Cardiovascular (Heart) Disease: Study of Possible Risk Factors Caused Heart Failures

Janith Perera

Introduction

Purpose of the Project

Predict a possible cardiovascular (heart) disease from risk factors which can be caused heart failures. In this project, prediction accuracy of Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Stochastic Gradient Descent, XGBoost algorithms will compare. Target variable of the dataset is a binary variable named as HeartDisease and contains 11 features such as Age, Sex, ChestPainType, Cholesterol, RestingECG and etc.

Significance of the Project

Heart disease are the number 1 reason of death worldwide. 4 out of 5 heart disease deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. People with heart disease or who are at high risk need early detection and management wherein a machine learning model can be a great help.

Research Question

Are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Stochastic Gradient Descent, XGBoost algorithms good models for predicting the possible heart disease based on clinical risk factors like chest pain type, cholesterol level, resting ECG, fasting blood sugar, maximum heart rate and etc.

Description of Dataset

Dataset downloaded from www.kaggle.com which contains one target variable and 11 input variables (5 categorical and 6 numerical variables) with 918 instances and no null-values.

Data Preprocessing

Data Preparation

All categorical features, Sex, ChestPainType, RestingECG, ExerciseAngina and ST_Slope, one-hot-encoded using pandas get_dummies() method. Numerical features, Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak, scaled using sklearn StandardScaler() method.

Exploratory Data Analysis

It is important to note that some of special features in the dataset while scrubbing the data. It seems to be men are more likely to have a heart disease and age 50 or older adults having more risk to have heart disease than younger persons. Further, it can clearly see Asymptomatic chest pain is more common among the patients with heart disease and there is no significant difference in chest pain types among non-heart disease patients. If ST segment is flat for any patient, they are more likely to having heart disease.

Visualization

Pair plots are not showing any valuable information but can visualize data points of heart disease and no heart disease are overlapping always. Box plot shows Cholesterol feature has some outliers and heatmap shows there is no any correlation in between numerical features which is a good trend in data and therefore it can use all features to further analysis.

Data Splitting

There is no requirement to balance the dataset since target variable is balanced while having 55% heart disease and 45% no heart disease samples. For the modeling purpose whole dataset split in to 70:30 ratio of train and test sets by using train_test_split() method.

Model Building and Evaluation

Model Building

In this project, primarily used the classification models to model the dataset because the target variable is categorical. Initially used six of well-known classification models in sklearn package. Those are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD) Classifier and XG Boost Classifier. Models compared with the overall accuracy, precision score, recall score, f1 score, ROC-AUC score and log-loss score.

Model Optimization and Model Selection

Three best fit models selected out of six models and remodeled with cross validation score and selected the best model out of three. Support vector machine classification model is the best performer among all models and tuned with hyperparameters to select optimal model with tuned parameters.

Model Comparison

At the initial step Decision Tree and Random Forest classifiers overfitted and SGD classifier had the poor performance than Logistic regression, SVC and XG boost classifier. Among those selected three models, SVC got the highest cross validation score and finally hyperparameter tuned SVC model gave 90% and 88% overall accuracy on training set and testing set respectively.

Conclusion

Conclusion

Support vector machine classifier worked best on this dataset and this model can be used to predict the patients who are having heart disease or not by analyzing risk factors. Cholesterol, Fasting blood sugar, Oldpeak (numeric value measured in depression), ST segment flat and ST segment upsloping are the most important features of predicting heart disease.

Lessons Learned

Age and two types of chest pain has negative values for permutation_importance() module which is in sklearn indicates that the predictions on the shuffled (or noisy) data are more accurate than real data. This means that those features do not contribute much to predictions, but random chance caused the predictions on shuffled data to be more accurate.

Recommendations

Since this dataset biased on gender, that means dataset has more instances of men than women, I recommend to populate this dataset with more female patients and remodel to check whether there is any difference on accuracy of the predictions.

Appendix

Figure 1: *Heart disease by gender*

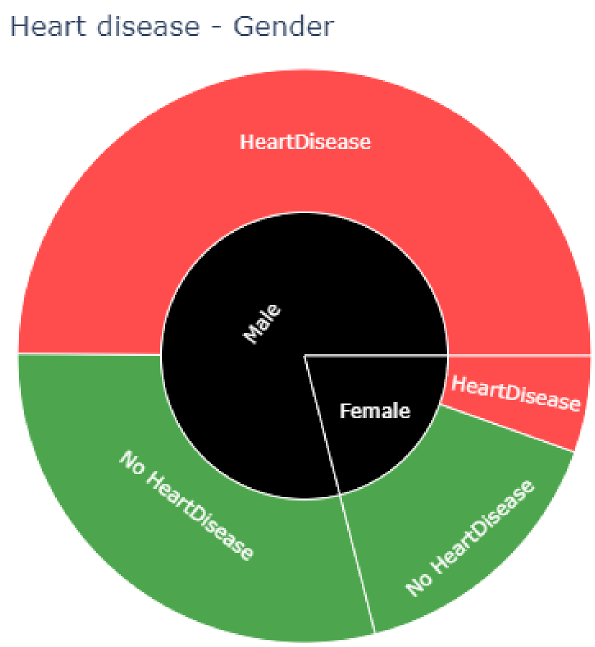

Figure 2: *Heart disease by age groups*

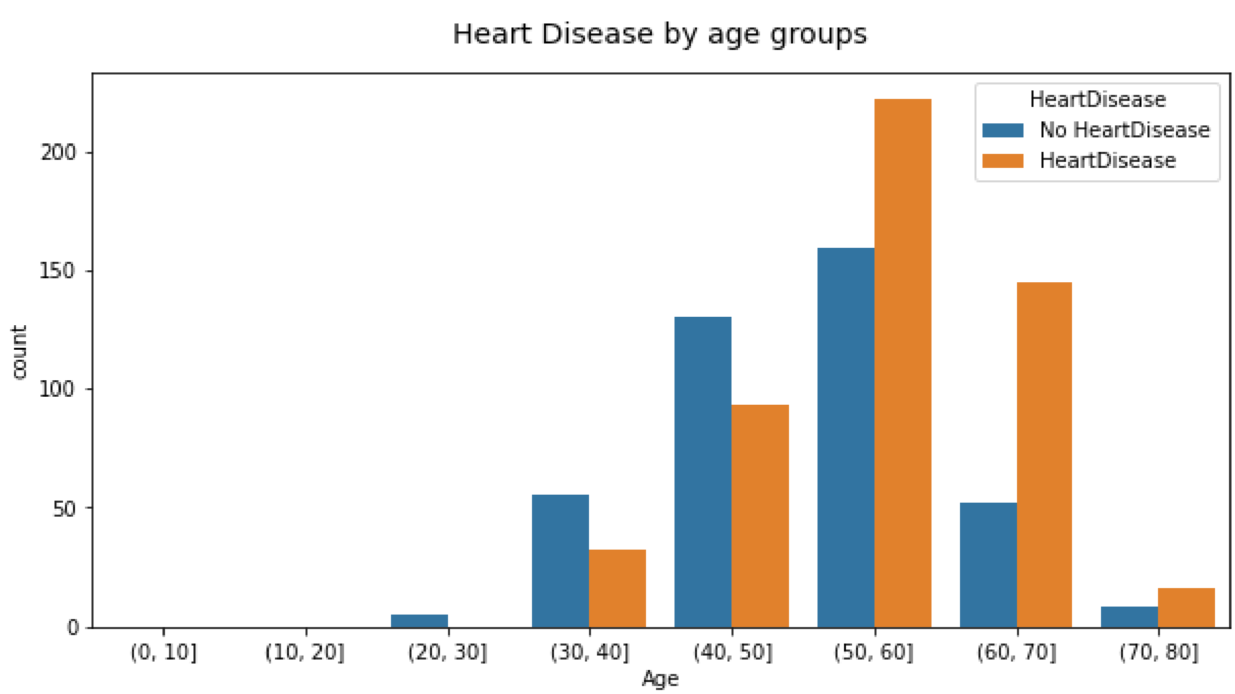

Figure 3: *Heart disease by chest pain types*

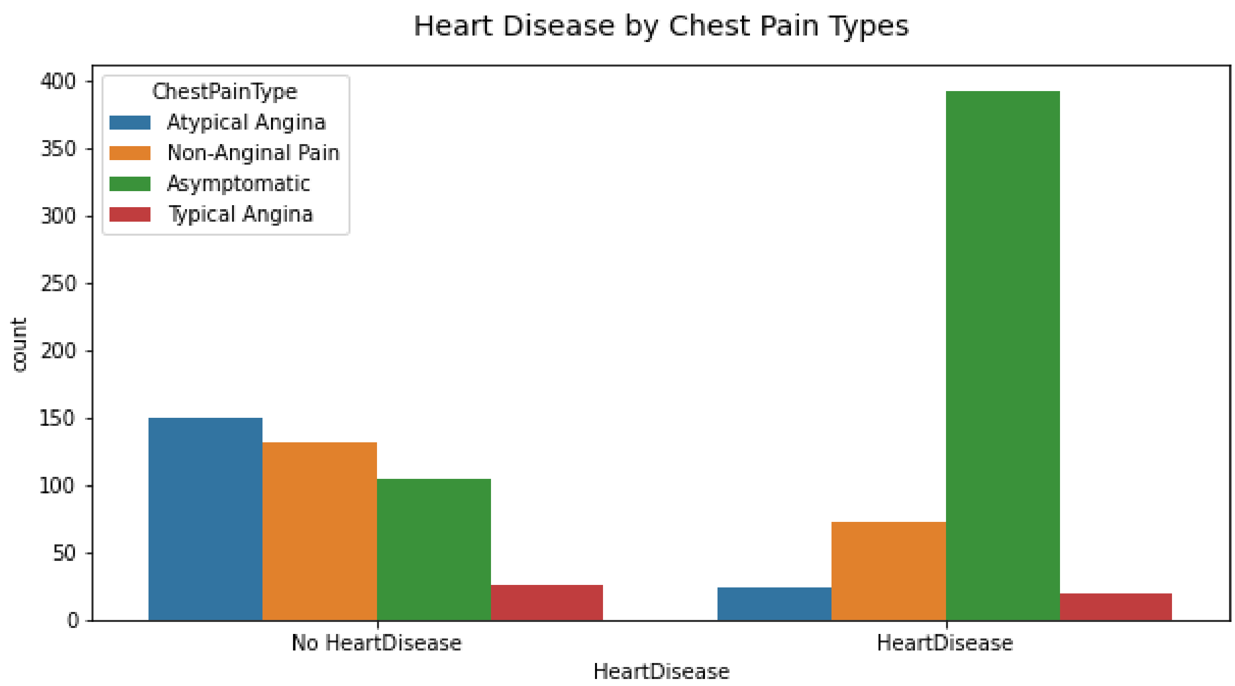

Figure 4: *Heart disease by resting ECG*

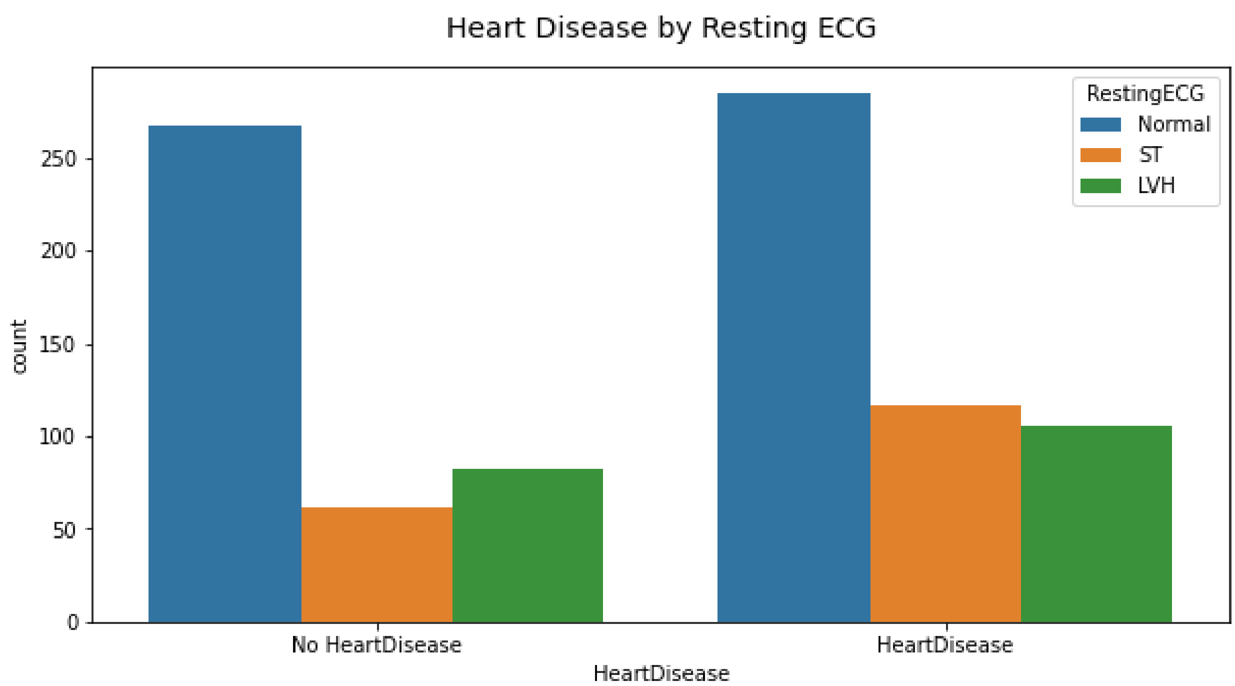

Figure 5: *Heart disease by the slope of the peak exercise ST segment*

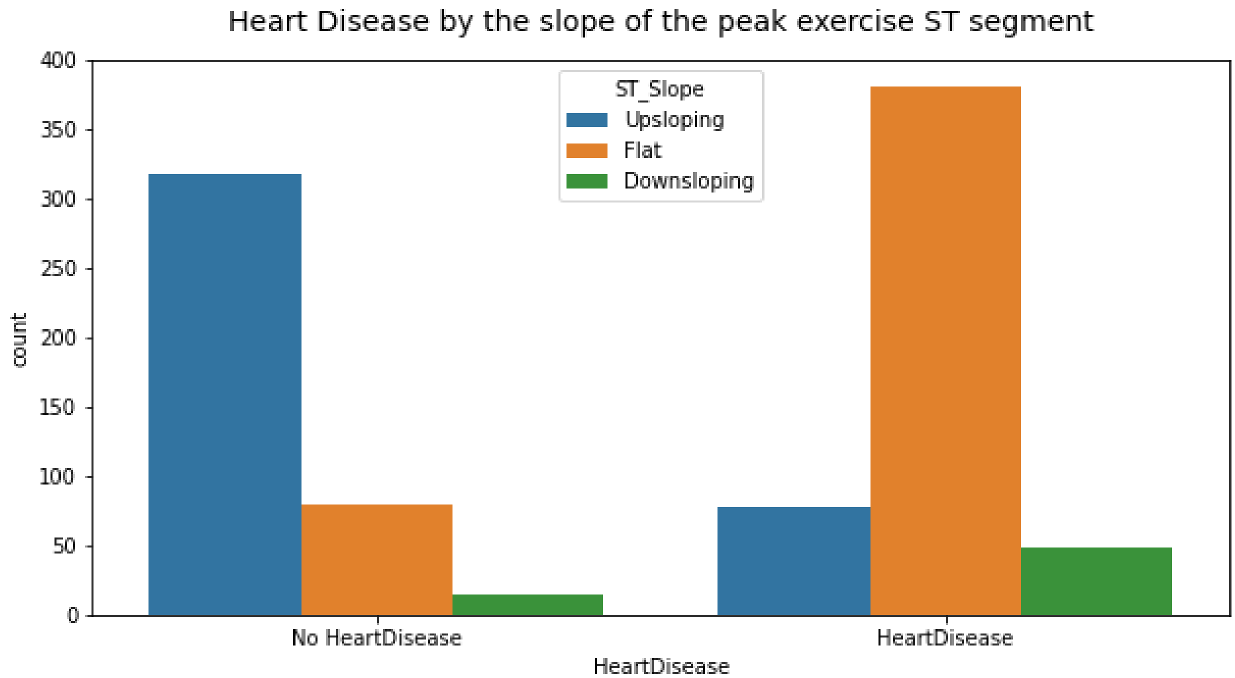

Figure 6: *Heart disease by exercise-induced angina*

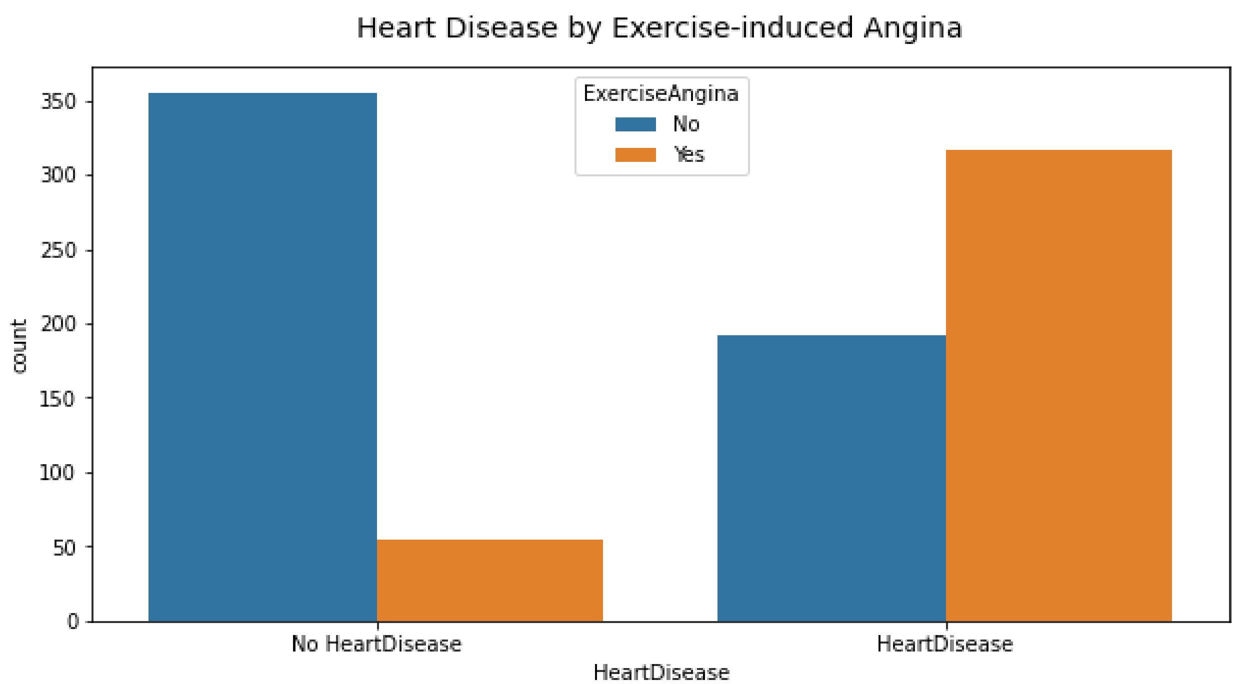

Figure 7: *Pair-plots on numerical features*

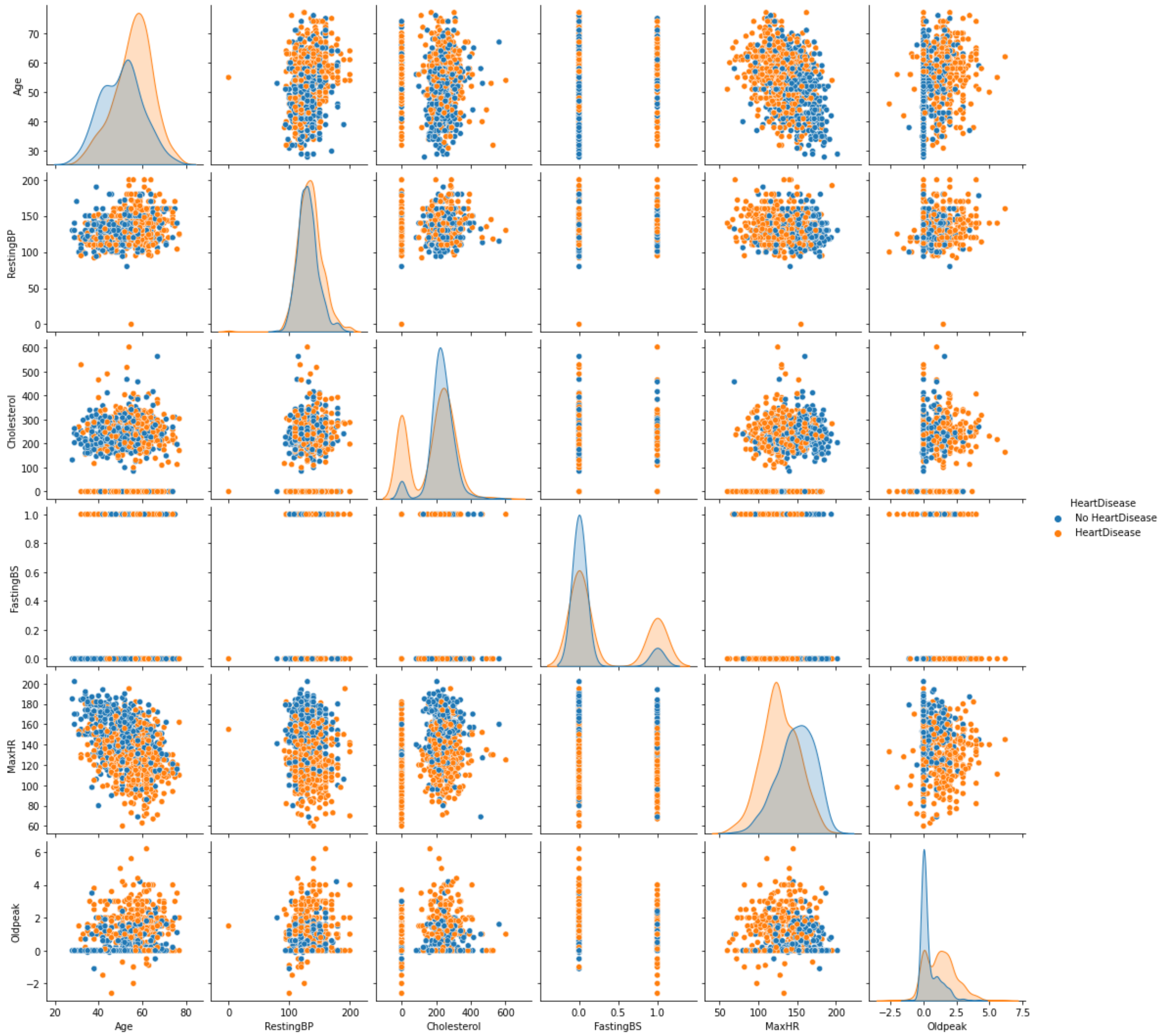

Figure 8: *Box plot on numerical features*

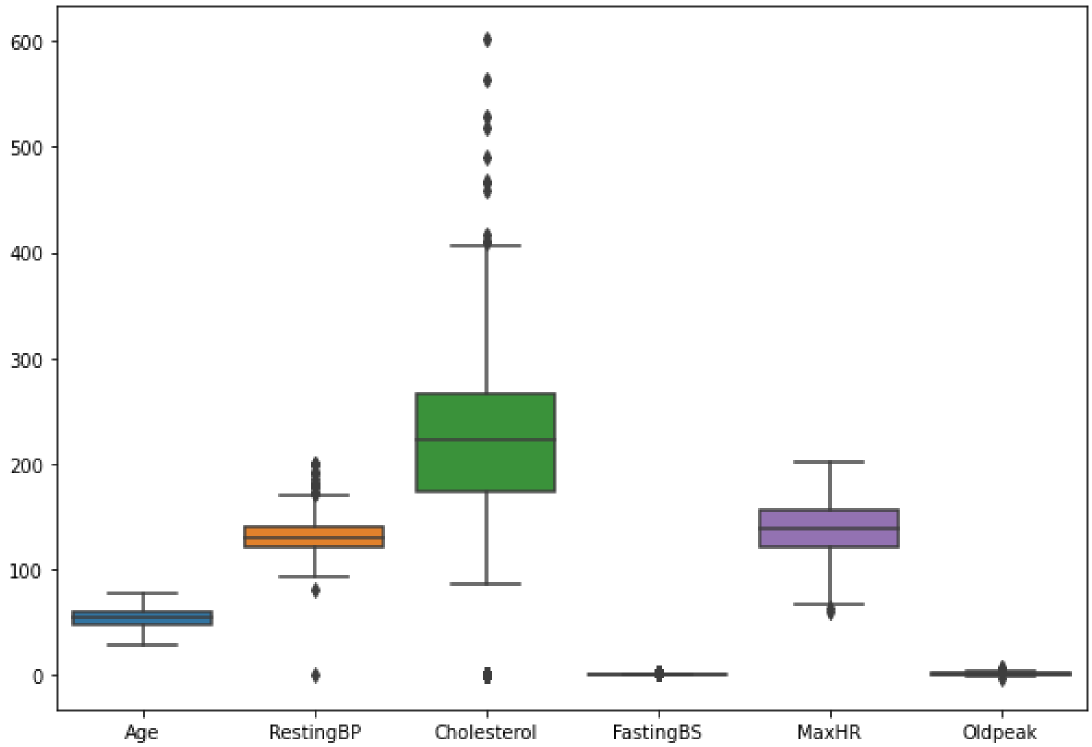

Figure 9: *Correlation heatmap on numerical features*

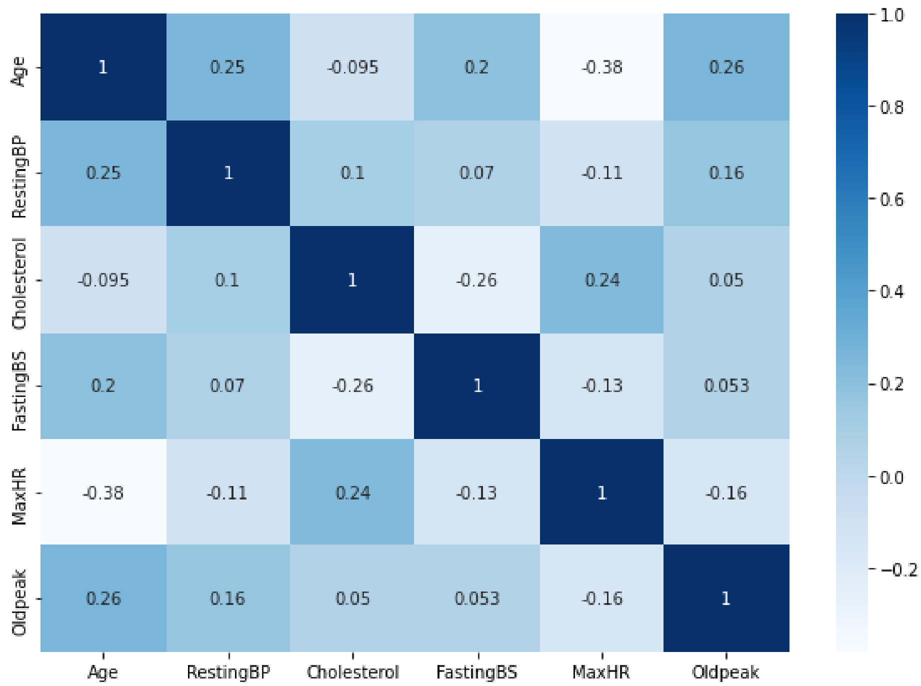

Table 1: *Comparison of accuracy scores*

Model	Set	Accuracy Score	Precision Score	Recall Score	F1 Score	ROC-AUC Score	Log-loss Score
LogisticRegression	Training	0.86	0.86	0.89	0.87	0.86	4.79
	Test	0.88	0.92	0.88	0.90	0.89	4.00
DecisionTreeClassifier	Training	1.00	1.00	1.00	1.00	1.00	0.00
	Test	0.78	0.88	0.73	0.79	0.79	7.76
RandomForestClassifier	Training	1.00	1.00	1.00	1.00	1.00	0.00
	Test	0.85	0.90	0.84	0.87	0.85	5.26
SVC	Training	0.90	0.89	0.93	0.91	0.90	3.44
	Test	0.88	0.90	0.91	0.90	0.88	4.00
SGDClassifier	Training	0.78	0.86	0.71	0.77	0.78	7.64
	Test	0.75	0.91	0.65	0.76	0.78	8.51
GradientBoostingClassifier	Training	0.94	0.93	0.96	0.95	0.94	1.99
	Test	0.88	0.93	0.86	0.89	0.88	4.25

Table 2: *Comparison of cross validation scores*

Model	CV Score
LogisticRegression	0.85
SVC	0.87
GradientBoostingClassifier	0.86

Table 3: *Classification report on training set with hyperparameter tuned model*

	precision	recall	f1-score	support
0	0.92	0.86	0.89	298
1	0.89	0.93	0.91	344
accuracy			0.90	642
macro avg	0.90	0.90	0.90	642
weighted avg	0.90	0.90	0.90	642

Table 4: *Classification report on testing set with hyperparameter tuned model*

	precision	recall	f1-score	support
0	0.86	0.85	0.86	112
1	0.90	0.91	0.90	164
accuracy			0.88	276
macro avg	0.88	0.88	0.88	276
weighted avg	0.88	0.88	0.88	276

Figure 10: *Heatmap on confusion matrix*

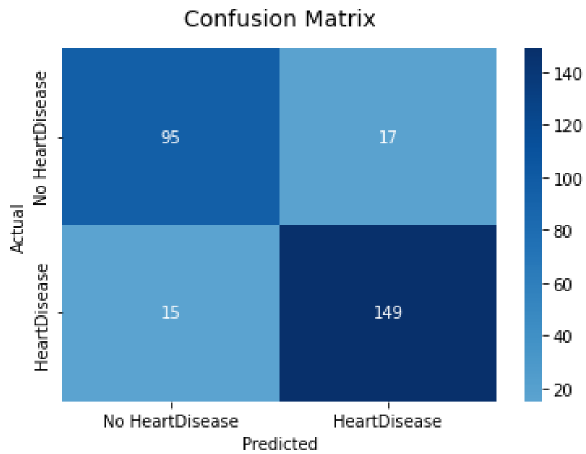

Figure 11: *Feature importance*