameter values are adjusted and relevant features undergo selection is conducted.

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.

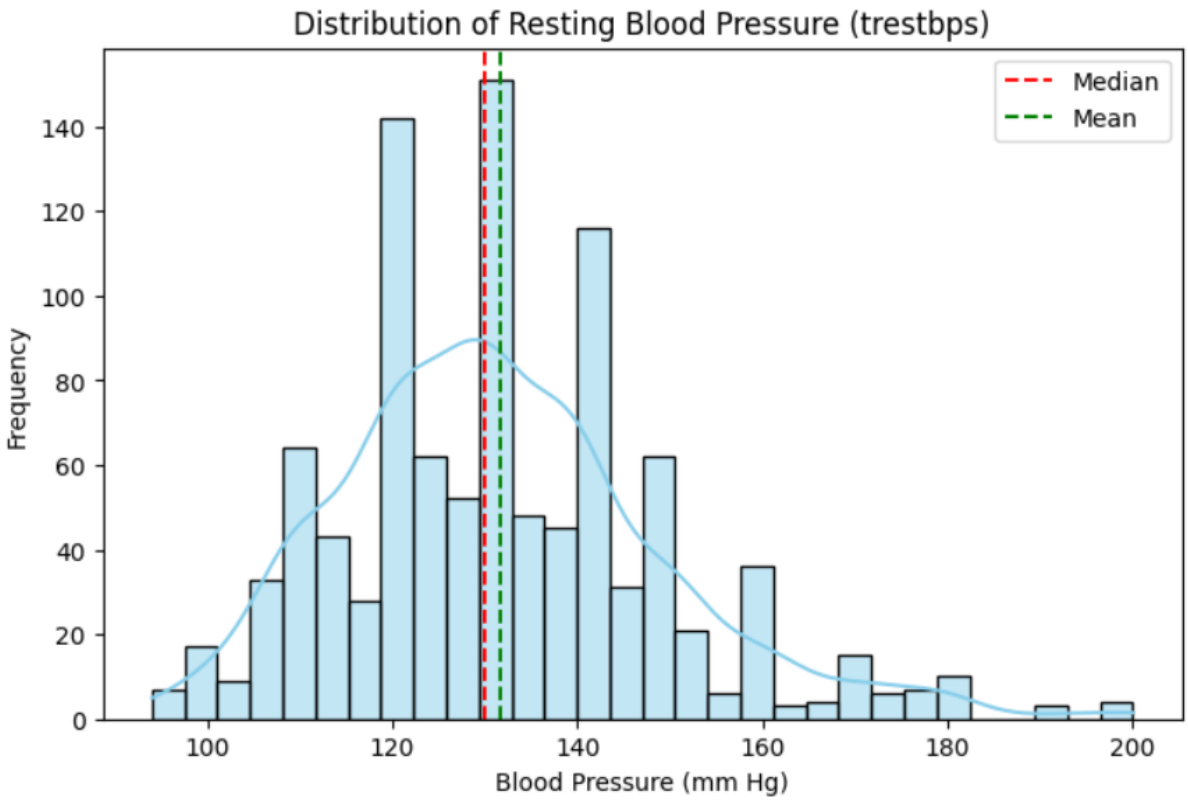

Figure 1: Distribution of Resting Blood Pressure

The data shows resting blood pressure (trestbps) distribution through both a histogram and kernel density estimate (KDE) to visualize its overall shape. The horizontal x-axis shows blood pressure values (in mm Hg) running from 100 to 200 mm Hg simultaneously as the vertical y-axis displays the number of observed occurrences per defined interval. The bars from the histogram show blood pressure value frequencies but the smooth blue curve from KDE functions as a probability density function which indicates a roughly normal distribution concentrated near 120 mm Hg. The histogram features two reference lines to show central tendency measures through the red dashed line at 130 mm Hg for the median that separates the data into two equal parts while the green dashed line marks the mean above the median level suggesting a positive skew. The skewing of the distribution shows that blood pressure readings at 120 to 140 mm Hg form the majority group although extreme higher readings elevate the average value. The distribution pattern shows typical clustering of values between certain limits but higher outlier observations alter the mean computation. The visual

presentation makes it possible to comprehend both the extent of blood pressure dispersion and the typical blood pressure values contained within the data collection.

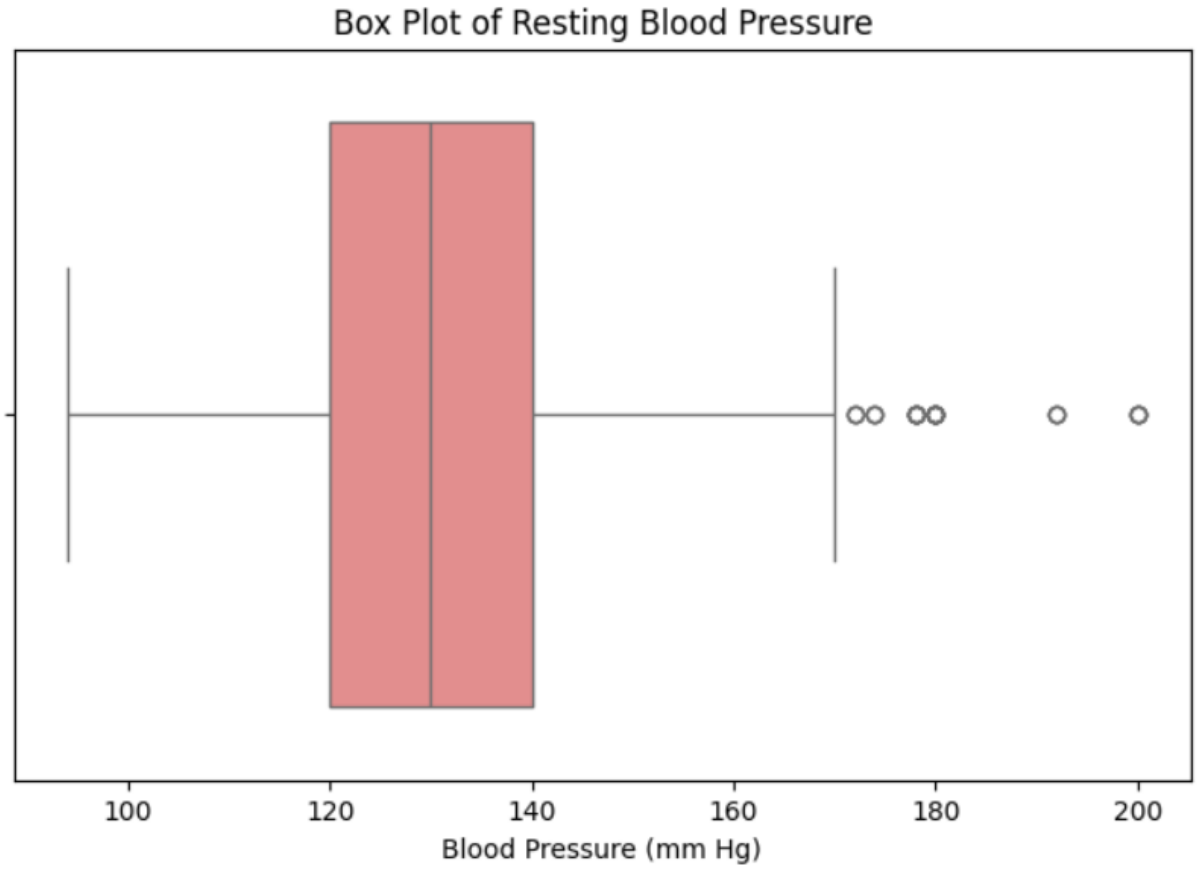

Figure 2: Boxplot of resting blood pressure

The resting blood pressure distribution (trestbps) appears in this box plot together with its spread characteristics and central tendency along with its group of extreme values. The median value of 130 mmHg rests inside the box as the interquartile range (IQR). This range extends between the 25th percentile (Q1) and the 75th percentile (Q3) where the middle 50% of data lies. One vertical line inside the box displays the median (50th percentile) value at 130 mm Hg. The vertical extent of the whiskers spans from the smallest value at 100 mm Hg to the largest value at 160 mm Hg and includes most of the recorded measurements. Several extreme cases show abnormally high blood pressure readings by exceeding 160 mm Hg, thus appearing above the maximum whisker. A majority of resting blood pressure measurements between 120-140 mm Hg form the main distribution pattern as shown by the data points having

a symmetrical spread. Medical analysis considers high outliers as important indicators of hypertension risk in specific patient groups.

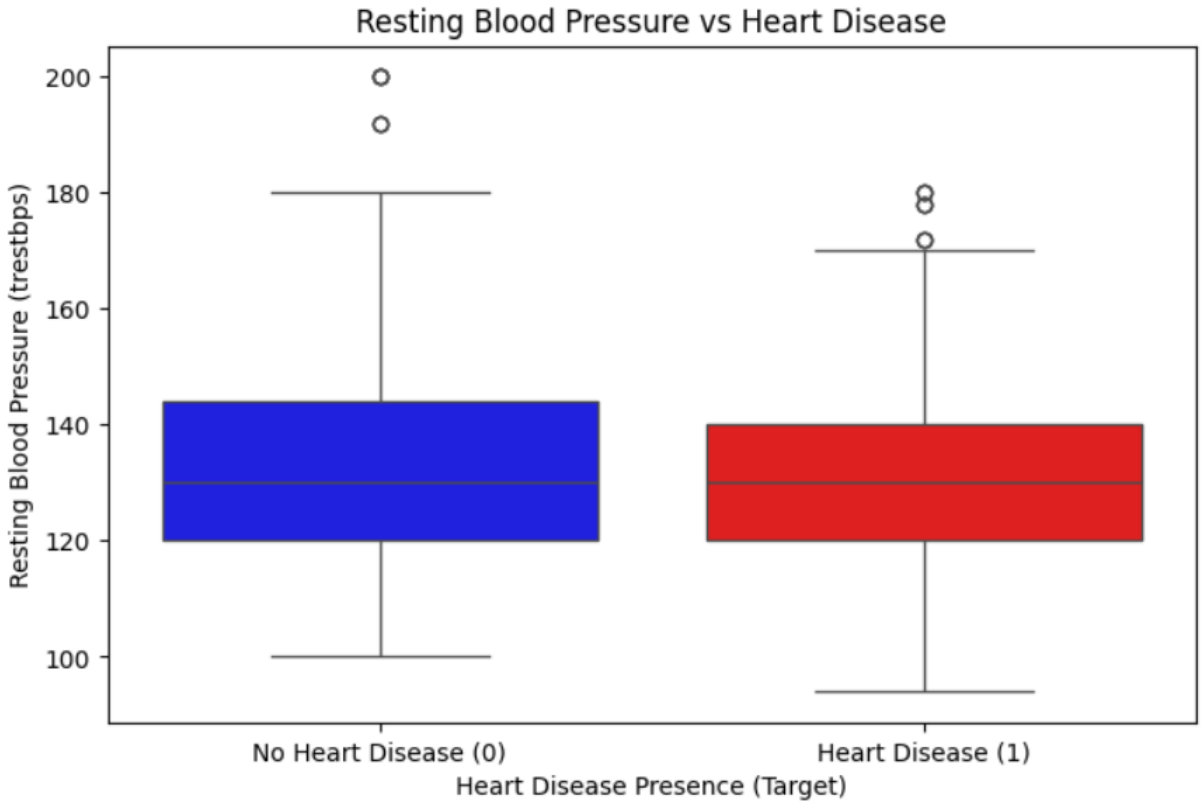

Figure 3: Resting Blood Pressure vs Heart Disease

The box plot shows resting blood pressure trestbps frequencies among heart-disease-positive and negative participants to differentiate their measurement patterns. The blue box highlights people without heart disease who exhibit a median blood pressure level of 130 mm Hg with a distribution ranging from 120 to 140 mm Hg and extending from 100 mm Hg to 160 mm Hg with a few points exceeding 160 mm Hg. Heart disease patients exhibit a similar IQR to healthy persons but demonstrate greater variability in blood pressure levels since their median rests at about 140 mm Hg while extending to larger values. The heart disease group contains more instances of people with extremely elevated blood pressure measurements which suggests an association between high hypertension and heart disease presence. The blood pressure ranges of both groups intersect although statistical measures show heart disease patients have elevated median pressure and more extreme blood pressure readings

than individuals without heart disease so there is a potential correlation between hypertension and heart disease.

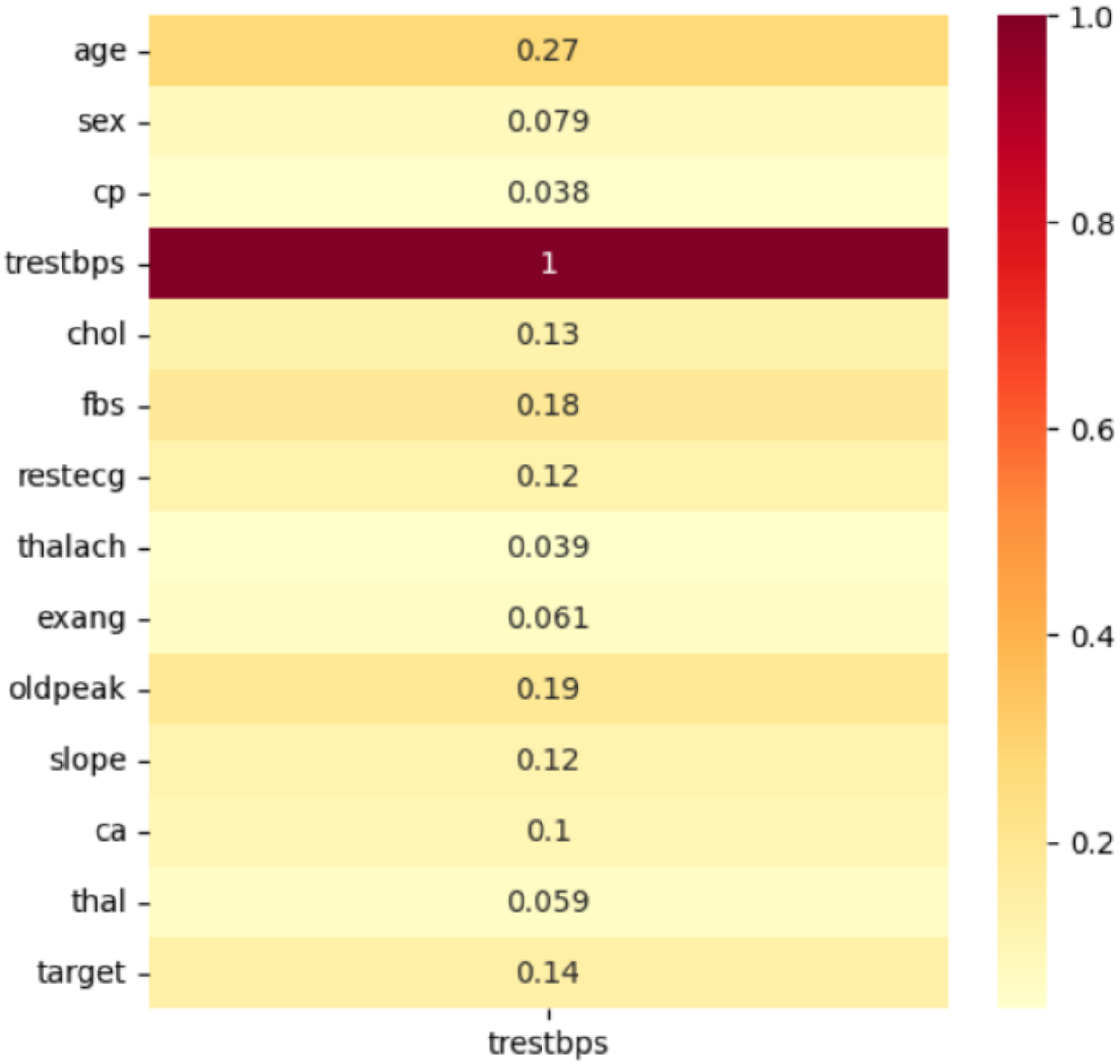

Figure 4: Correlations between features

The map displays visual correlations that exist between features age, sex, cp and trestbps with respect to the feature trestbps which stands for resting blood pressure. The color scale located on the right side of the heatmap uses yellow tones for indicating weak relationships and red tones for showing stronger positive correlations.

The self-correlation value of "trestbps" becomes the reference point for measuring all other correlations because it is perfectly correlated (1). The strength of association among other features appears moderate to low when compared to resting blood pressure. The

correlation measurement between age and resting blood pressure equates to 0.27 which demonstrates an existence of a not particularly robust positive relationship between increasing age and increasing blood pressure levels. The relationship between sex and resting blood pressure measures 0.079 thus demonstrating minimal connection between the two variables.

The heatmap data indicates that "trestbps" possesses strong self-correlation but the remaining features do not demonstrate substantial linear relationships with it. According to this heatmap these features are unlikely to operate as independent forecasting variables for resting blood pressure values. Correlation itself does not prove causation since the relationship between two variables does not establish cause-effect connections even when they demonstrate statistical correlation. Linear relationships form the basis of understanding from this heatmap because it does not detect non-linear relationships between variables.

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.

Linear Regression Performance

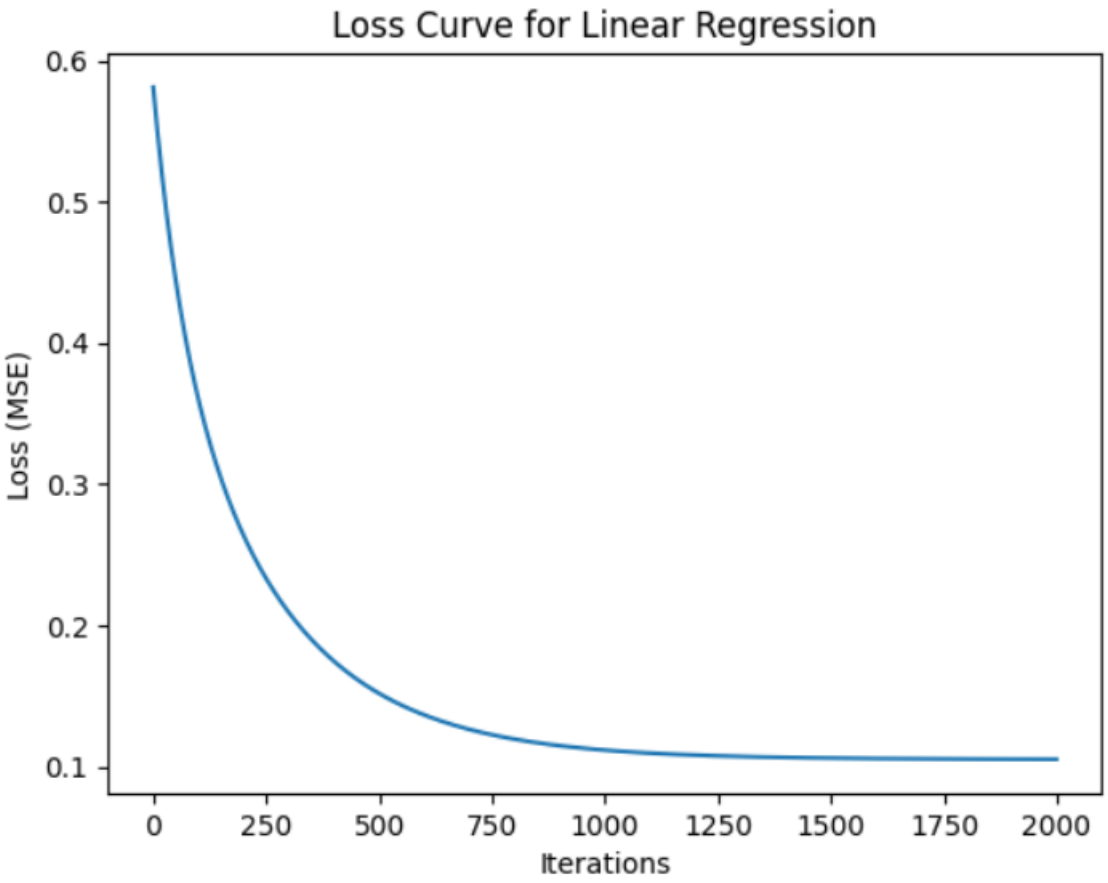

Figure 5: Loss Curve for Linear Regression

The linear regression loss curve begins at a high value then declines swiftly in the first part of the training due to rapid model adjustment and performance enhancement. The reduction in loss rate slows down during successive training steps because the model evolves towards its best possible outcome. When the loss curve stops decreasing the model reaches convergence since significant improvements stop occurring. Gradient descent optimization follows this pattern through which models efficiently optimize Mean Squared Error (MSE) to obtain the best-fitting line while additional performance gains become minimal.

The performance metrics offered detailed understanding regarding how well the model functions. The Mean Absolute Error (MAE) in the training data showed predictions deviating from actual values to 0.2511 while Root Mean Squared Error (RMSE) displayed 0.3242 thus indicating an average deviation of 0.25–0.32 units. The R^2 score stands at 0.5703 which indicates that the model understands and accounts for 57% of the data variations yet leaves

a major portion of unexplained data variations. The test set shows lower predictive performance since its errors reach an MAE of 0.3006 and RMSE of 0.3747 and R^2 score of 0.4368. The model demonstrates an overfitting effect because it achieves improved results during training yet it fails to maintain similar performance when processing new data.

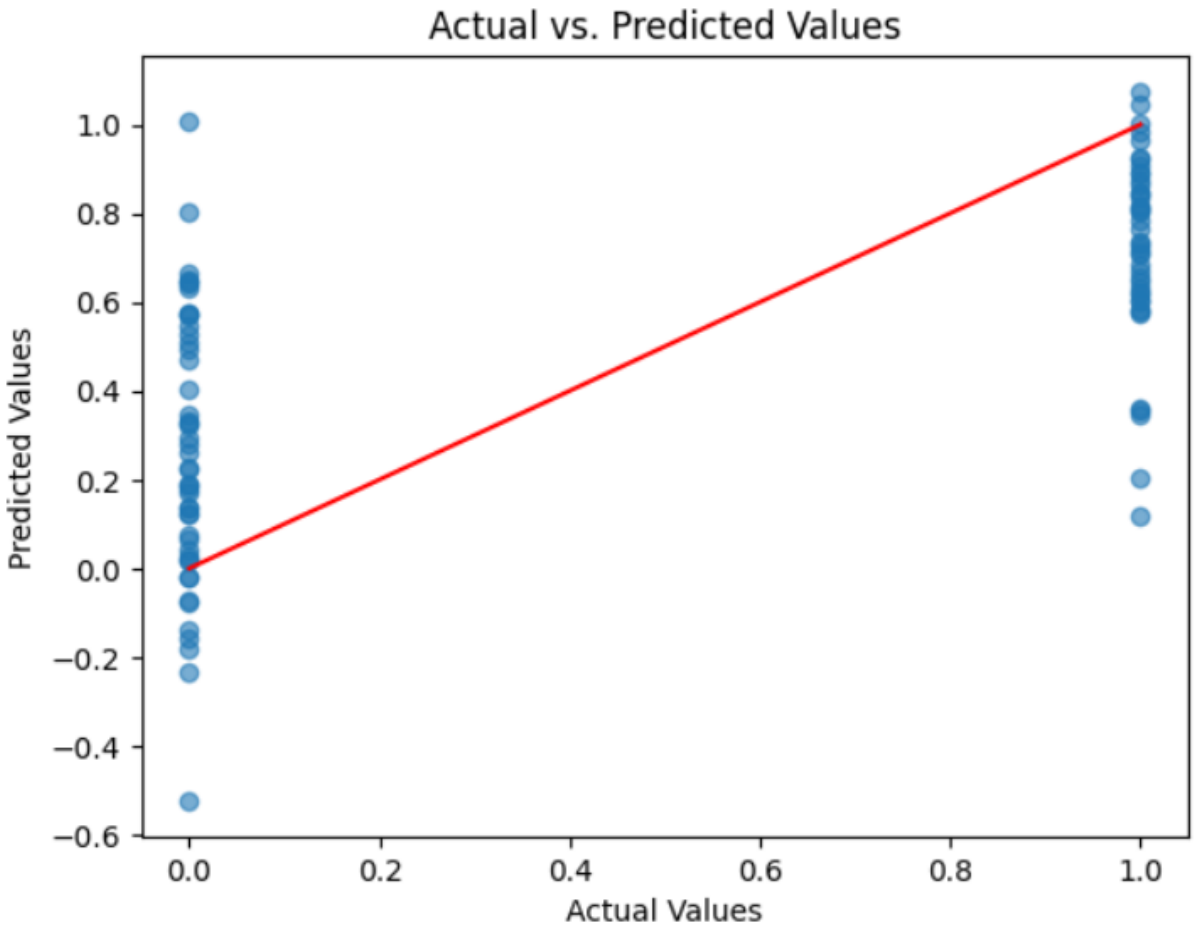

Figure 6: Actual vs Predicted values

The analysis presents actual data points alongside predicted output values from the linear regression model. The distribution of most data points around the origin (0, 0) demonstrates accurate prediction of small values yet numerous points remain clustered in the top-right corner (1, 1) where predicted values reach the red line. Some points are not aligned with the red line although the model shows good general fit by tracking the overall pattern. Linear regression does not seem to be the right model for this task.

Primary Models: Decision Tree & Random Forest Regressor

The results indicate that Linear Regression was not capable of capturing the underlying patterns in the dataset, and this necessitated the exploration of more complex models. The Decision Tree Regressor had severe overfitting, achieving a perfect training R^2 of 1.0 and a very low test R^2 of 0.0741, meaning that it had successfully memorized the training data but failed to generalize. On the other hand, the Random Forest Regressor provided a more balanced performance with a training R^2 of 0.9191 and a test R^2 of 0.4620, with much better generalization and less error. This shows that Random Forest is the best model so far, with more variance in the test data explained without the overfitting seen in the Decision Tree.

Hyperparameter Optimization and Feature Selection

Following hyperparameter tuning, Decision Tree Regressor was also tuned using GridSearchCV and it gave back the optimal parameters as criterion='squared_error', max_depth=5, max_features='sqrt', max_leaf_nodes=20, min_samples_leaf=2, min_samples_split=5, min_weight_fraction_leaf=0.0, and splitter='random'. For Random Forest Regressor, RandomizedSearchCV was utilized for exploring the space of hyperparameters in an efficient manner, and it gave back the optimal set of parameters: n_estimators=400, min_samples_split=2, min_samples_leaf=2, max_features='sqrt', max_depth=20, and bootstrap=True.

I first applied the filter approach to remove highly correlated features by detecting and removing Thal_type_2, Electrocardiographic_1, Electrocardiographic_0, and Thal_type_3 to remove redundancy and increase model effectiveness. Then, I performed feature importance for the Decision Tree Classifier and Random Forest Regressor to select the most significant features. The Random Forest model ranked Chest_Pain_0, ca, oldpeak, age, thalach, chol, trestbps, sex, slope, and exang as the top 10 most important features. The Decision Tree model, however, ranked Chest_Pain_0 as the most significant feature, followed by oldpeak, trestbps, age, thalach, chol, ca, sex, Thal_type_1, and exang.

Rebuilding the best model with optimized features and hyperparameters

After the reconstructed Random Forest model using the tuned hyperparameters and selected features, the performance metrics have improved slightly from the initial results. The training Mean Squared Error (MSE) is currently at 0.0475, and the training R^2 score is 0.8059, indicating a good fit on the training data, though the model is not overfitting as much as before. The test Mean Squared Error (MSE) is 0.1341, and the test R^2 score is 0.4618, showing a small improvement in the test performance, with the model continuing to have a modest ability to generalize to unseen data. In comparison with the last performance where the training R^2 was 0.9191 and the testing R^2 was 0.46197, the model here is more balanced in its performance with less overfitting, hence a better model (but not the best) without losing its generalization capability for the test data.

Conclusion

In general, the Random Forest model showed a tremendous improvement in performance compared to earlier versions, especially in training results. However, one thing must be remembered that the data for this analysis had only 302 rows, which is too few to train a robust model. Because of this limited dataset, the model could not generalize well, which most probably led to the average testing performance. The outcomes are fine but express the problems of dealing with a small dataset. This can lead to overfitting, where the model will not generalize well to new data. In the future, I understand that I need to try out other models, implement cross-validation techniques, and improve the manner in which I select features. In addition, learning more about the specific area will be very important in understanding the trends of the data. This can be utilized to choose the right models and make them perform better. Trying out different things and applying expert knowledge will be needed to overcome the problems posed by small datasets. This will help achieve better generalization and prediction accuracy.