Heart Disease Prediction

Sandun Gunasekara

Assignment 1

1. **Steps followed.**

At the beginning, 'processed.cleveland.data' was loaded from the provided resource dataset and some data processing was performed to clean the data. Firstly, all values containing non-numeric values were replaced with 'NaN', as there were some missing values. Next, all rows containing 'NaN' values were removed from the dataset, as they were considered garbage values.

As part of the feature selection process, a feature correlation matrix (Figure 1) was drawn and analyzed to determine whether there were highly correlated features to remove. A threshold of 0.7 was used for the correlation, but no such features were found in this dataset.

One hot encoding was then performed to process the categorical features ('sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'). One hot encoding allows categorical data to be represented numerically in a way that can be used as input for machine learning algorithms.

After all these steps, all the features were scaled using 'StandardScaler' for data normalization. This is a common technique used in machine learning to standardize the range of features, make them more comparable, and improve the performance of machine learning algorithms. Data normalization can also be used to detect outliers.

Once the data processing was completed, the next step was to create training and testing datasets. In this scenario, a split threshold of 0.25 was used. The dataset was then split, with 75% used as training data and the remaining 25% used as testing data.

Then, the model was trained using the created datasets. A RandomForestClassifier was used instead of logistic regression methods, as random forest provides greater accuracy than logistic regression.

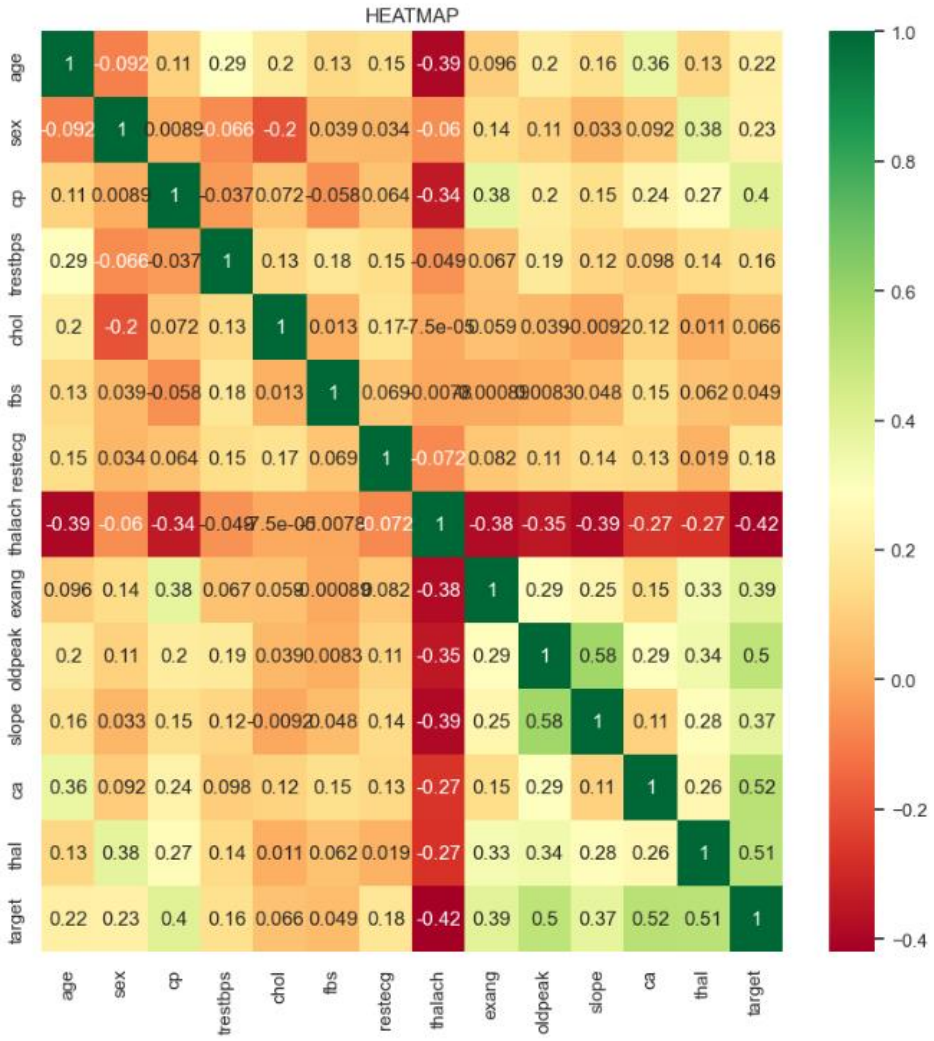

Figure 1: Correlation Matrix

Additionally, cross-validation methods were also used during model training since the dataset was limited. Cross-validation is a technique used in machine learning to evaluate the performance of a model on a limited dataset. The basic idea behind cross-validation is to split the dataset into two or more parts: one part is used to train the model, and the other part is used to test the model's performance.

Finally, predictions were made using the test dataset, and the accuracy of the model was compared with the target test values. The best accuracy achieved so far was 56%. However, the approach could be further optimized using various machine learning techniques.