**PROJECT REPORT: HEART DISEASE PREDICTION**

# **Data Understanding & Preparation:**

## **Data Preparation:**

The data needs to be prepared before performing any kind of analysis. A number of data preprocessing steps are applied to clean and prepare the data for analysis. The structure of the dataset is checked and the figure attached below shows that there are 303 observations & 14 attributes including the target variable. Most of the variables which needs to be of type factor (as they are categorical) are of type integer.

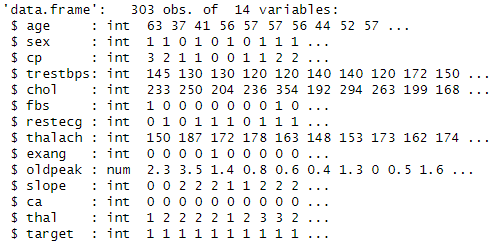


Figure : Glimpse of the dataset

First of all, the dataset is checked for missing values and it can be seen that there are no missing observations in the dataset.

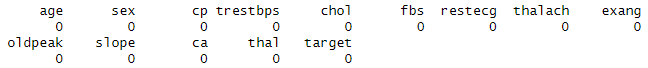


Figure : Count of missing values in dataset

The boxplot for numerical columns in the dataset are created to see whether there are any evidences of outliers or not. Based on the boxplots, the age column has no outliers. However, the resting blood pressure, cholesterol & maximum heart rate achieved columns has outliers. These outliers are removed from the dataset using inter quartile range method with a cut off value of 0.25 & 0.75. The IQR method remove all those observations which are either above the third quartile or below the first quartile. After removing the outliers from the dataset the dataset is now reduced to 288 observations. This means that 15 observations are removed which contained outliers.

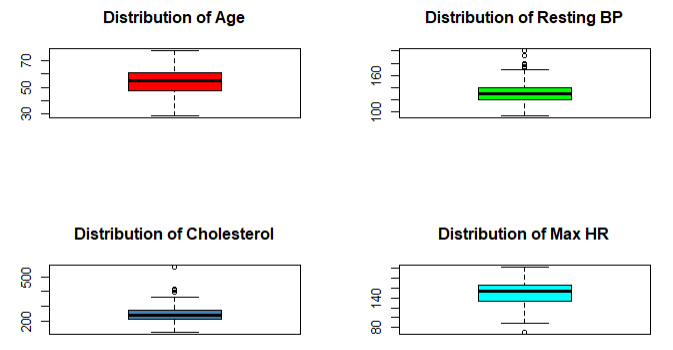


Figure : Checking outliers in numerical columns

The rest of the categorical variables are converted to factors and the glimpse of final cleaned dataset is attached below.

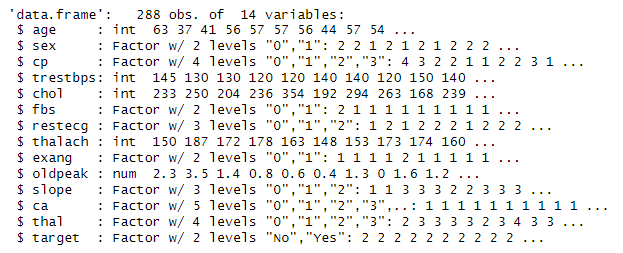


Figure : Glimpse of cleaned dataset

## **Data Understanding:**

In order to understand the data, the exploratory data analysis is performed. Three boxplots are created which shows the distribution of age, cholesterol, maximum heart rate & old peak grouped by the target variable which shows whether the presence or absence of heart disease. The boxplot attached below shows that the median age of those who have a heart disease is lower compared to those who do not have any heart disease. The median cholesterol level of those who have a heart disease is lower compared to those who do not have a heart disease. The median maximum heart rate achieved is higher for those who have a heart disease compared to those who do not have any heart disease. Finally, the median ST depression induced by exercise relative to rest is lower for those with a heart disease compared to those who do not have a heart disease.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure : Distribution of age, cholesterol, heart rate & ST depression by target variable

Moving forward a few stacked bar charts are also created which are attached below. The charts shows that those who have a heart disease, 43% are female while 57% are male. This shows that the prevalence of heart disease among male is higher compared to female. The highest percentage of heart disease patients are those who have a chest pain of type 2, followed by type 1, type 0 and type 3. The highest percentage of patients are those who have a slop of type 2, followed by type 1 & type 0. Finally, the distribution of the target variable shows that the frequency of those who have a heart disease in the dataset is higher compared to those who do not have a heart disease. The data is a very little imbalance and this imbalance can be ignored as it is very minimal.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure : Proportion of patients having heart disease by Sex, Chest Pain Type & ST Slope

# **Algorithms Techniques Chosen & Implemented:**

Three different machine learning algorithms are trained in order to predict the heart disease using all features that are present in the dataset. The dataset is divided into training and testing set with 70% of the data being used for training the model while 30% is used for testing the performance of the model. A logistic regression model is trained, a naïve bayes model is trained and a decision tree model is trained. All default parameters are used to train the models. The performance of all three models is tested on unseen 30% test set. The performance evaluation metrics that are used includes accuracy, kappa statistics, confusion matrix, sensitivity and specificity.

# **Results Discussion:**

Three models for predicting heart disease include Logistic Regression, Naive Bayes, and Decision Tree. These were examined and compared using performance criteria. Each model takes a unique approach to prediction, which is reflected in its accuracy, sensitivity, specificity, and other statistical characteristics.

Logistic Regression had the highest accuracy of 83.91%, which means it accurately predicted heart disease in 83.91% of cases. This model also demonstrated a strong balance in identifying patients with cardiac disease (sensitivity: 83.78%) and those without (specificity: 84.00%). The kappa score of 0.6731 indicates moderate agreement between the model's predictions and actual results, implying that the model is credible for this classification task. Overall, Logistic Regression performs well with little trade-offs, making it a promising contender for heart disease prediction.

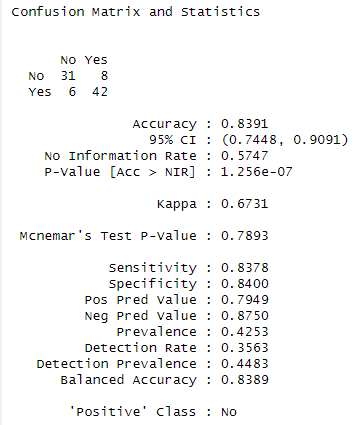


Figure : Performance of Logistic Regression Model

In contrast, Naive Bayes had a little lower accuracy of 81.61%, showing that it successfully predicted heart disease somewhat less frequently than Logistic Regression. However, it had a high specificity of 86.36%, indicating that it was quite successful at detecting people without cardiac disease. The sensitivity was lower (76.74%), indicating that the model missed some cases of heart disease. Naive Bayes has a kappa value of 0.6317, indicating moderate agreement but somewhat less than Logistic Regression. This model may be preferred when it comes to correctly identifying those who do not have the condition over recognizing those who do.

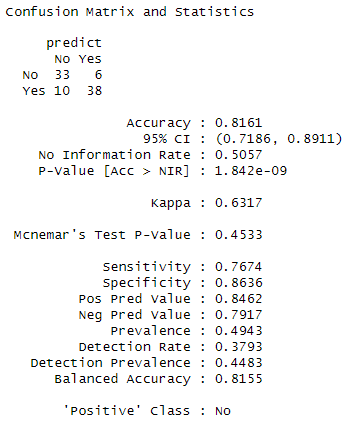


Figure : Performance of naive Bayes Model

Finally, the Decision Tree model had the lowest accuracy (77.01%), indicating that it was the least dependable in predicting heart disease accurately. The sensitivity was 74.36%, indicating a reduced capacity to detect individuals with heart disease, while the specificity was 79.17%, indicating intermediate efficiency in detecting those who did not have the ailment. The kappa value of 0.5353 indicates a reasonable level of agreement between predictions and actual results, but it is lower than the other two models. Although the Decision Tree is a simple and understandable model, its performance is not as strong, making it a less appealing option for heart disease prediction.

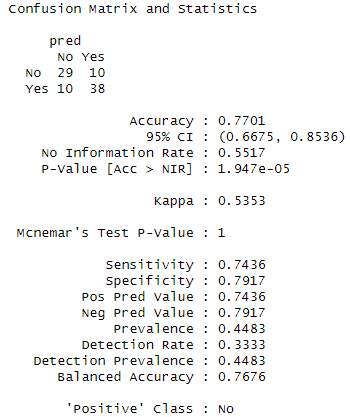


Figure : Performance of Decision Tree Model

In conclusion, while all three models have unique capabilities, Logistic Regression is the most balanced and accurate model for heart disease prediction in this comparison, followed by Naive Bayes and Decision Tree.