Final Project

Heart disease prediction using machine learning

Naveen Teja Reddy Majji

Webster University

03/13/2025

**Table of Contents**

[**Executive Summary** 4](#_Toc190356951)

[**Introduction** 4](#_Toc190356952)

[**Data Description** 5](#_Toc190356953)

[**Dataset Overview** 5](#_Toc190356954)

[**Data Preprocessing** 5](#_Toc190356955)

[**Exploratory Data Analysis (EDA)** 6](#_Toc190356956)

[**Univariate Analysis** 6](#_Toc190356957)

[**Bivariate Analysis** 10](#_Toc190356958)

[**Outlier Detection** 12](#_Toc190356959)

[**Modeling** 12](#_Toc190356960)

[**Logistic Regression** 12](#_Toc190356961)

[**Decision Tree** 13](#_Toc190356962)

[**Random Forest** 14](#_Toc190356963)

[**Performance Evaluation** 14](#_Toc190356964)

[**Accuracy Metrics** 14](#_Toc190356965)

[**Confusion Matrices** 15](#_Toc190356966)

[**ROC and AUC Analysis** 18](#_Toc190356967)

[**Conclusion** 20](#_Toc190356968)

[**References** 21](#_Toc190356969)

**Table of Figures**

[Figure 1: Data Processing 6](#_Toc190356929)

[Figure 2: Age Histogram 6](#_Toc190356930)

[Figure 3: Age Boxplot 7](#_Toc190356931)

[Figure 4: Cholesterol Histogram 8](#_Toc190356932)

[Figure 5: Cholesterol Boxplot 8](#_Toc190356933)

[Figure 6: Max Heart Rate Histogram 9](#_Toc190356934)

[Figure 7: Max Heart Rate Boxplot 10](#_Toc190356935)

[Figure 8: Cholesterol Levels by Heart Disease (true/false) 11](#_Toc190356936)

[Figure 9: Correlation Heatmap 11](#_Toc190356937)

[Figure 10: Outlier Check 12](#_Toc190356938)

[Figure 11: Logistic Regression Model 13](#_Toc190356939)

[Figure 12: Decision Tree Model 13](#_Toc190356940)

[Figure 13: Random Forest Model 14](#_Toc190356941)

[Figure 14: Logistic Regression Accuracy 14](#_Toc190356942)

[Figure 15: Decision Tree Accuracy 15](#_Toc190356943)

[Figure 16: Random Forest Accuracy 15](#_Toc190356944)

[Figure 17: Logistic Regression Matrix 16](#_Toc190356945)

[Figure 18: Decision Tree Matrix 17](#_Toc190356946)

[Figure 19: Random Forest Matrix 18](#_Toc190356947)

[Figure 20: Linear Regression ROC and AUC 19](#_Toc190356948)

[Figure 21: Decision Tree ROC and AUC 19](#_Toc190356949)

[Figure 22: Random Forest ROC and AUC 20](#_Toc190356950)

# **Executive Summary**

The research below explores leveraging machine learning in heart disease prediction using clinical and other factors. The study utilizes three main models for the analysis of patient data. The models used are logistic regression, decision tree and random forest in the identification of risk factors. Logistic regression is utilized to provide a baseline for the other models allowing for more accurate results. The preliminary analysis results show that random forest model outperforms other models in predictive accuracy. Logistic regression on the other hand provides insights into feature performance. The study thus serves to demonstrate the potential of ML-driven diagnostics in healthcare.

# **Introduction**

Heart disease is one of the leading causes of mortality worldwide due to various causes such as predisposition to the condition. Detecting heart disease is often daunting as most patients only exhibit such conditions during the late stages. Early detection and early prediction are thus essential to improving the patient outcomes. The use of traditional diagnostic methods such as clinical imaging and evaluation are often time-consuming and expensive with no tangible results. The recent advancements in data analytics have shown machine learning as a viable tool for identifying patterns in medical data further enabling early diagnosis and risk assessment (Shah et al., 2020). This paper thus focuses on the development of machine learning models to predict heart disease using parameter such as age, cholesterol levels and blood pressure. The objective is to leverage machine learning techniques to enhance heart disease prediction accuracy to aid healthcare professionals in making data-driven decisions. The research is important as it has the potential to reduce diagnostic errors, optimize treatments and enhance preventive healthcare.

# **Data Description**

## **Dataset Overview**

The heart dataset is sources from a public data repository i.e., Kaggle where it has over 1000 observations and includes the following columns.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| age | Age of the patient (in years). |
| sex | Sex of the patient (1 = Male, 0 = Female). |
| cp | Chest pain type (0 = Typical angina, 1 = Atypical angina, 2 = Non-anginal pain, 3 = Asymptomatic). |
| trestbps | Resting blood pressure (mm Hg). |
| chol | Serum cholesterol level (mg/dL). |
| fbs | Fasting blood sugar (>120 mg/dL) (1 = True, 0 = False). |
| restecg | Resting electrocardiographic results (0 = Normal, 1 = ST-T wave abnormality, 2 = Left ventricular hypertrophy). |
| thalach | Maximum heart rate achieved. |
| exang | Exercise-induced angina (1 = Yes, 0 = No). |
| oldpeak | ST depression induced by exercise relative to rest. |
| slope | Slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping). |
| ca | Number of major vessels (0–3) colored by fluoroscopy. |
| thal | Thalassemia type (1 = Normal, 2 = Fixed defect, 3 = Reversible defect). |
| target | Presence of heart disease (1 = Disease, 0 = No disease). |

## **Data Preprocessing**

The data preprocessing step includes checking the structure of the data and finding errors such as missing values. The first step is converting any numerical operators to binomial as most column types are interpreted using default values. The declare missing operator is used to detect any nominal variables with missing data which are then filtered using the filter operator. This results in a well-structured dataset ready for EDA process.

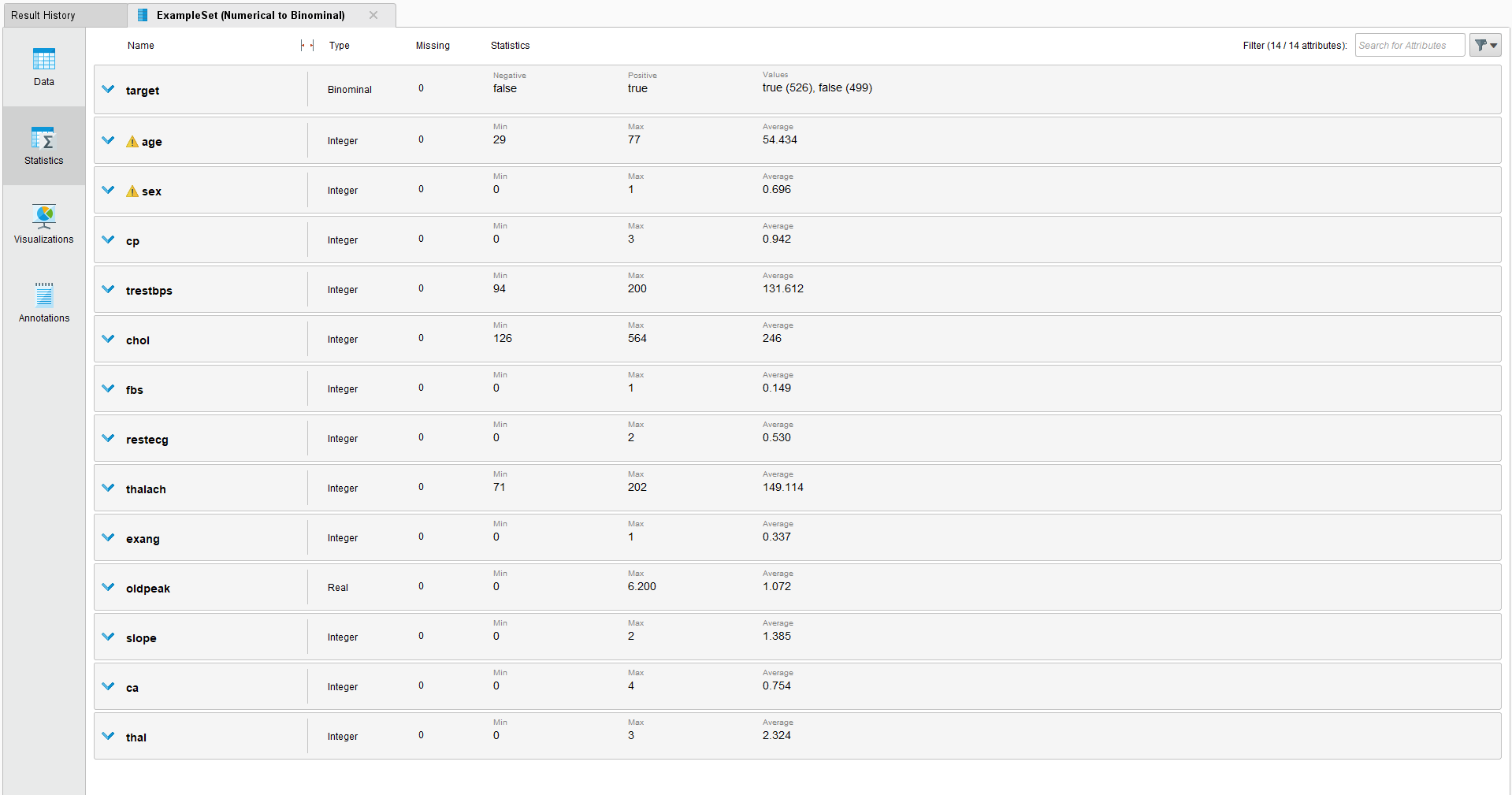


Figure : Data Processing

# **Exploratory Data Analysis (EDA)**

## **Univariate Analysis**

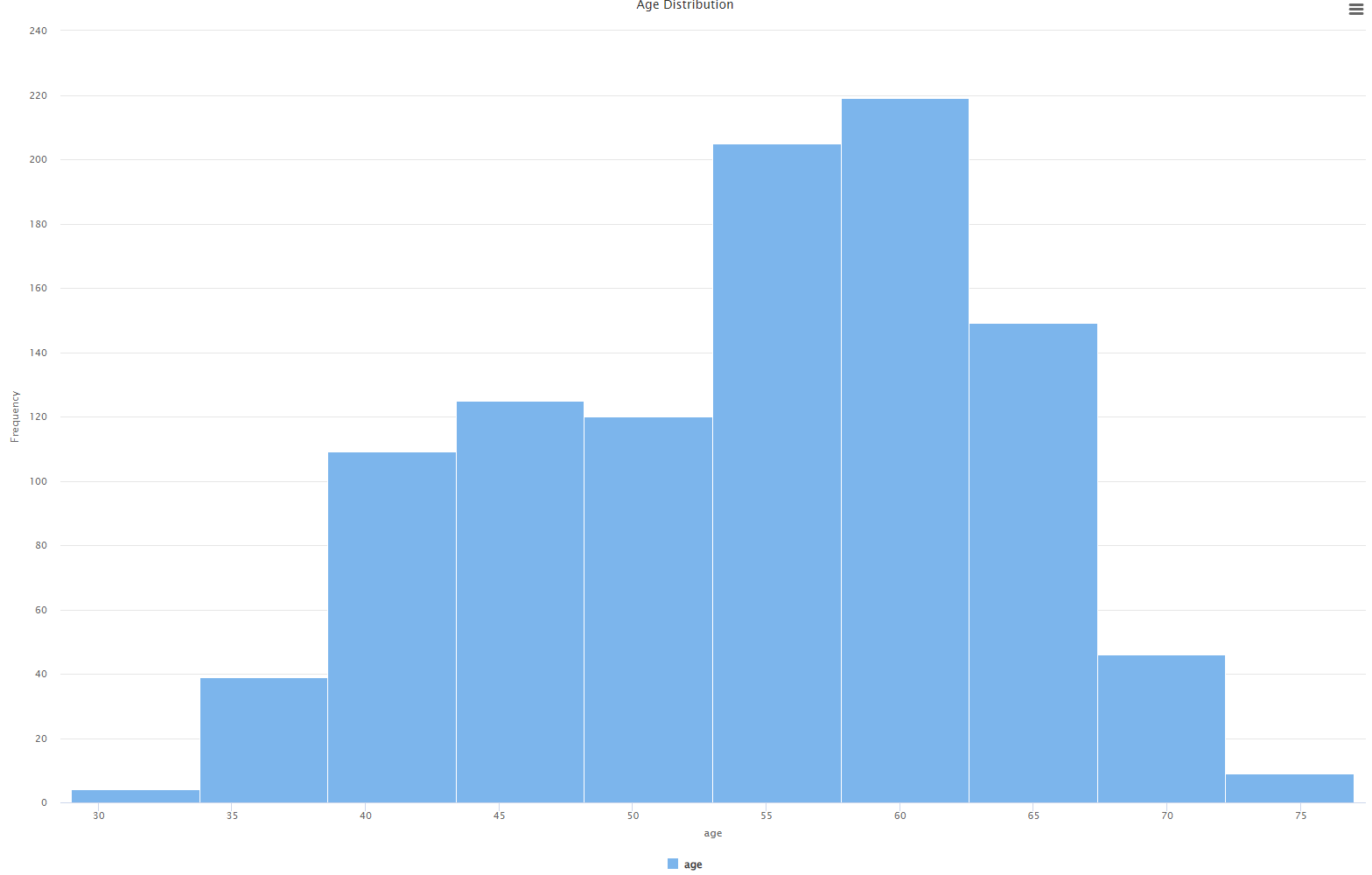


Figure : Age Histogram

The analysis of the age distribution shows that the age data is left skewed as most patients have been between 55 and 65 years old. The distribution also shows some interesting information in that most of the individuals are above 35 years old. These observations can be confirmed in the boxplot below where the median age is around 55 years.

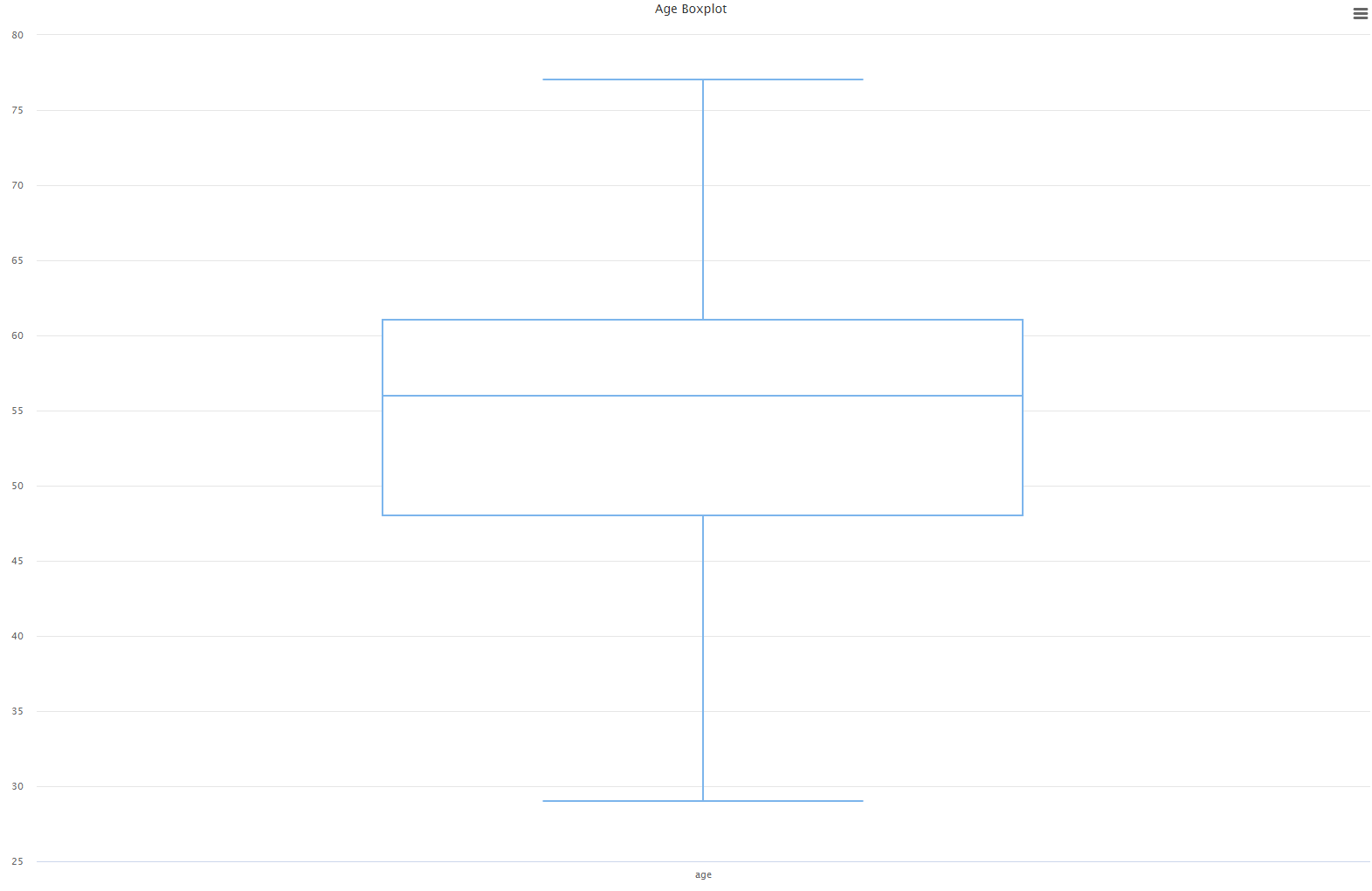


Figure : Age Boxplot

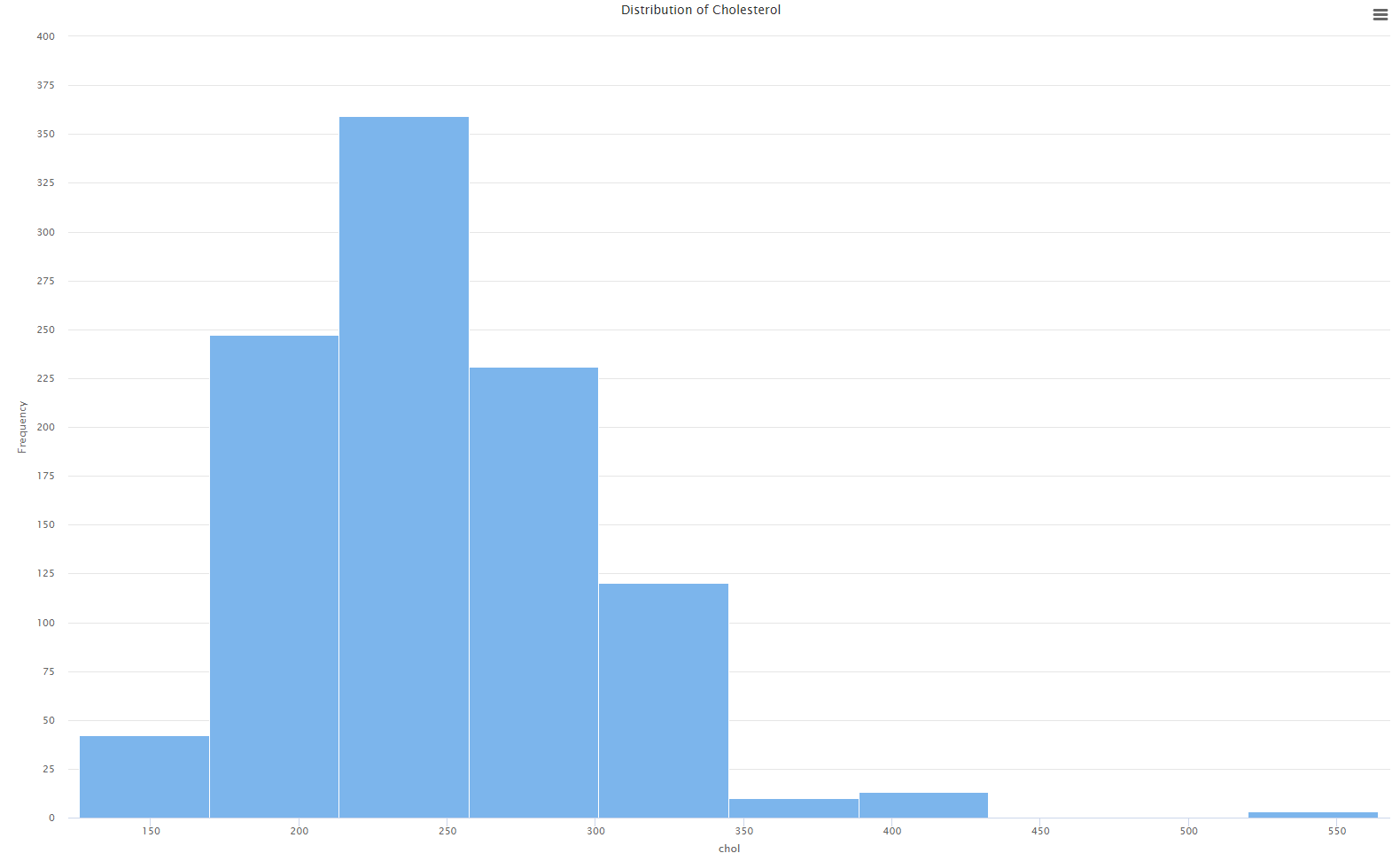


Figure : Cholesterol Histogram

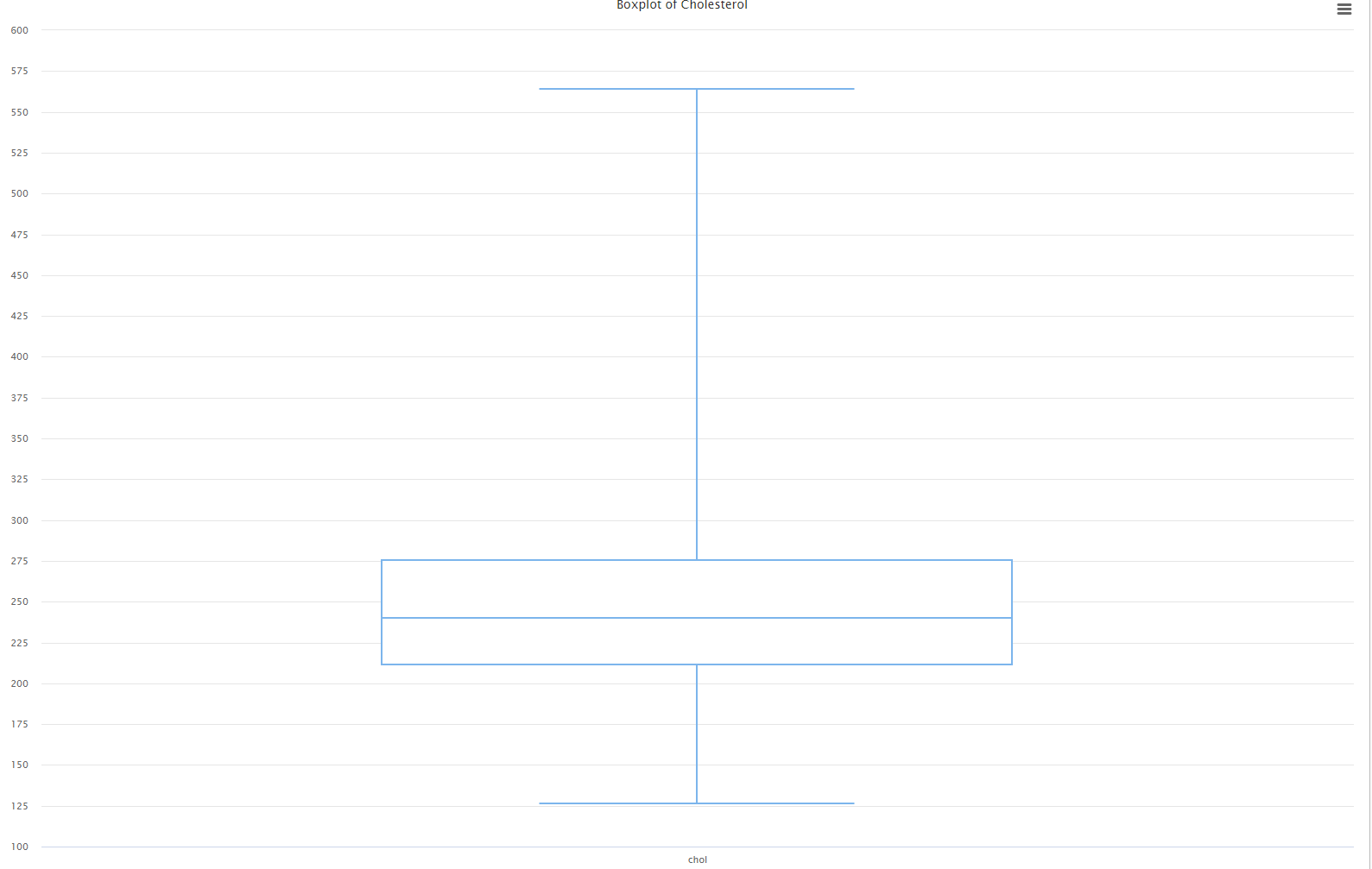


Figure : Cholesterol Boxplot

The analysis of the cholesterol shows a right skewed distribution meaning that the data is not normally distributed. The cholesterol levels are mostly on the low end except for a visible case where some patients have extremely high cholesterol. This is also shown by the longer tail of the box plot above.

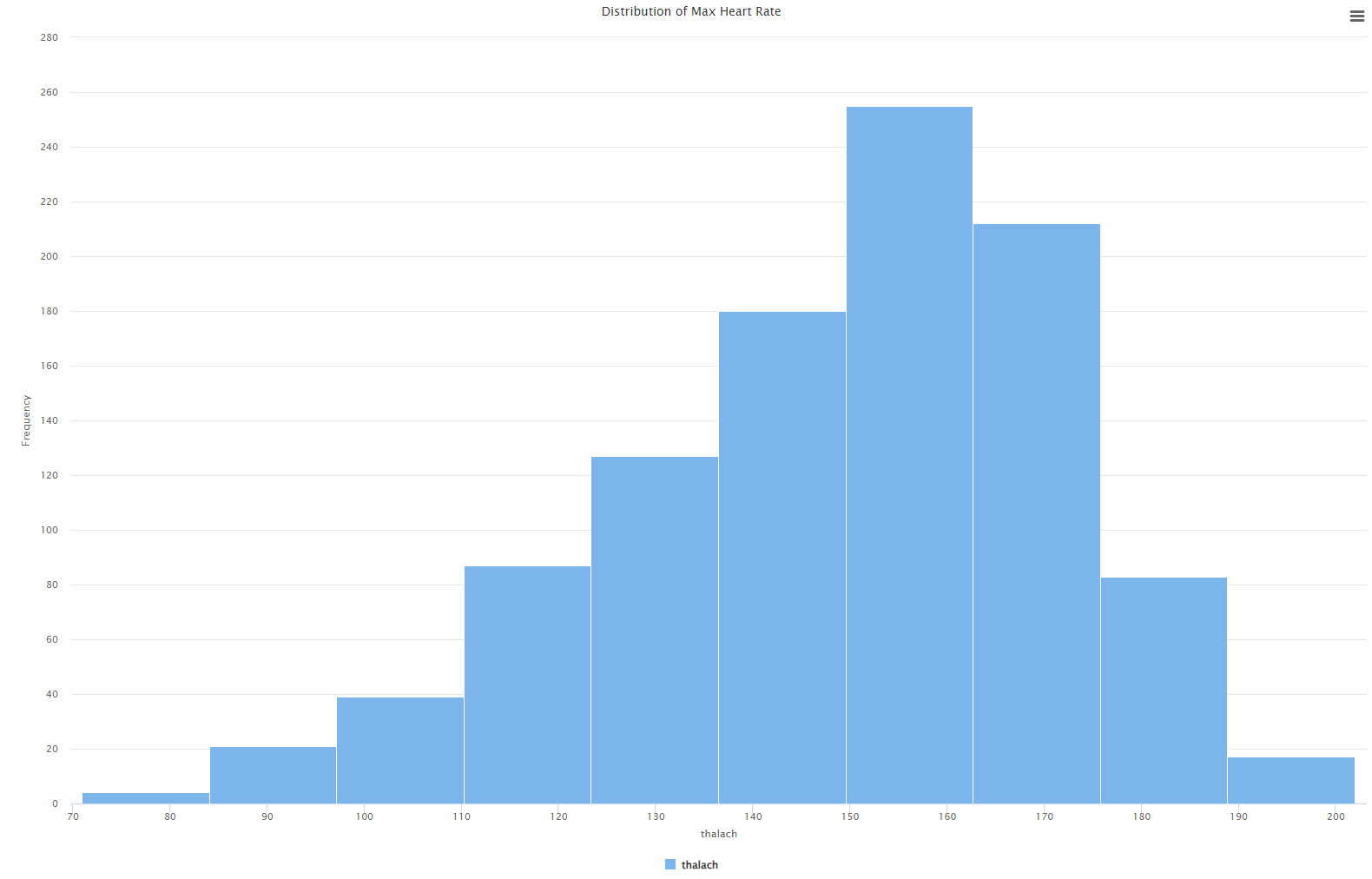


Figure : Max Heart Rate Histogram

The analysis of maximum heart rate also shows that the distribution is not normal. The histogram shows that the data is skewed to the left meaning that the majority of patients had a heart rate closer to the maximum recorded.

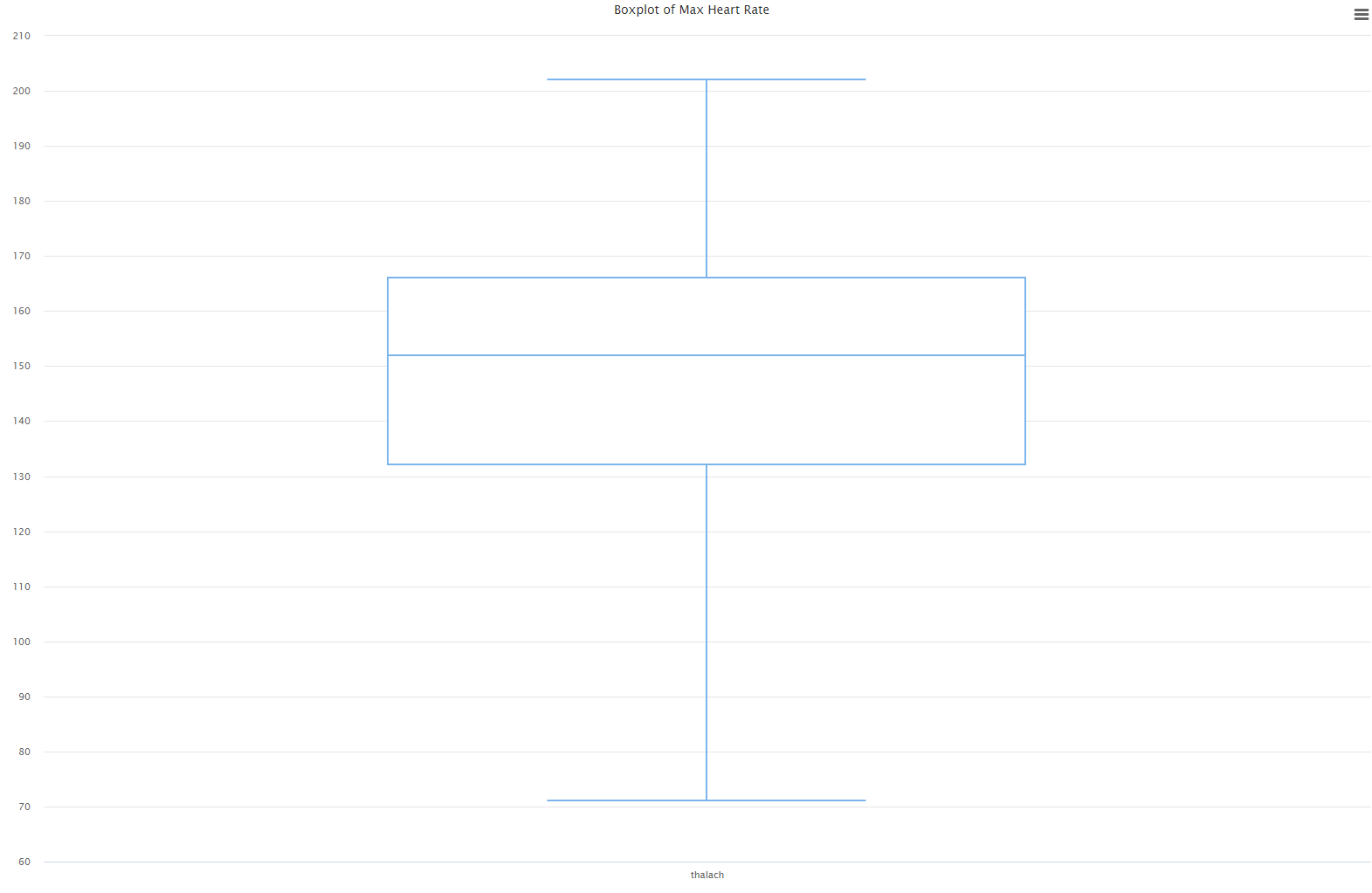


Figure : Max Heart Rate Boxplot

## **Bivariate Analysis**

The scatterplot shows the distribution of heart disease based on the cholesterol levels. The analysis shows that there are many patients with recorded heart disease that have high cholesterol levels. The higher number of green dots shows this as more individuals tested with cholesterol levels between the observed values have heart disease (Jindal et al., 2021). The correlation analysis shows weak correlations between most of the variables. This means that conducting a multivariate analysis using a scatterplot for the dataset makes no sense.

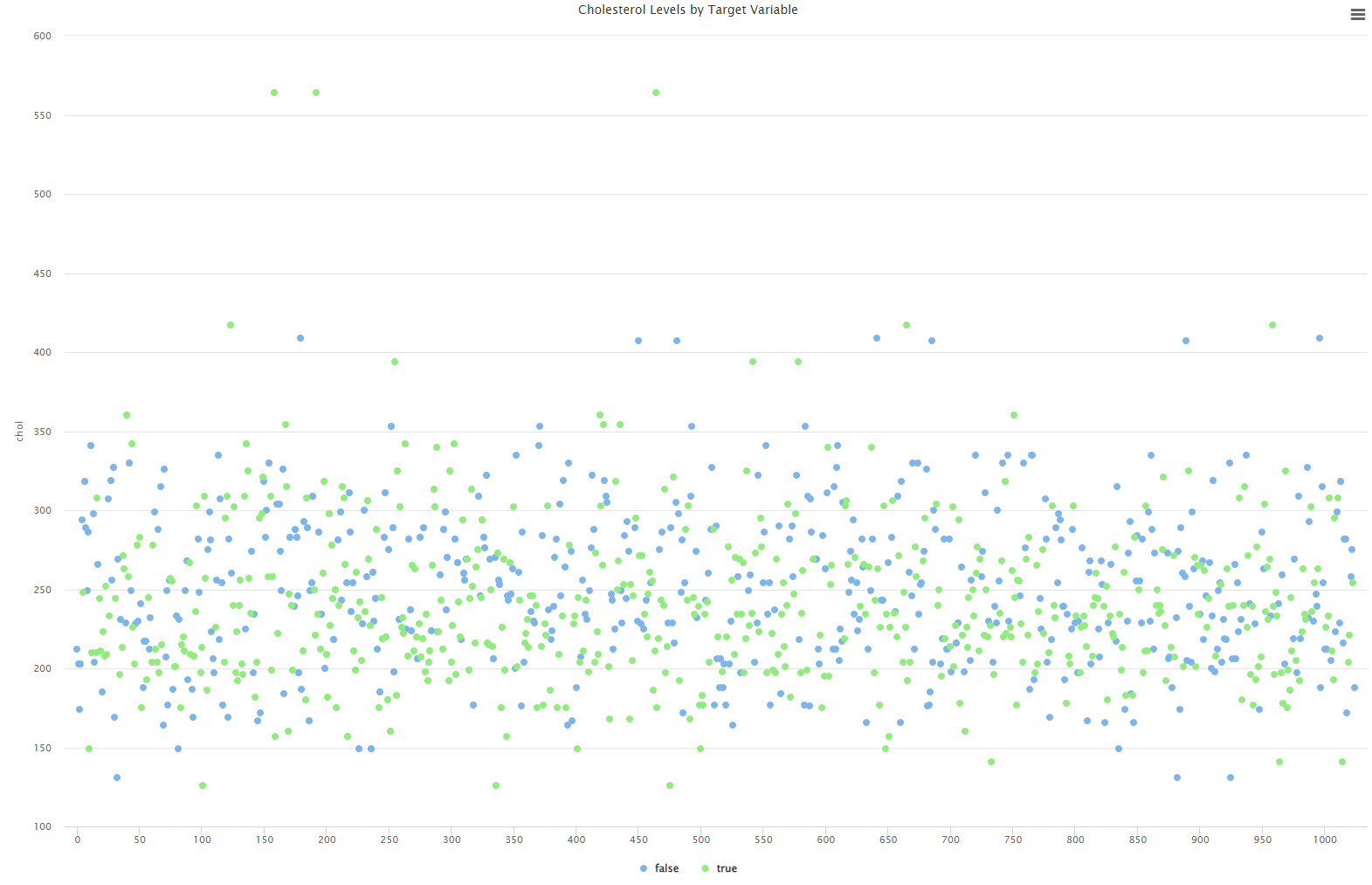


Figure : Cholesterol Levels by Heart Disease (true/false)



Figure : Correlation Heatmap

## **Outlier Detection**

The boxplot plot shows the distribution of all the numerical variables with the aim of finding outliers. The plot shows that there are no visible outliers in the dataset making it a suitable candidate for the analysis below.

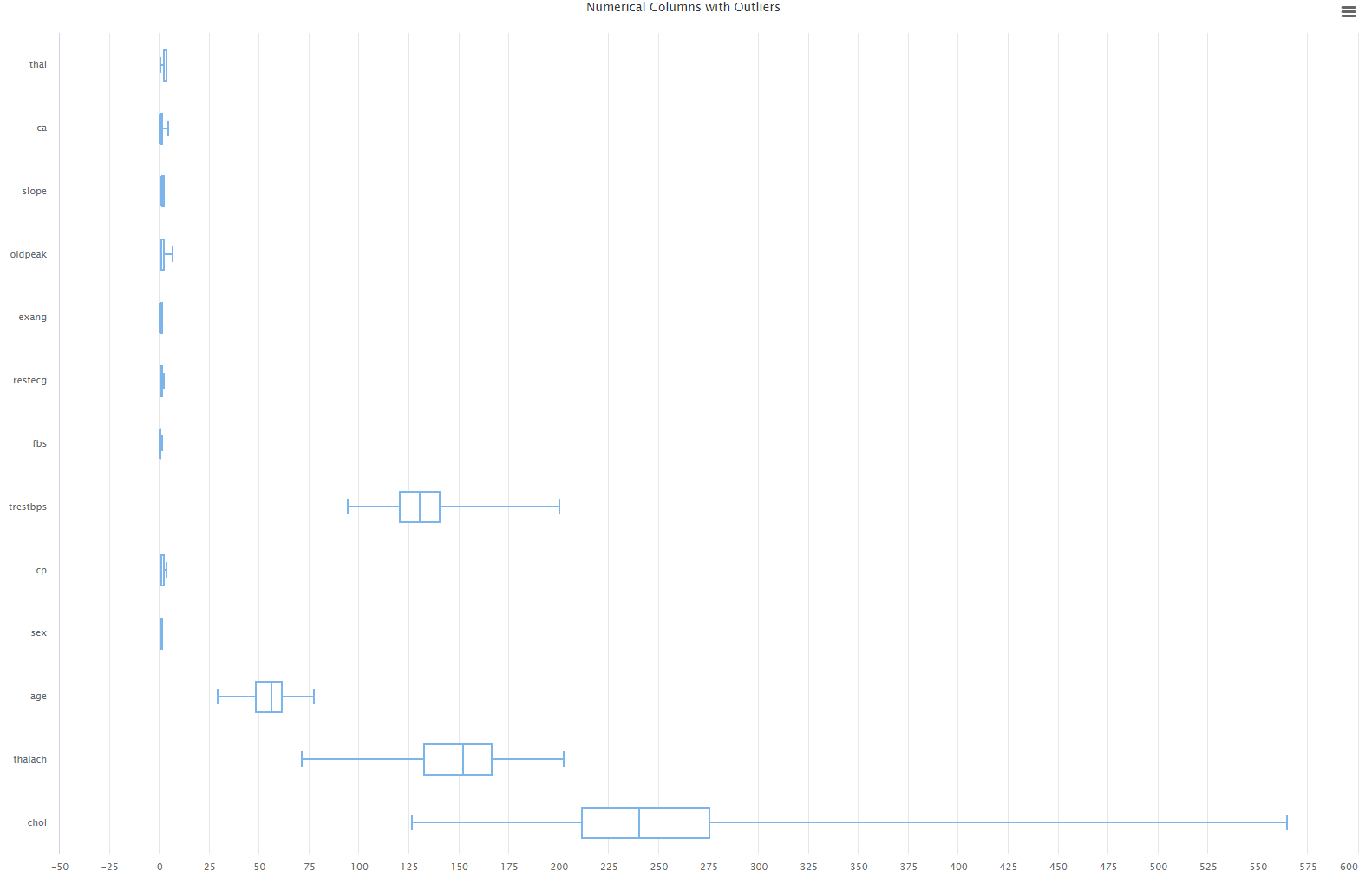


Figure : Outlier Check

# **Modeling**

## **Logistic Regression**

Logistic regression model is created in this case to help in the estimation of the probability of heart disease using the logistic function. The model is created using operators in Rapid Miner as shown in figure 11 below. The model is trained through a split set of training, testing and validation sets created. The training process involves the normalizing of numerical features in the set to improve convergence. The model is then trained using the maximum likelihood estimation method with the use of regularization to prevent overfitting. The resulting model is applied to the test and validation sets and the performance in each case is measured.

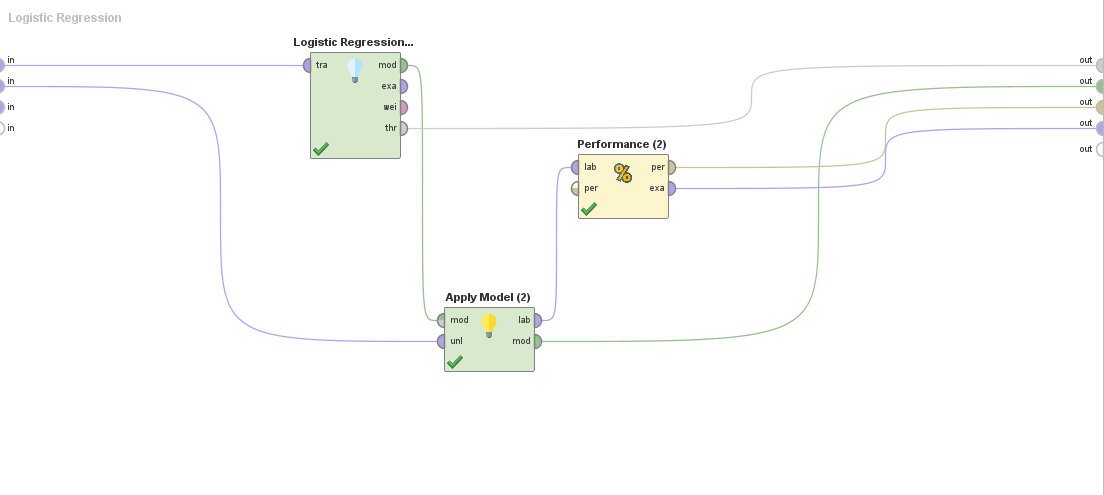


Figure : Logistic Regression Model

## **Decision Tree**

The decision tree model is also used as it provides a different approach to the prediction of heart disease. The mode works by splitting the data based on feature conditions to make predictions. The training process for this model involves selection of the best feature at each node which is done using Gini index. The data is split recursively until leaf nodes are reached which in this case is the stopping criteria being met such as maximum depth or minimum samples. The tree is further subjected to pruning operations to reduce overfitting and ensure the results are accurate. Figure 12 below shows the resulting model from the analysis carried out in Rapid Miner.

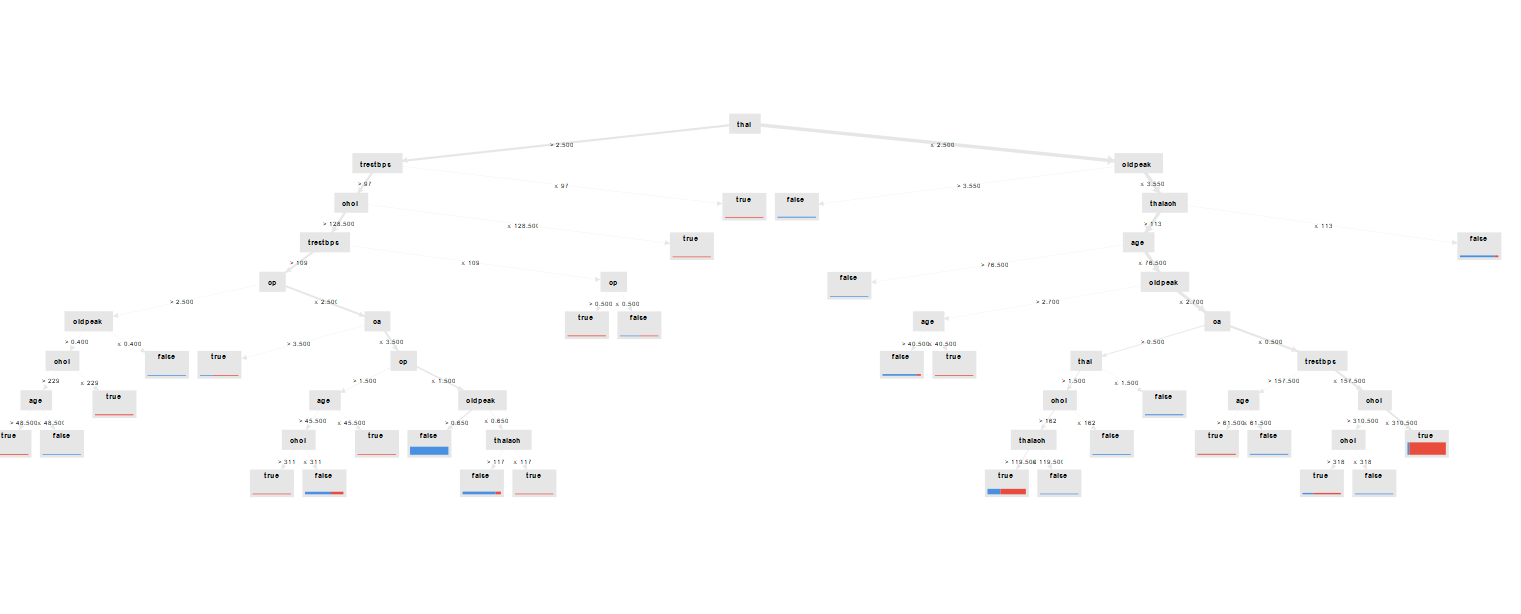


Figure : Decision Tree Model

## **Random Forest**

The random forest model is created based on the decision tree model and thus is an ensemble model. The model is created through the combination of multiple decision trees for higher accuracy and robustness. The training process involves creating multiple decision trees using the bootstrap sampling method (Sharma et al., 2020). This method often referred to as bagging allows for each tree to be trained on a random subset of features which greatly reduces variance. The aggregate predictions from all the trees are thus used as the majority vote for the classification result. Figure 13 below shows one of the many trees created as part of the random forest model.

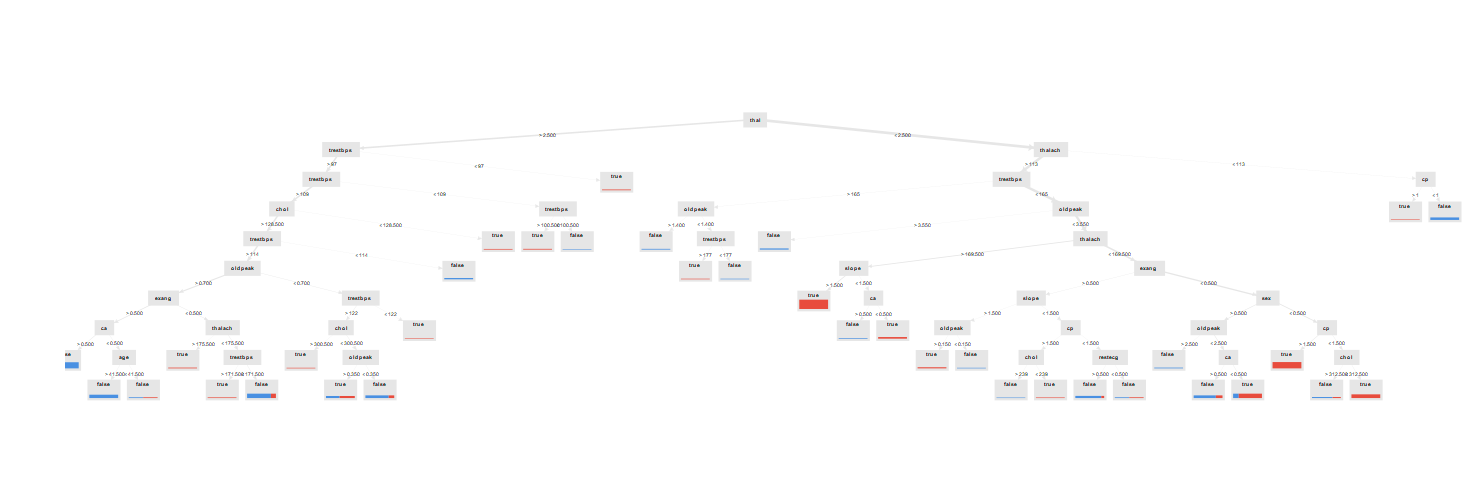


Figure : Random Forest Model

# **Performance Evaluation**

## **Accuracy Metrics**

The images below show the different accuracy metrics for each of the three models when they are applied. The results show that the random forest model outperformed the logistic and decision tree models. The model achieved a 94.53% accuracy which is quite indicating the capabilities in capturing complex patterns in the data. The decision tree model performed relatively well with an accuracy of 85.55%. This was fairly better than the 84.38% resulting from the logistic regression model. This suggests that the decision tree might have had overfitting issues affecting the outcome. The logistic regression model thus provided a solid baseline with good interpretability for the analysis.

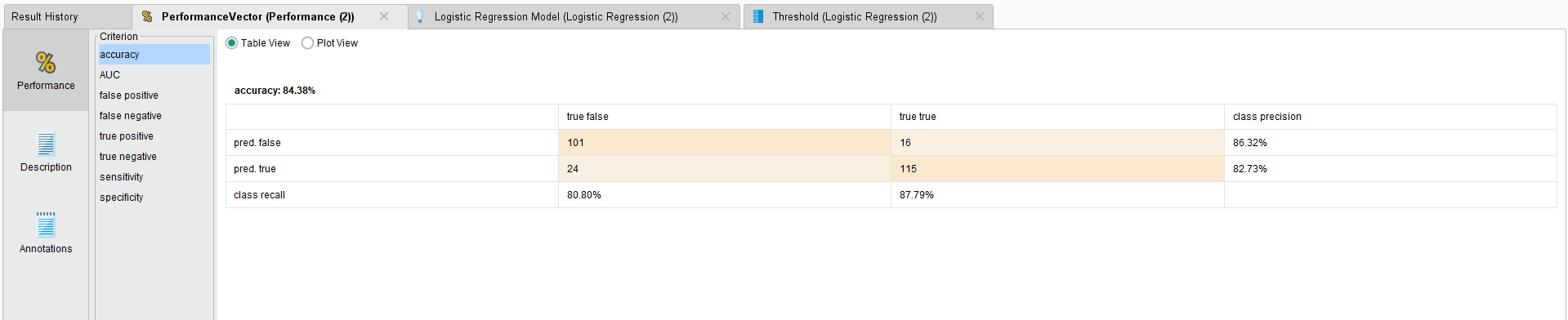


Figure : Logistic Regression Accuracy

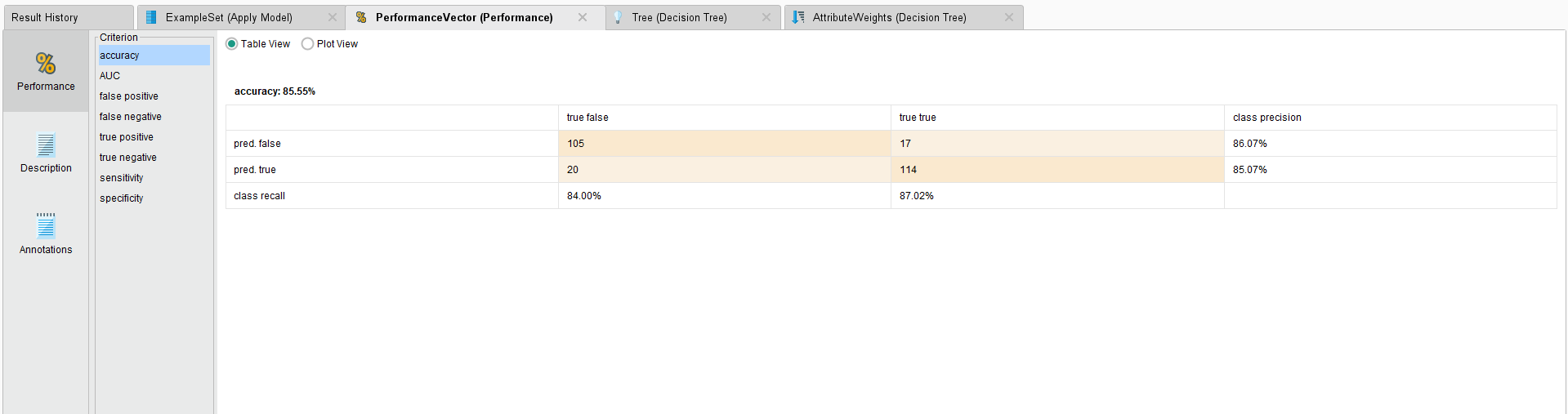


Figure : Decision Tree Accuracy

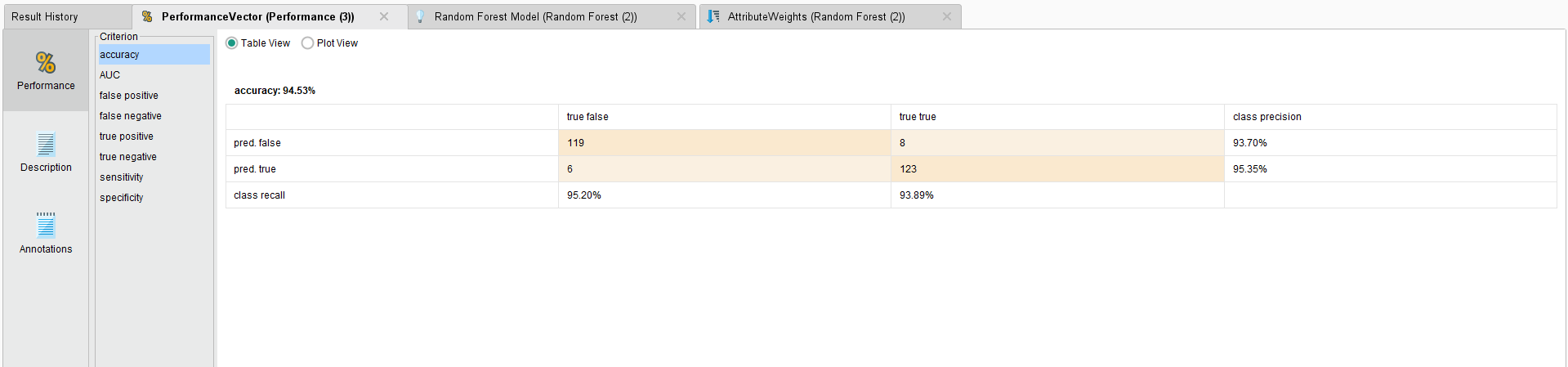


Figure : Random Forest Accuracy

## **Confusion Matrices**

The three models each resulted in a confusion matrix providing detailed breakdown of the predictions. The breakdown includes instances of correct and incorrect classifications as shown below. The measures also provide an overview through accuracy score, precision, sensitivity and specificity. These measures show that the logistic regression and decision tree models had moderate recall and precision (Bhatt et al., 2023). This indicates that the models were able to correctly classify most cases with existence of both false negatives and positives. The random forest model on the other hand exhibited extraordinary performance with fewer misclassifications. This is shown by the high recall and specificity which are over 90%. The performance indicates the existence of minimal false positives and negatives.

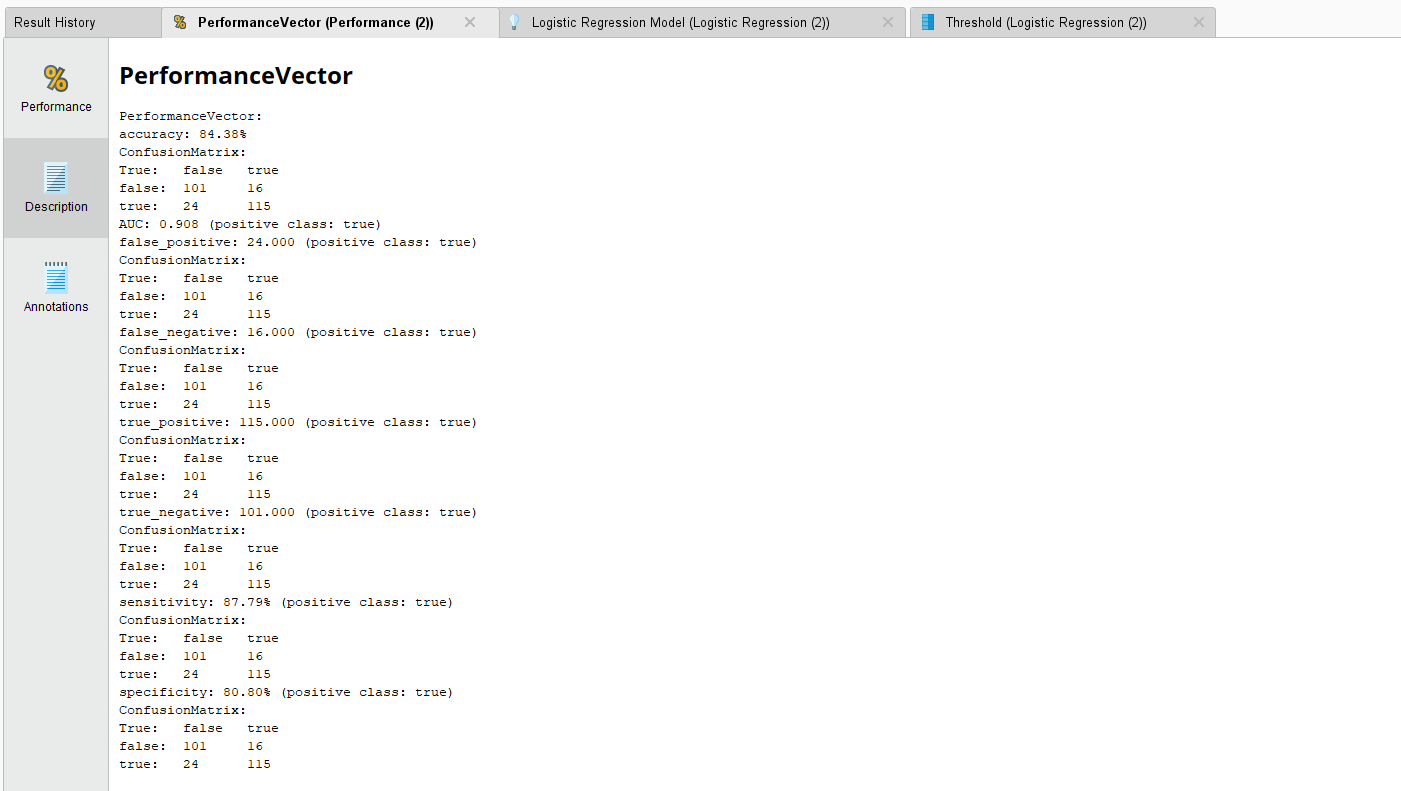


Figure : Logistic Regression Matrix

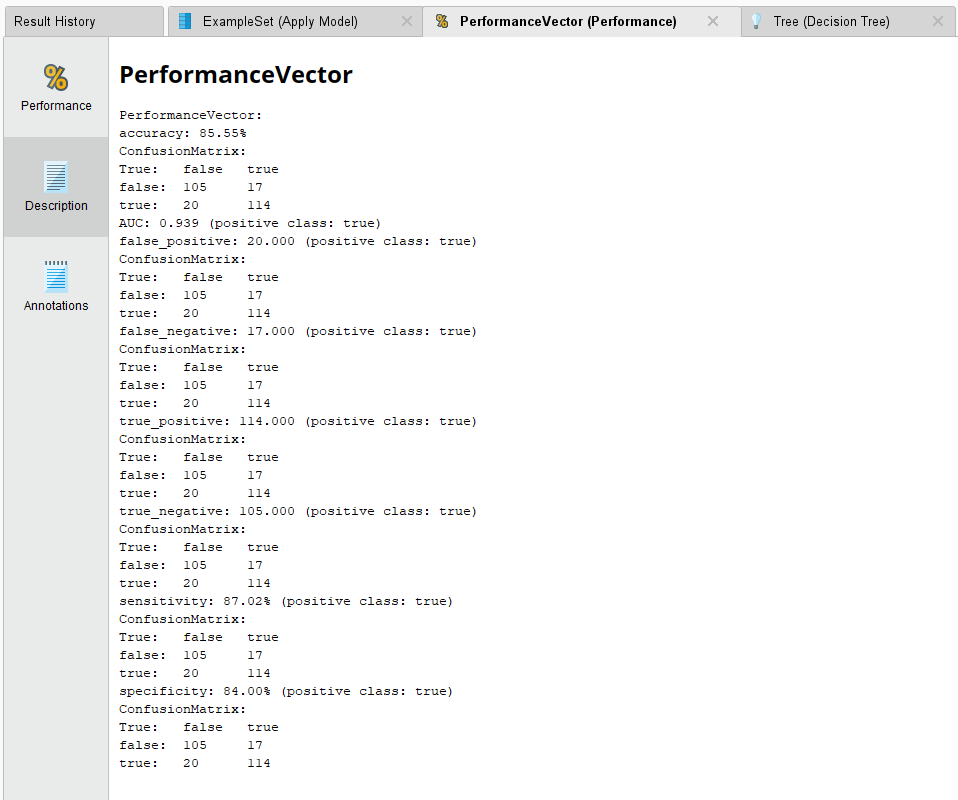


Figure : Decision Tree Matrix

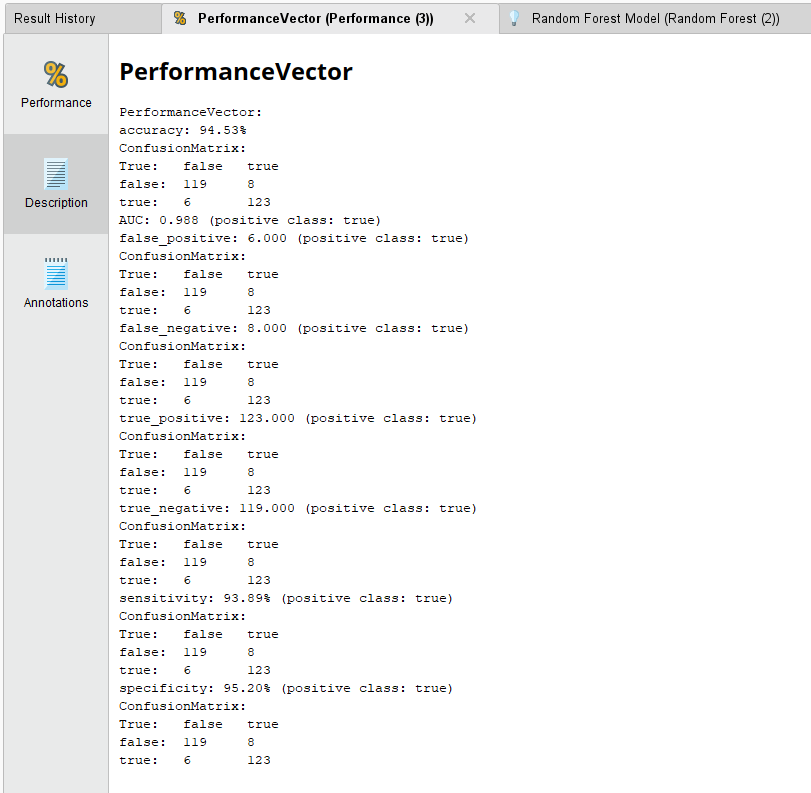


Figure : Random Forest Matrix

## **ROC and AUC Analysis**

The ROC curves for the different models show interesting results where the random forest model is a near-perfect model with a sharp rise in the curve to (0,1). This differs from the logistic and decision tree models where the rise in not sharp. Since the curves are further from the top-left corner we conclude that they have lower performance when compared to the random forest model. The AUC further proves these observations as the random forest model has an AUC of 0.988 which is closest to 1 indicating perfect classification. The logistic and decision tree models do not however fall behind as they have quite high AUCs though smaller.

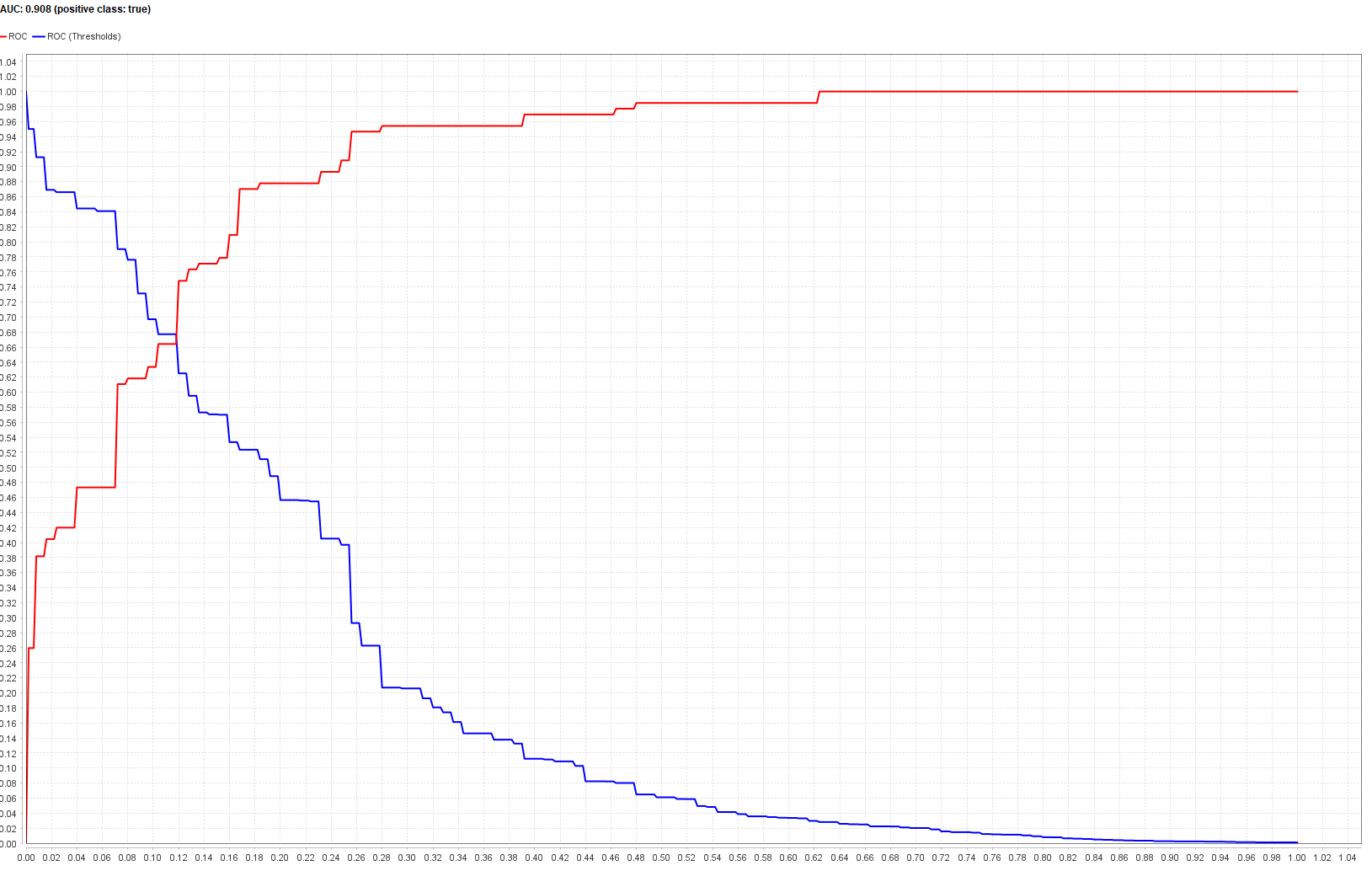


Figure : Linear Regression ROC and AUC

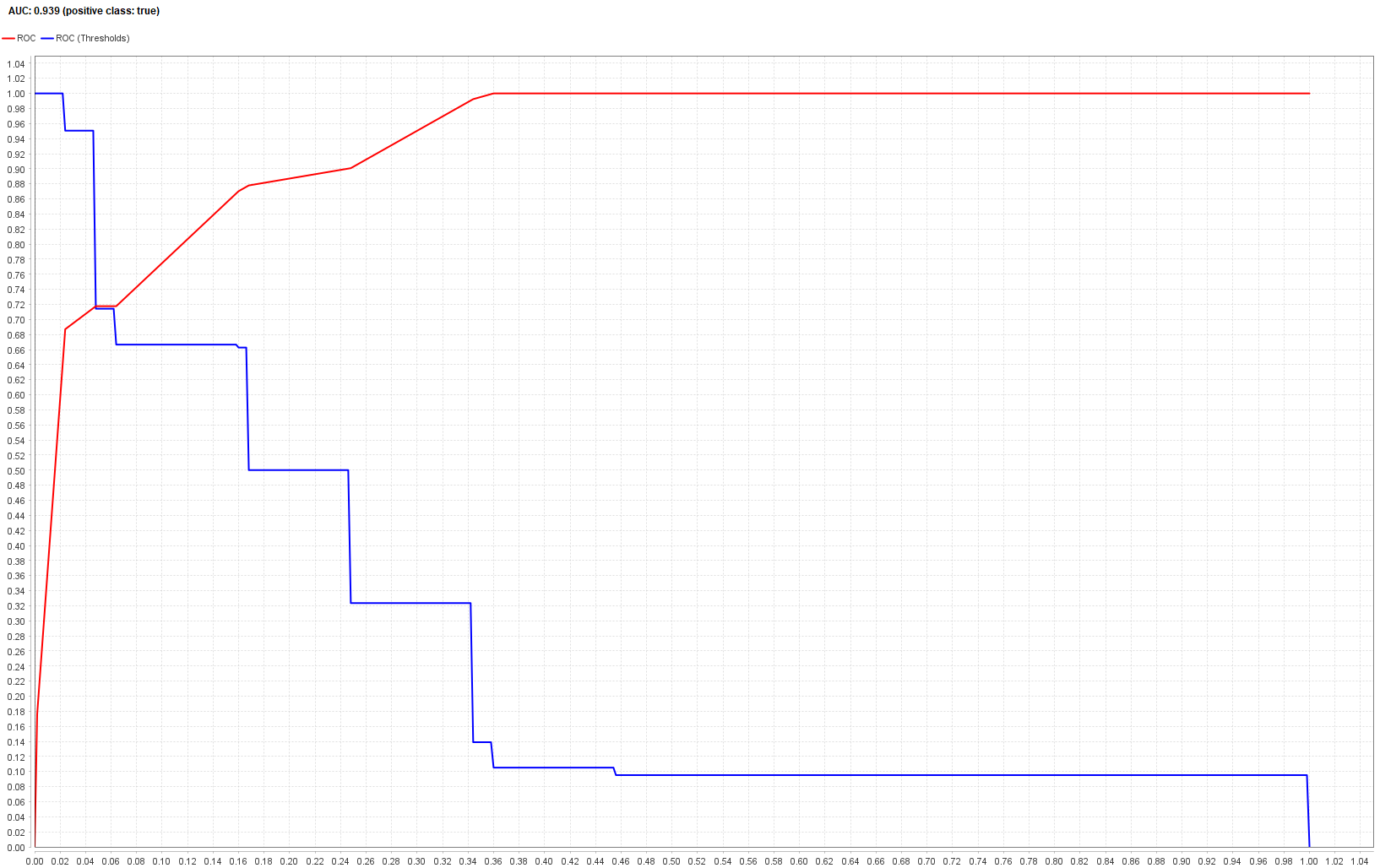


Figure : Decision Tree ROC and AUC

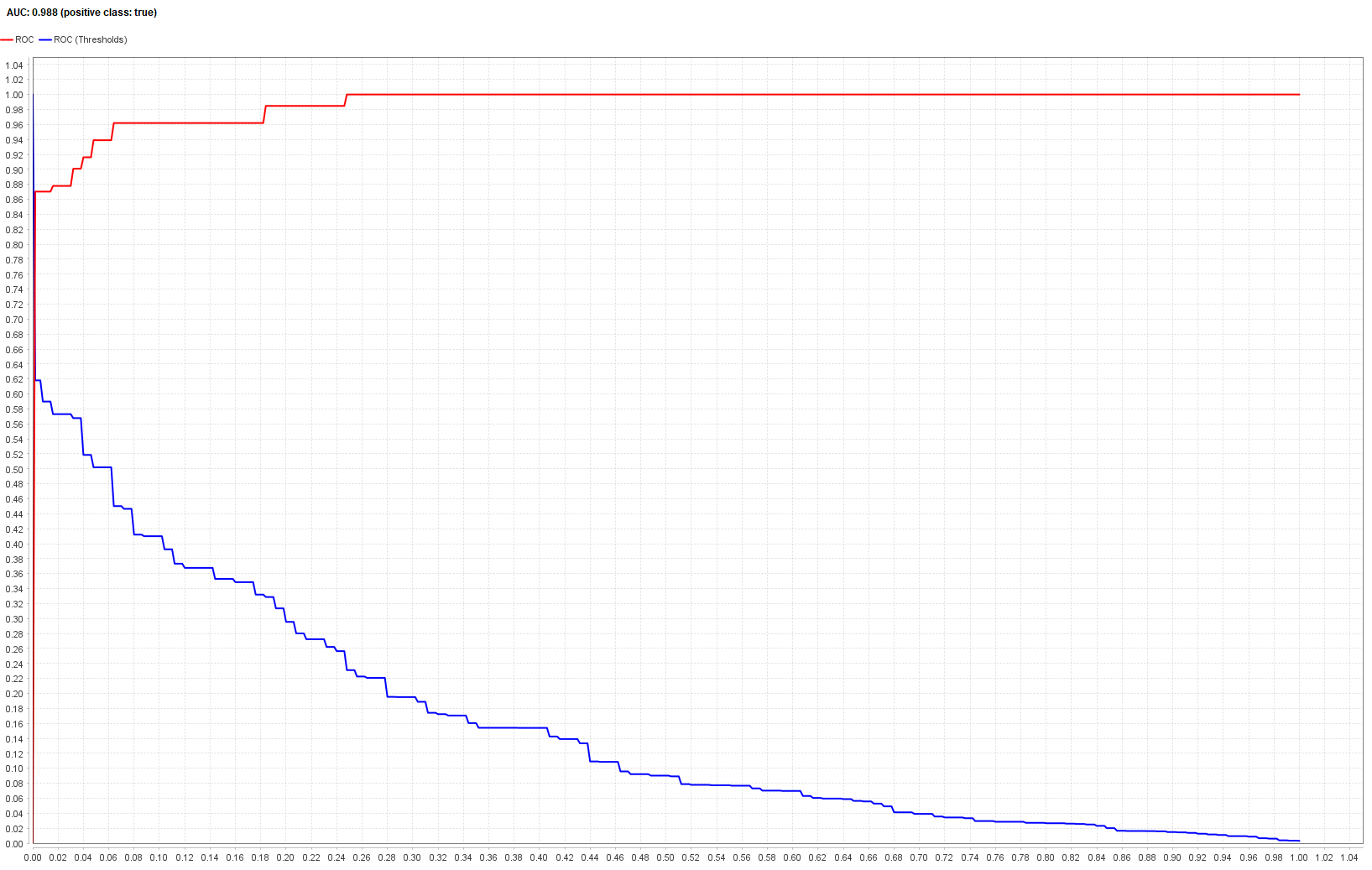


Figure : Random Forest ROC and AUC

# **Conclusion**

The research demonstrates the effectiveness of machine learning in the prediction of heart disease conditions. The models tested include logistic regression, decision tree and random forest among which random forest stood out. The model significantly outperformed the rest with an accuracy of 94.53% highlighting its capability in capturing complex patterns and reducing overfitting. The findings thus suggest ensemble methods are the most reliable in capturing medical patterns and predictions. They are essential in medical predictions as they provide a strong foundation for early diagnosis and risk assessment. This is especially useful from a business and healthcare perspective as implementing such models can enhance decision making hence reducing misdiagnosis.

# **References**

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345.

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. Algorithms, 16(2), 88.

Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In 2020 2nd international conference on advances in computing, communication control and networking (ICACCCN) (pp. 177-181). IEEE.