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

**CRN:** 81941

**Professor**: Marco Montes De Oca

**Final Project: Predicting Heart Condition using Neural Network**

**Submitted By:**

**Rishhikumar Janakiraman** [janakiraman.r@husky.neu.edu**]**

**Nikhil Sanjay Thorat [**thorat.n@husky.neu.edu**]**

**Megha Ravi [**ravi.m@husky.neu.edu**]**

1. **PROJECT DESCRIPTION**

The dataset “Heart Disease” has been taken from Kaggle. The origin source of the data is the UCI machine learning repository containing the results of 303 patients on tests such as blood pressure, cholesterol levels, heart rate resulting in patients being classified based on a positive or negative heart condition. The main idea behind choosing this dataset is to predict the chances of the patients getting heart diseases using various machine learning and deep learning models. The number of cases for heart disease is increasing day by day due to the modern lifestyle and food habits. The diagnosis of heart disease was a challenging task and the classification models will predict whether the patient has heart disease or not based on various conditions/symptoms of their body.

1. **DATASET - INDEPENDENT AND DEPENDENT VARIABLES**

The dataset has multiple variables which are classified as independent variables and a dependent / target variable and all these variables are used during predictions in each model. The dataset contains303 samples with 13 independent feature variables which are - Age, Sex, Chest Pain type, Blood pressure, Cholesterol, Fasting Blood sugar level, resting electrocardiographic results, Depression induced by exercise, angina, max heart rate achieved, slope, major vessels and blood disorder called thalassemia. The target contains one variable - “target” which is a categorical data showing 0 for negative on heart condition and 1 for positive. There is no missing values in these feature variables.

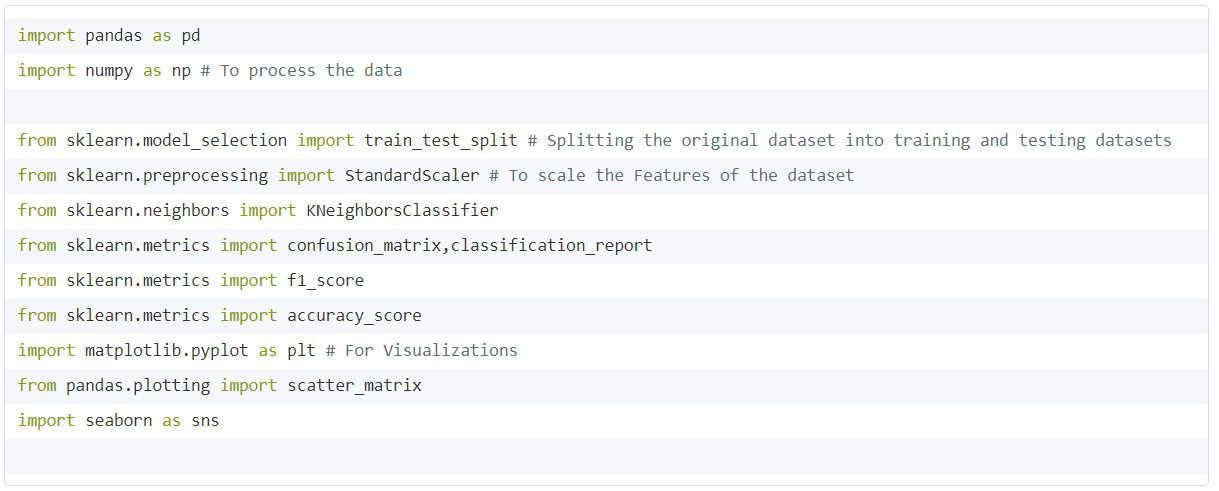
1. **DATA ANALYSIS**

The analysis part of the heart disease dataset has been carried out in Python by initially installing all the necessary libraries that are required to create the ML models and further investigation of the data has been made with making visualizations that highlights the importance of the feature variables. Below are the steps involved in working with the data:

* Data Exploration
* Feature Engineering
* Modelling
* Optimization
* Results of the models to predict the accuracy

**LOADING LIBRARIES**

Python Jupyter notebook has been used to make the analysis and this environment comes with many analytical libraries installed and below are some of the packages and library functions required to run our models.

****

**EXPLORATORY DATA ANALYSIS AND VISUALIZATIONS ON THE DATASET**

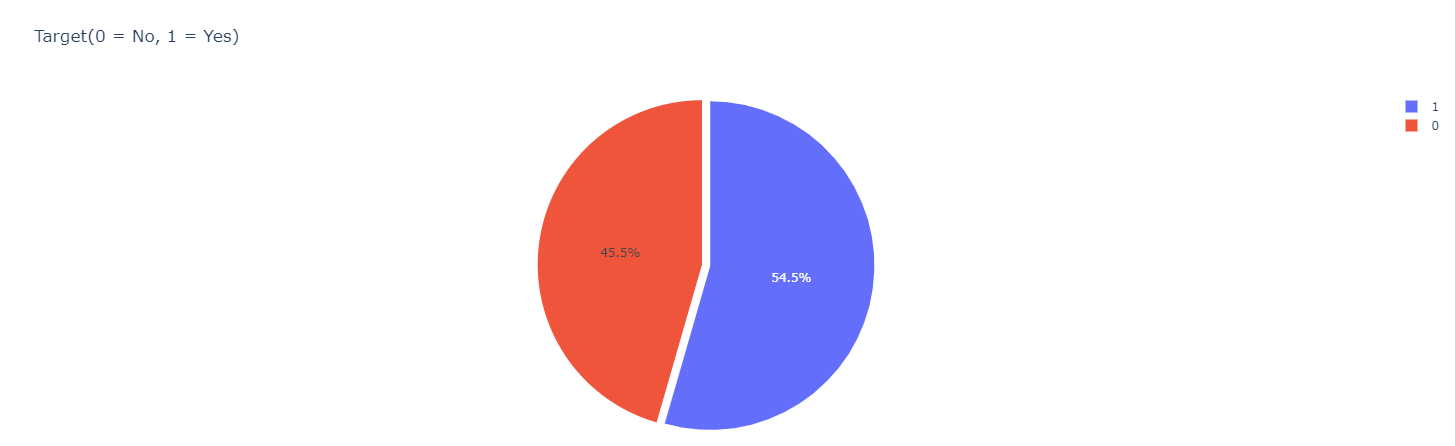
The goal is to explore the data and understand the nuances which helps us to get closer to the main objective of predicting the heart disease in patients. These EDA was useful to identify if there were any missing values or outliers that affected the performance of the ML models that we had created to predict the outcome. Some of the EDA has been demonstrated in the form of visualizations for better understanding.

**To Show the Number of Patients having Heart Disease**

Based on the project objective we have considered “target” column as our dependent target variable which shows whether the patients have heart disease or not. From the below pie chart, we can see that almost 45% of the admitted patients does not have the heart disease, while on the other hand more than 54% of the patients have heart disease.

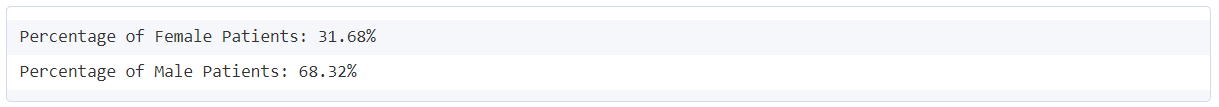
****

Pie Chart that visually represents the Target variable :

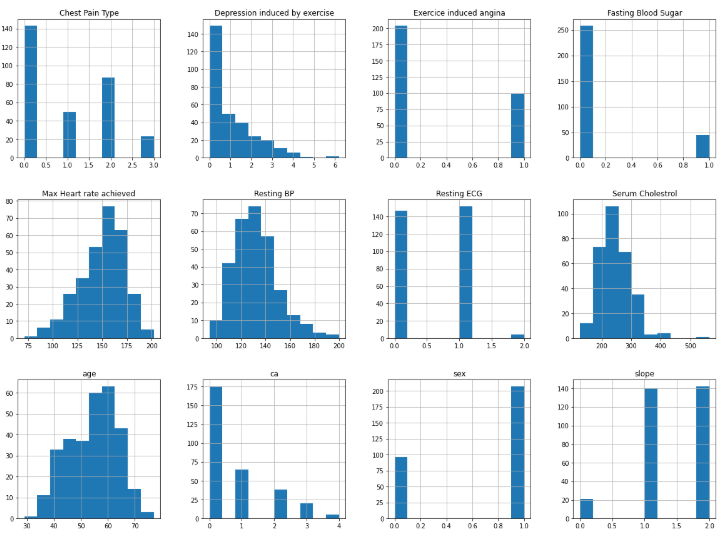


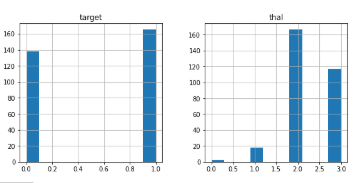
**To Show the comparison of Male and Female patients:**

The proportion of male and female observations recorded in the dataset acts as an important factor while making conclusions on the ML results. There are 31.68% of female records and 68.32% of male records.

****

**Histograms of the feature and target variables:**

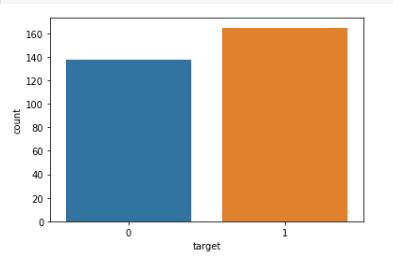
****

****

The above histogram shows that each feature has a different range of distribution and it is very important to scale these feature variables before passing through the model to avoid overfitting or unwanted redundancy of the model performance.

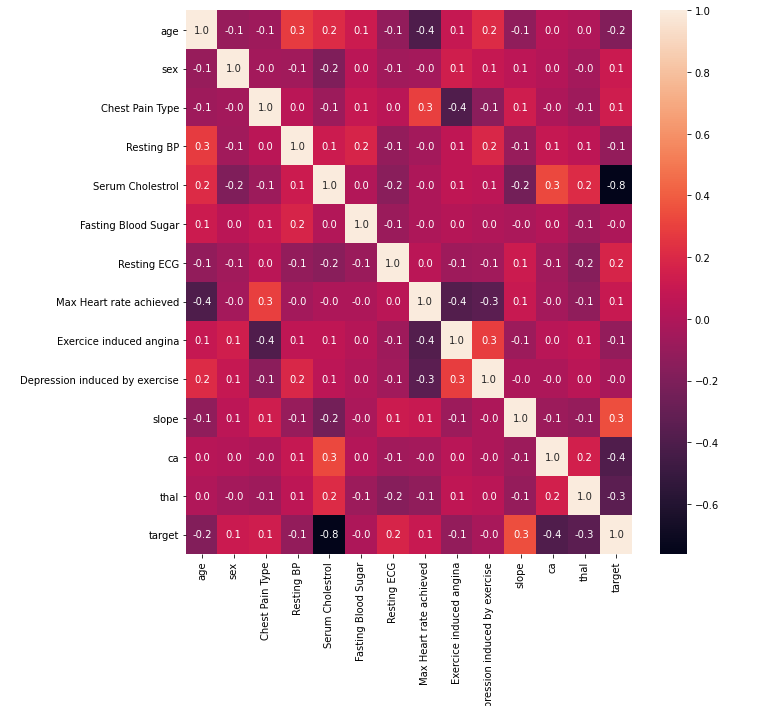
**To Show the Comparison levels of the Target variables**

The target variables are classified into 2 class labels, 0 and 1, 0 being negative cases and 1 being positive cases. There is a total of 138 negative cases and 165 positive heart condition cases for all observations.

****

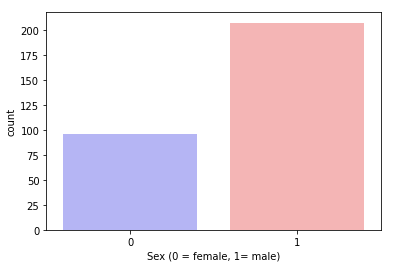
**To Show Correlation plot of the dataset**

This is a correlation plot of the feature and target variables in the dataset and it is seen that there is no feature with 0.5 correlation which shows the dependency of these variables on the accuracy of the models created.



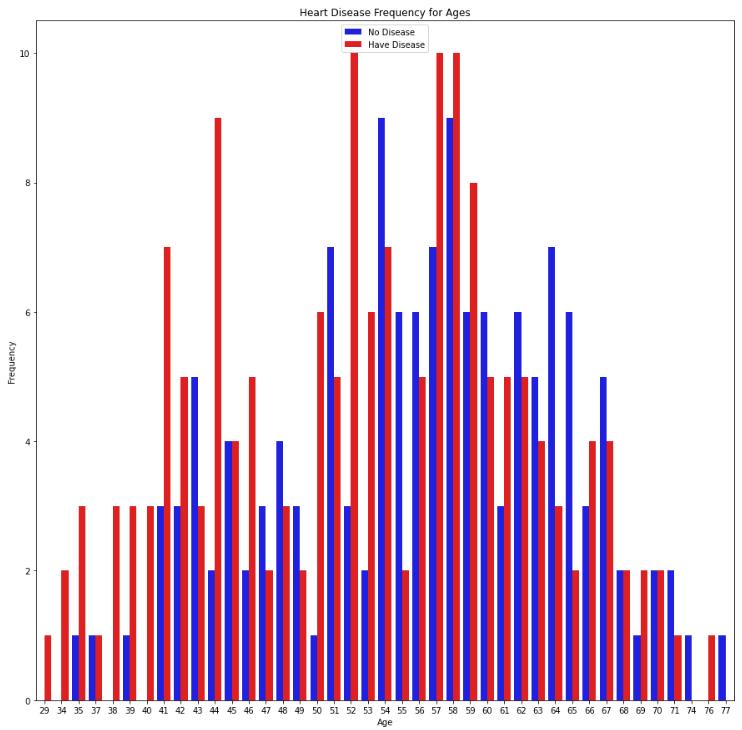
**Male and Female count visualization**

The bar chart below shows the number of male and female observations recorded for a positive heart condition. The total number of females in the dataset with a heart condition are 96 and the total number of males in the dataset are 207.

****

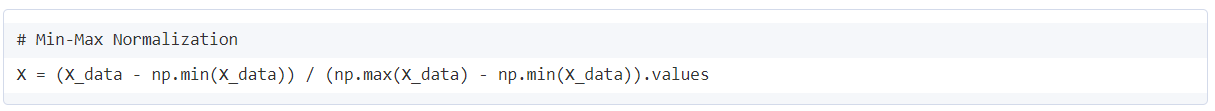
**Heart Disease Frequency by Age**

The below bar graph shows the frequency of positive heart conditions in the population with respect to age. People in the age gap of 50 to 60 are most susceptible to a heart condition when compared to anyone in the range of 30 to 40 or 67 to 77. The heart conditions gradually increase and reach the highest at 52 and gradually decrease later. Mostly after 60 (closer to 60), the proportion of people having no heart condition is more when compared to people having a heart condition and becomes equal after the age of 70.

****

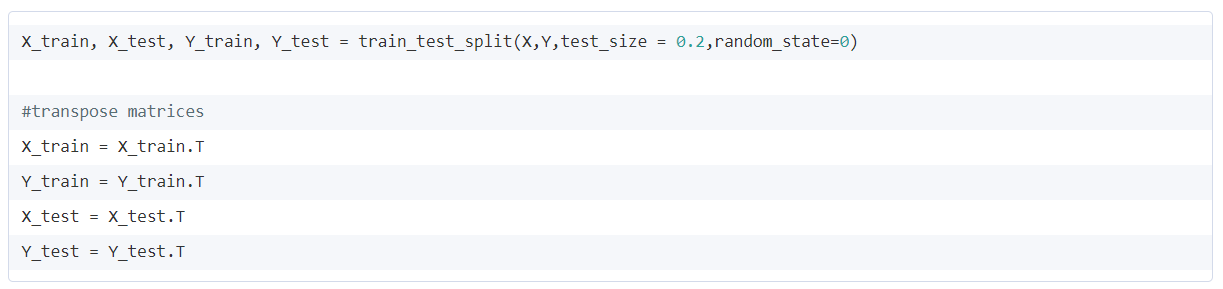
**NORMALIZATION OF DATA**

After exploring the dataset, it was important to convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models. By doing this, we aim at reducing redundancy in the dataset and to improve data integrity. First, I'll use the get\_dummies () method to create dummy columns for categorical variables.

****

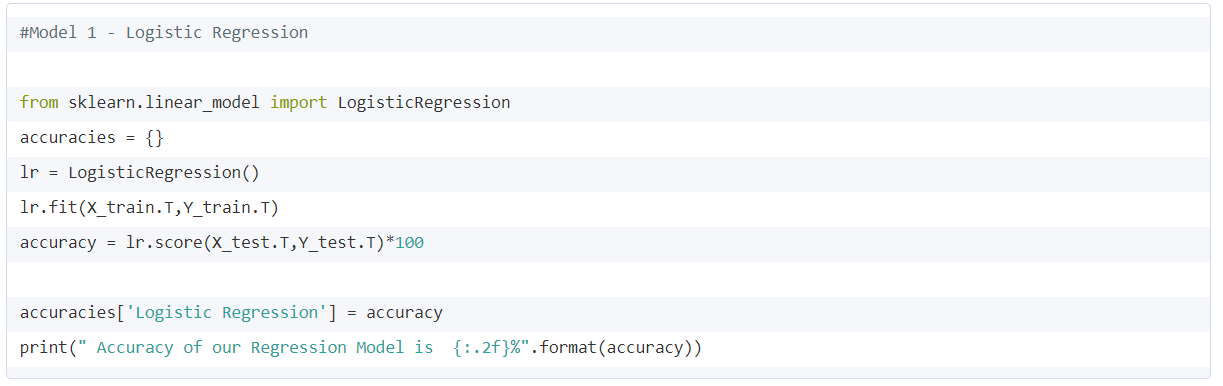
**SPLITTING OF TRAINING AND TEST VARIABLES FOR THE MODEL**

Splitting the dataset in 80/20 proportion with 80% train data and 20 test data. X\_train and X\_test having independent variables and Y\_train and Y\_test variables hold the target variables.

****

**MODEL 1 - LOGISTIC REGRESSION**

The classification model is generally used when the target variable is binary i.e., have values which are 0/1 or yes/no. Since the model is built to predict if an individual will have a heart condition, it is vital to use all the features recorded.

We have trained the Logistic regression model to find the patterns on the training set using the LogisticRegression() classifier function and then later fit the model to determine the trends in the testing set which predicts the outcome. ****

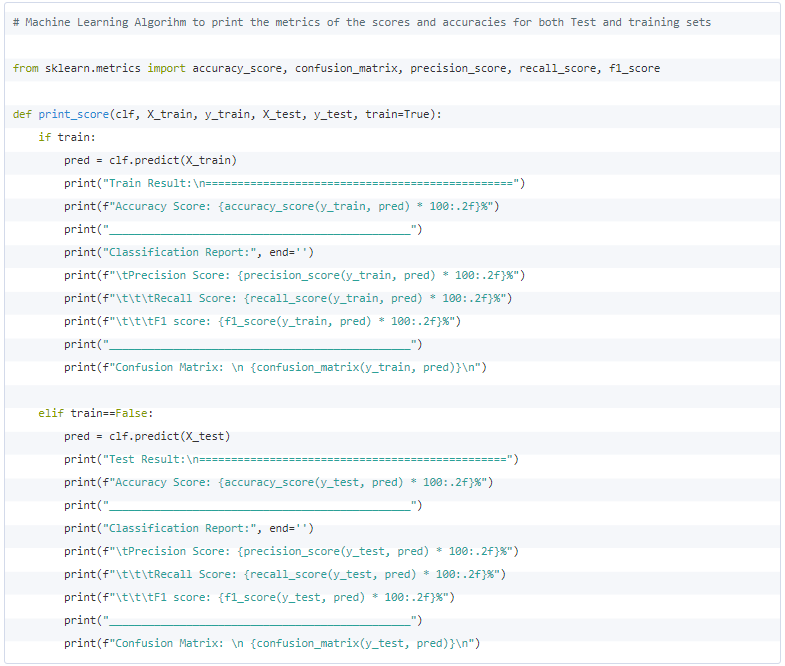
**MODEL OUTPUT**

Our regression model showed an accuracy of 85% which showed the smallest difference between the predictions of both training and test data.

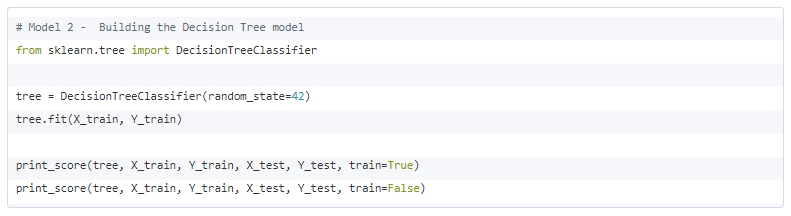
****

**MODEL 2 - DECISION TREE MODEL**

The next ML model designed to predict the heart disease outcome is by using the classification algorithm of decision trees. Below is the code which defines the prediction metrics for the model and they are passed as objects.



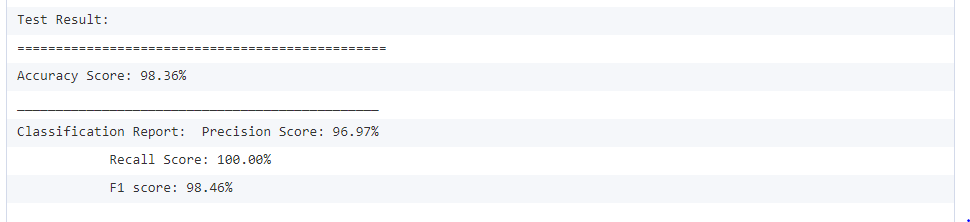
Decision tree classifier is a tree in which internal nodes are labeled by features and the classifier categorizes an object by recursively testing for the weights that the features labeling the internal nodes have in vector, until a leaf node is reached. The dataset is again split into 2 parts where the training data is passed through the classifier and the other will evaluate it.



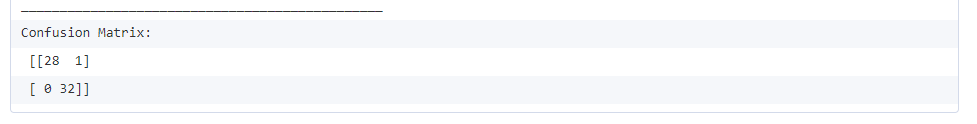
This model combines multiple nodes of different depths in predicting the model. Here we have used decision trees for improving our accuracy of the model as it reduces the overall complexity of the model that is being built.

**MODEL OUTPUT**

The classifier model output for the decision trees gives the accuracy of 98.46% which is obtained without hyperparameter tuning. These results highlight that the variables passed through the model are not dependent on each other which affects the accuracy of the model. We observed the bagging effect most strongly with random forests model because the base-level trees trained with random forests have relatively high variance due to feature subsetting and that is the main reason for the accuracy to be very high.

****

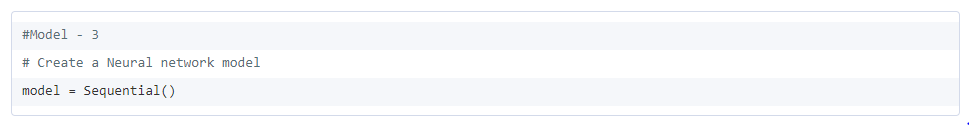
**CONFUSION MATRIX FOR TREE MODEL**

****

**MODEL 3 - NEURAL NETWORKS**

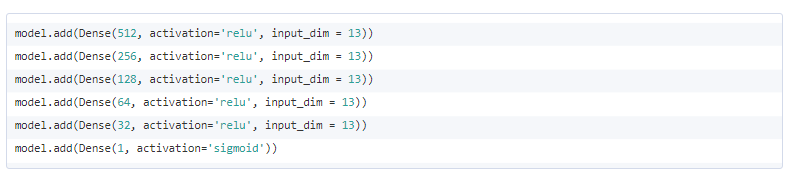
The last model for this project is to build a neural network that can help the classification problem and in turn predict the causes for heart diseases with at most precision by training the nodes of the network.

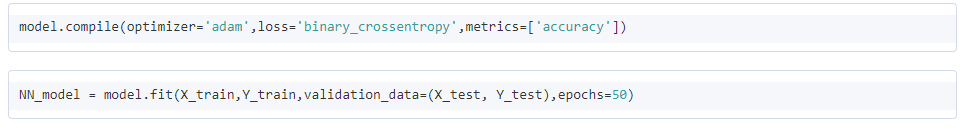
The model we have considered for the NN model is Sequential () with many hidden layers and 1 one output layer.

****

We have used ‘relu’ activation function for all the hidden layers except for the output which is defined as the positive part of the argument and the activation function for the output is ‘sigmoid’ as the variables have either 0 or 1 to predict the outcomes of heart disease prediction.

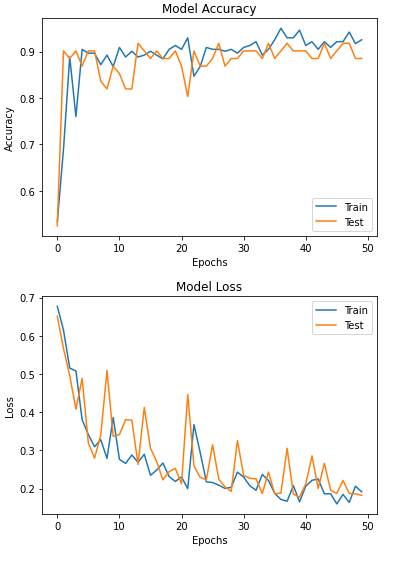
These 6 hidden layers are trained for multiple neuron values and iterated using epoch and multiple combinations have been tried to determine the accuracy of the model.

For our NN model , we have instantiated the optimizer before passing through the compiler and we used ‘adam’ optimizer which follows the Gradient descent method and the loss function used is ‘binary\_crossentropy’ since the output variable is binary values and the metric used to determine the accuracy is ‘accuracy’. After compiling the data, the variables are fitted for various iterations specified as ‘epochs’. On increasing the epochs, the layers and the number of neurons per layer resulted with 90% accuracy and not any higher.

****

**MODEL OUTPUT:**

The below graphs show the model accuracy and the loss of data occurred when passing through the neural network model. As we can see, the accuracy of the model increases with the increased layers and iterations on the other hand decreasing the data loss. By increasing the number of layers the network learns the algorithm with better precision and as the iterations increases the accuracy tends to increase. The maximum accuracy reached is 90%.

****

Although our model has strived to achieve promising results, there are still large error with the loss of data for every iteration. This could be because it is very difficult to distinguish between the different severity levels of heart disease.

**SUMMARY**

Machine learning and AI technology have been widely used in the medical industry to enhance patient experience and for better diagnosis, hence we’ve selected this dataset to get a hands-on how implementations are done. The objective of this project is to clean the dataset by checking for outliers and missing values, classify the variables as independent and dependent and understand how each independent variable affects the target variable by applying 3 supervised machine learning models, logistic regression, Decision tree classifier and Neural networks. The accuracies of the predictions from different models are compared to check which gives better prediction.

* Logistic regression had an accuracy of 85.25%
* Decision tree classifier had an accuracy of 98.36%.
* Neural Networks had an accuracy of 90%.

Decision Tree Classifier gives the highest accuracy when compared to the other two models. Decision Tree Classifier is one of the best supervised machine learning algorithms giving higher accuracy. The function used is automated to select features and a point where each feature is categorized as more or less from the selected point and builds the tree involving all the input features.

**RESULTS AND COMPARISONS**

The highest accuracy of the model was received when Decision Tree classification was applied when compared to Logistic regression or Neural networks. As mentioned, Decision tree classifier has a track record of resulting with higher accuracy because of its modeling structure. Techniques like feature engineering or hyper parameter tuning could be used but the accuracy of the models were high without implementing them, showing the importance of all the variables involved in the classification. The result also shows higher true positive and true negative cases when compared to the other 2 models.

**OBSERVATIONS FROM WORKING WITH THE DATASET**

1. From the EDA and visualizations, we have observed that there are 165 patients with heart disease and 138 patients without heart disease.
2. All other variables have a significant correlation with the target variable.
3. Women's hearts are affected by stress and depression more than men's heart and Depression makes it difficult to maintain a healthy lifestyle.
4. The person having the heart rate over 140 is more likely to have heart disease, therefor we conclude that we must check our heart rate monthly if it’s over the thalach 140 then we have to consult the doctor and much conscious to the health.
5. The person’s age between 40 to 65 is more likely to be affected by heart disease, therefore they need to be more conscious about their health.
6. 68% of male have a heart condition and 32% of female have a heart condition with highest accuracy got from Decision tree model.

**REFERENCES**

[1] Cardiovascular diseases. (n.d.). Retrieved from<https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1>

[2] Real Python. (2020, January 13). Logistic Regression in Python. Retrieved May 30, 2020, from <https://realpython.com/logistic-regression-python/>

[3] 1.10. Decision Trees¶. (n.d.). Retrieved June 25, 2020, from <https://scikit-learn.org/stable/modules/tree.html>

[4] Brownlee, J. (2020, April 23). How to Choose Loss Functions When Training Deep Learning Neural Networks. Retrieved from <https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/>

[5] Yiu, T. (2019, August 04). Understanding Neural Networks. Retrieved from <https://towardsdatascience.com/understanding-neural-networks-19020b758230>

[6] Gandhi, R. (2018, May 17). Improving the Performance of a Neural Network. Retrieved from <https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d>