PROJECT REPORT CS401

Bones Dataset Classification

MU Registration No:20250154 | Name: Abhishek Kumar Pandey

1. Data Preparation

The text file **'train-io.txt**' is imported into the Jupyter notebook and data preprocessing tasks like checking for null values, NA values and imbalance in the dataset is done. It is seen that the dataset doesn't contains any null/NA values. Also, there is no class imbalance in the target variable and same has been shown using seaborn countplot. The shape of the **'train-io.txt'** dataset is (300000,13) and for '**train-io.txt**' **is** (10000, 1).

Also, **Searborn** is used to check the correlation between the variables using heatmap.

Using **StandardScaler** from **sklearn** library is used to standardize the values as the mean values for the independent variables is close to 0 but the max values for some independent variables are larger than the mean value.

Also, **train_test_split** is being used again from the **sklearn** library to divide the 'train-io.txt' dataset in X_train, X_test, y_train, y_test respectively with training set split into 70% of the dataset and test set into 30% of the remaining dataset.

Similar steps as mentioned above have been performed for both the '**train-io.txt**' and '**test-in.txt**' **except the train_test_split step.**

2. Model Selection/Evaluation

After the data preparation step, I have selected few classification machine learning algorithms for our bone dataset. The algorithms are as follows:

1. Logistic Regression
2. Decision Tree
3. Naïve Bayes Algorithm
4. Artificial Neural Networks

Classification has been performed on the training dataset '**train-io.txt**' and confusion matrix/classification metrics from **sklearn** library has been used. ROC AUC score has been used to know the performance of the model. Also, ROC curve has been shown for all the models whether it has been rejected or accepted. Similarly, the confusion matrix for all the machine learning algorithms has been added the notebook.

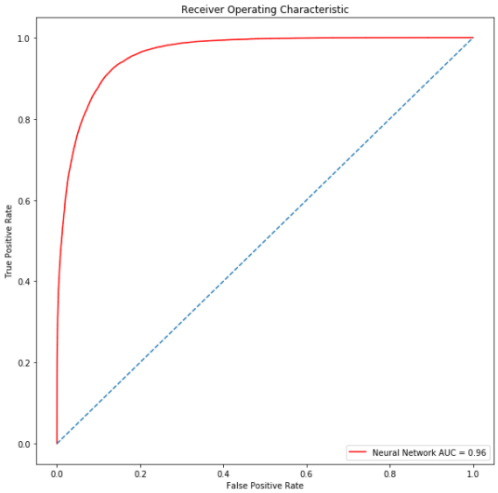
The ROC AUC scores for all the model has been mentioned below and in the jupyter notebook respectively:

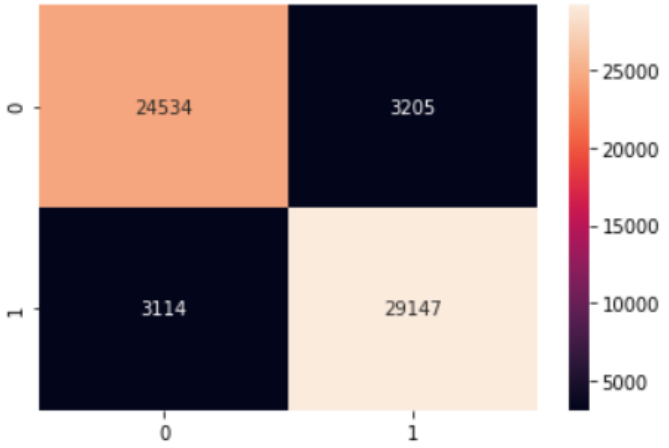
1. Logistic Regression- 0.4993874
2. Decision Tree- 0.5193449
3. Naïve Bayes Algorithm- 0.5659498
4. Artificial Neural Networks- 0.96088969

The more the AUC score, the better is the model distinguishing between positive classes and negative classes.

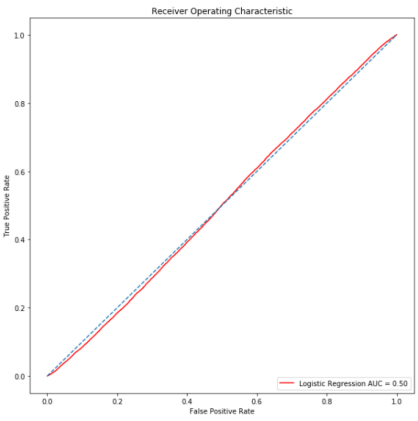
Considering the scores, it is seen that the Logistic Regression/Decision Tree/ Naïve Bayes doesn't perform nicely as they are not able to identify the binary classification. So, out of all the scores from the above, Neural Nets has outperformed all the other machine learning algorithm by producing a score of 0.96088.

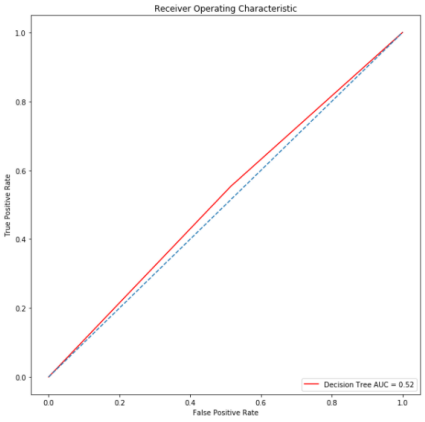
ROC curve and confusion matrix for the model being settled on **i.e; Neural Networks** is shown below:

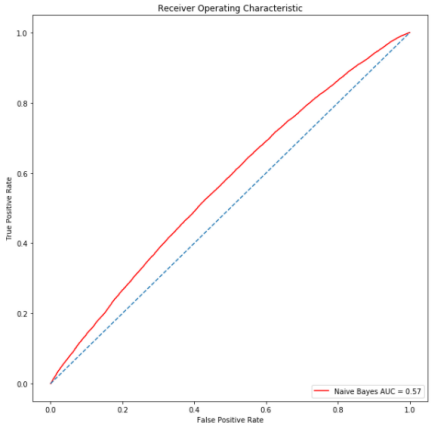


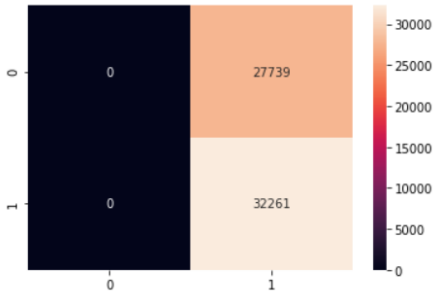


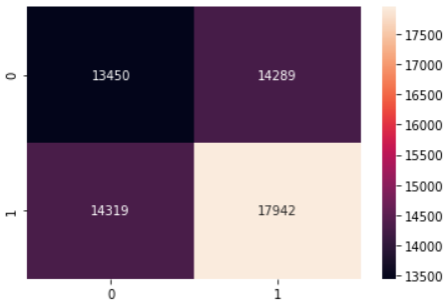
ROC curve & Confusion matrix for the rejected Logistic Regression, Decision Trees and Naïve Bayes classifier has been added respectively:

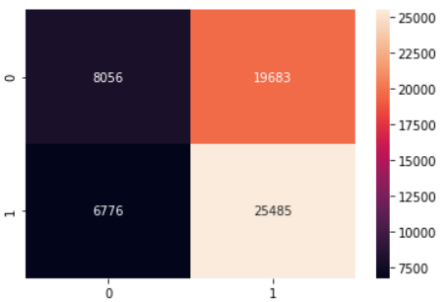












3. Predicting Output and saving it in **test-out.txt**

The Neural Networks were used to classify because of the outstanding performance w.r.t other machine learning algorithm. The predicted output was stored in **test-out.txt** file with the name on the first line followed by the binary output 0 or 1.