**HEART DISEASE PREDICTION**

**FINAL REPORT**

**ABSTRACT:**

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and relies on factors such as the physical examination, symptoms and signs of the patient. Factors that influence heart disease are body cholesterol levels, smoking habit and obesity, family history of illnesses, blood pressure, and work environment. Machine learning algorithms play an essential and precise role in the prediction of heart disease.

Heart disease can be predicted based on various symptoms such as age, gender, heart rate, etc. and reduces the death rate of heart patients.

This problem can be treated as a binary classification based problem on whether a person has a heart disease or not. Hence, Classification based Machine Learning Algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbours, Decision Tree, Support Vector Machines, etc can be used to solve this problem and the algorithm with the best accuracy can be chosen as the final solution based on different evaluation metrics.

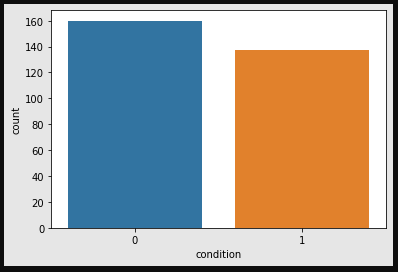
**Evaluation metrics:**

The Evaluation metrics, used to find out how good a model is, are given below,

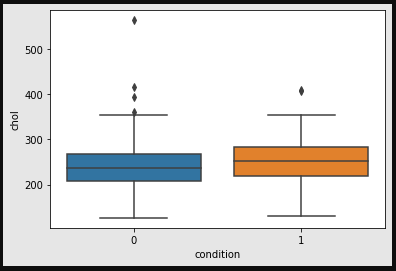
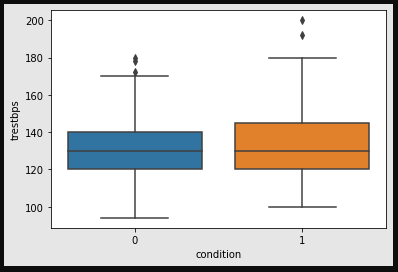
1. Classification report: The classification report function builds a text report showing the main classification metrics such as precision, recall, f1-score, support.
2. Confusion matrix: A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It shows the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN).
3. Accuracy score: Accuracy is one metric for evaluating classification models. It is defined as the fraction of predictions our model got right.

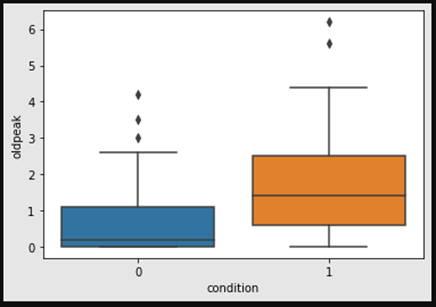
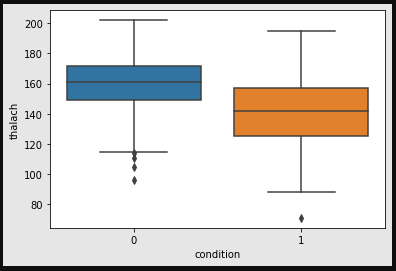
**Exploratory Data Analysis:**

* The dataset contains no null values.
* The dataset balanced. The ratio of number of people that have heart disease to the number of people that don’t have heart disease is 137:160

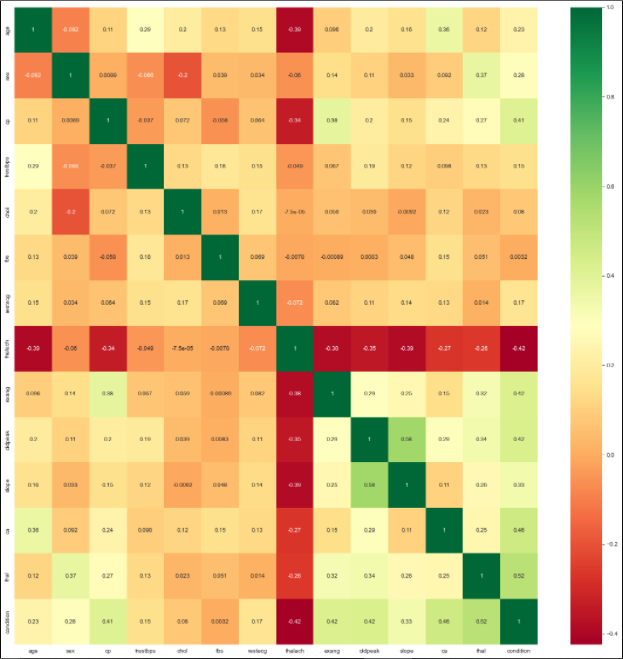


* There are a small number of outliers present in some features in the dataset.





* The feature ‘THALCH’ is negatively co-related with every other feature. by less than 50% .



**Solution:**

**Step 1: Data Cleaning**

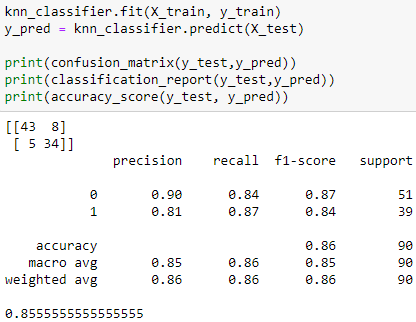
From EDA we learn that the dataset is very clean and hence there is no need for data cleaning or data manipulation to be done.

**Step 2: Data Processing**

All the categorical features are removed and are converted into quantitative features by making dummy features. The remaining features are scaled using the Standard Scalar. The dataset is Split into training and test set in the ratio of 70:30 respectively.

**Step 3: Find the model with highest accuracy**

Using the Evaluation Metrics such as classification report, confusion matrix, accuracy score, the accuracy of the models such as Logistic Regression, Naive Bayes, K-Nearest Neighbours, Decision Tree, Support Vector Machines is compared and the model with the highest accuracy is chosen. The model with the highest accuracy of 85.55% was the K-Nearest Neighbours classifier.

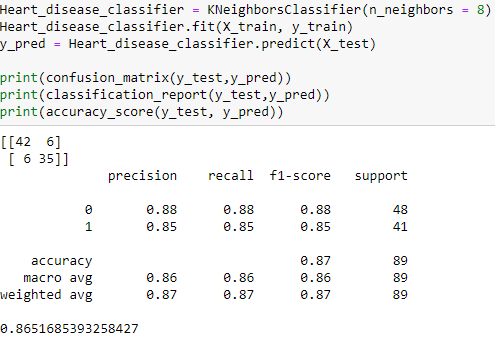


**Step 4: Check if removing outliers increases accuracy**

The feature ‘CHOL’ contains an outlier whose value is very high compared to the 75th percentile value of that feature. When we remove that outlier from the dataset and use the K-Nearest Neighbours classifier on the dataset, the accuracy of the model increases to 86.51%. Therefore, removing the outlier increased the accuracy of the model.

**Step5: Choosing the final model**

Since the K-Nearest Neighbours classifier model trained on the dataset with the outliers removed has better accuracy, that model will be used as the final model for Heart Disease Prediction.



**Limitations:**

* Since we are using a K-Nearest Neighbours classifier model, finding the correct value of ‘K’ for a large dataset is time expensive and computationally expensive as well.
* Even though the model is 86% accurate, it still generates a small amount of True negative. This can be reduced by increasing the accuracy to as close as 100% as possible