IEE 520 – FALL 2020-FINAL PROJECT REPORT

Submitted to:

Dr. Runger
School of Computing, Informatics, and Decision Systems Engineering.

Submitted by:

Pranay Reddy Vancha
Industrial Engineering
ASU ID: 1217852909

Objective:

Our Project objective was to build a classification model for the dataset given based on the methods described in the course. We need to Pre-process the data before building any classification model.

Pre-Processing of Data:

For the pre-processing of data, we had to do some simple manipulations to the dataset given. Here are the steps done to pre-process the data.

1. Firstly the instance column is removed from the labelled dataset given as this attribute should not affect the outcome of our classification model.

2. There are some null values in the dataset. Column 1-12 were categorical data and there were some null values in column 1, column 4 and column 8. These Null values were replaced with mode of that column.

Model Building:

Using the pre-processed data classification models were built and the accuracy of each classifier were compared. We had the following accuracies for the classifiers we built.

Classifier	Accuracy
Support Vector Machine	85%
Decision tree	87%
Random Forest	89%

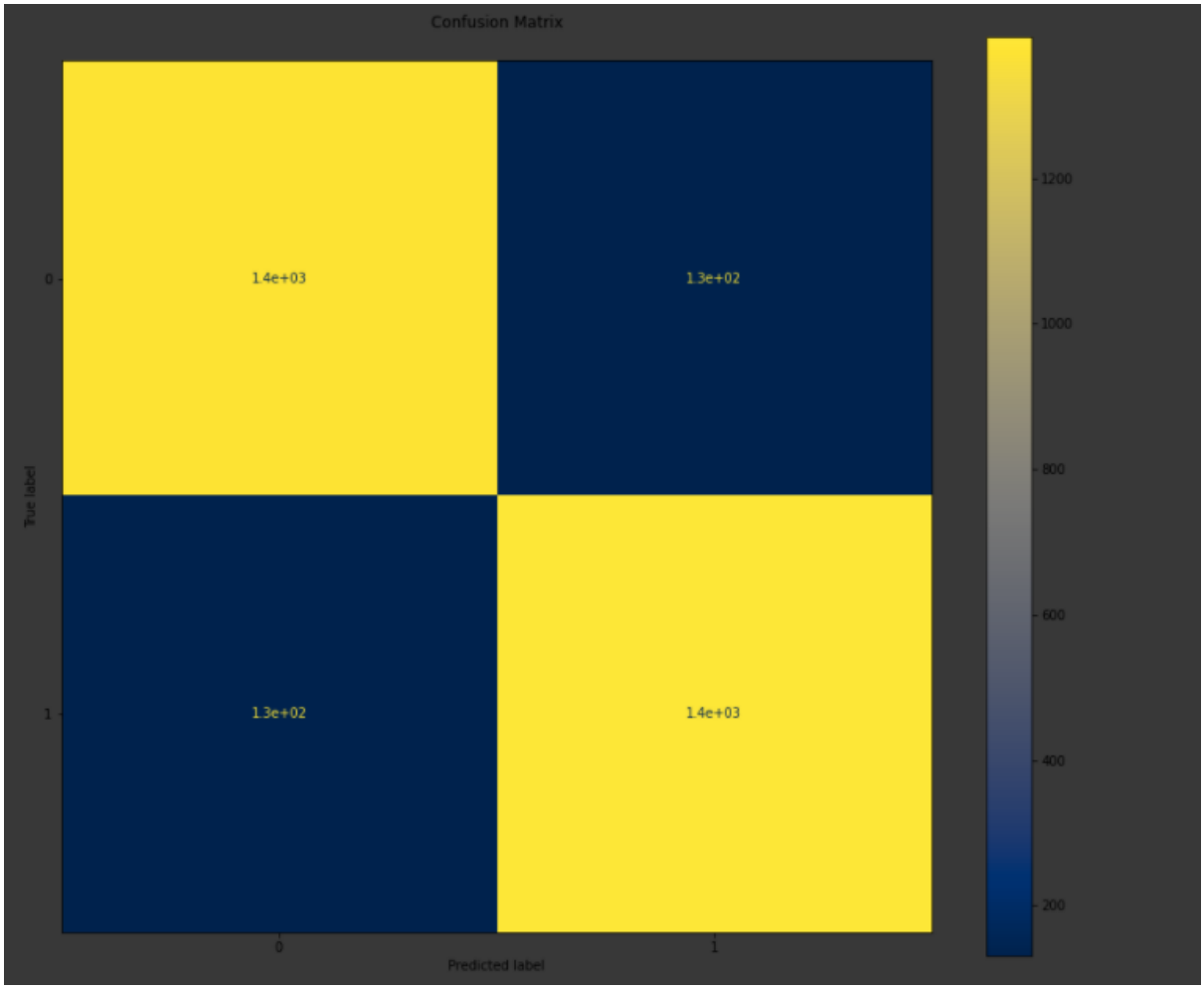
As the Random forest which can be considered as the bagging of Decision Tree Classifiers which resulted in the increase of accuracy by 2 percentage points, I also tried the boosting of Decision tree Classifier.

We used the AdaBoostClassifer model from the Scikit package. GridsearchCV was used to identify the best parameters for the base Decision Tree classifier and those parameters were used to define the base classifier to be used in the AdaBoost classifier.

The parameters that were used are:
Algorithm: SAMME.R
base_estimator__criterion: entropy
base_estimator__splitter: random
learning_rate: 0.01
n_estimators: 150

We have the generalization error to be **8.60%** from the accuracy of the model on the test data. The accuracy is **91.39%** for the boosted model with a base Decision tree classifier.

The Confusion matrix for the model is given below:



The balanced Error rate is then calculated from the confusion matrix as:

$$\text{Balanced Error Rate} = \left(\frac{(130)}{1400+130} + \frac{(130)}{1400+130}\right)/2$$

which gave us the error rate of **8.4967%.**

Prediction

Based on the balanced error rate and accuracy Adaboost classifier built on the basic decision tree is used for the classification of the test dataset.

The Adaboost Classifier applied on the basic Decision tree has been used to predict the unlabelled test data and the output was taken in the form of a .csv file. Out of the 5000 instances in the test data 3946 of the instances were classified as Class ‘0’ and the remaining instances were classified as Class’1’. The resulting .csv file is attached along with the project report.