Altrium Machine Learning Bootcamp - Assignment 1

REPORT

Dasun Theekshana

1. Dataset

Heart Disease Data Set – UCI Machine Learning Repository

No. of Records – 302

No. of variables – 14

- Used the processed Cleveland Dataset

2. Exploratory Data Analysis

- Used Matplotlib, seaborn and Plotly libraries
- Categorical Data Distributions visualized by barcharts
- Numerical Data Distributions visualized by barcharts and box & whisker plots.
- Feature-target relationships of categorical variables visualized by barcharts.
- Feature-target relationships of categorical variables visualized by barcharts box & whisker plots.

3. Feature Selections

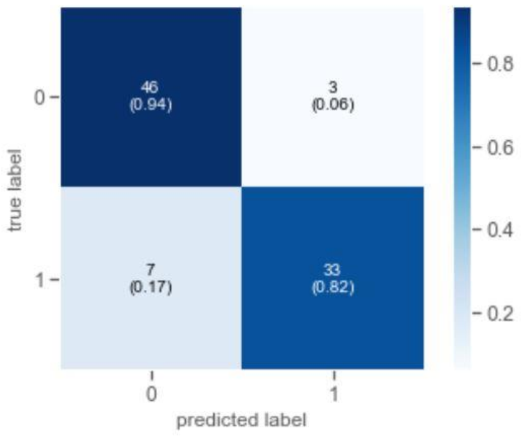
- Correlations of variables were analysed using a heatmap
- Correlations of feature variables with target variables was analyzed using bar chart
  - *Thalassemia* had a strong positivd correlation with the target variable.
  - *max_heart_rate_achieved* had a strong negative correlation with the target variable.

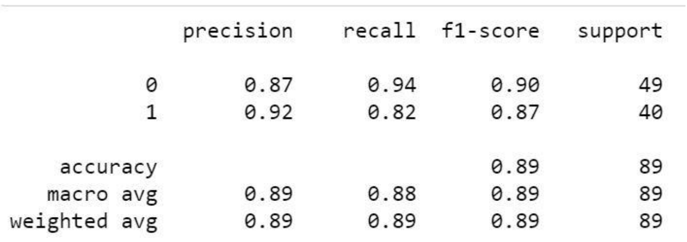
  - *cholesterol, fasting_blood_sugar & resting_blood_pressure* had the weakest correlations with target variable
- Variable Inflation factor (VIF) test was carried out to determine the multicollinearity of variables. All variables had VIF scores of <5 therefore no significant multicollinearities existed in feature variables. As a result univariate feature selection methods were considered.
- ANOVA feature selection was carried out with selectKBest algorithm to select the 10 best features
  - ***cholesterol, fasting_blood_sugar & resting_blood_pressure*** were eliminated during feature selection.
- Ultimately, ***cholesterol, fasting_blood_sugar & resting_blood_pressure*** *were eliminated from feature variable set.*

4. Model Training

- The data was split with a 70:30 split between train and test sets.
- StandardScaler was used for data normalization.
- The data was trained using the Logistic Regression Algorithm.
- Logistic regression is a statistical model used to predict binary outcomes or perform binary classification tasks. Since the target variable expresses whether a person has heart disease or not, logistic regression can be employed in this particular use case.
- Hyperparameter tuning was performed using the Grid search Cross validation technique to find the best estimators.
- The performance of the model is as follows.

Training set score: 0.8309
Test set score: 0.8876





```
              precision    recall  f1-score   support

           0       0.87      0.94      0.90        49
           1       0.92      0.82      0.87        40

    accuracy                           0.89        89
   macro avg       0.89      0.88      0.89        89
weighted avg       0.89      0.89      0.89        89
```