Weak Pattern Theory

**Project report**

**COSC 6342: Machine Learning**

**Submitted by**

**S M Salah Uddin Kadir (1800503)**

**Rubayat Jinnah (1891217)**

Classify any dataset with perfect accuracy! In this paper, We are going to propose a theory that can be used to classify a dataset to get the perfect accuracy. We know, strong always ***dominates*** the weak by nature. When we build a classifier from a dataset, the classifier represents the maximum examples, ***strong pattern***, by dominating the other minor groups, ***weak patterns***, of examples. So, the idea is that if we can separate the weak pattern examples from the dataset and build a different classifier then it is possible to achieve a very good result close to perfect accuracy.

To verify the idea, We have used the Census Income Data Set [1] where the accuracy of the dataset using a single classifier is low.

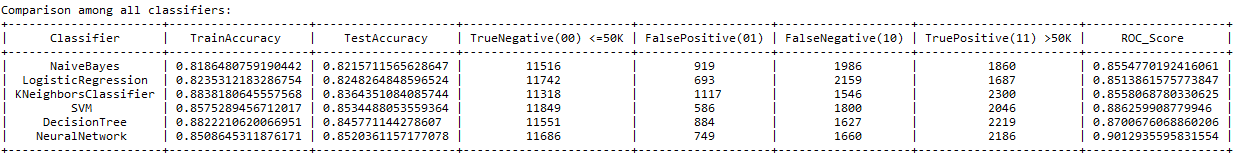
**The Analysis on the Dataset**

Our used algorithms to test the dataset,

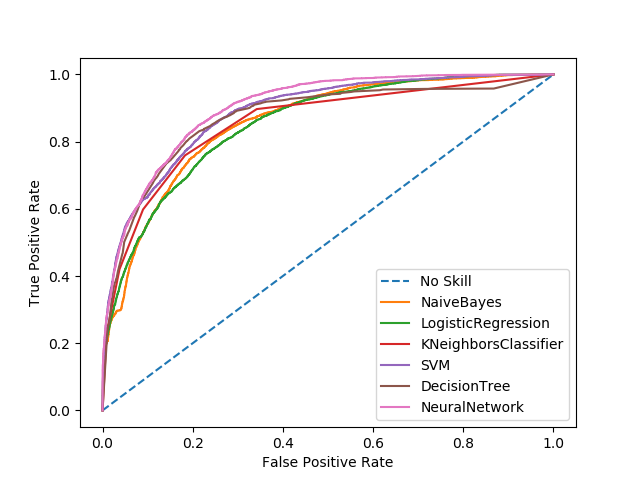
* Naïve Bayes,
* Logistic Regression
* k-nearest Neighbors
* Support Vector Machine
* Decision Tree and
* Neural Network

First, we will have a comparison table of performance among those algorithms,

(*Please zoom in to read the table, sorry for the inconvenience. We made this to keep all things together for the easy comparison.*)



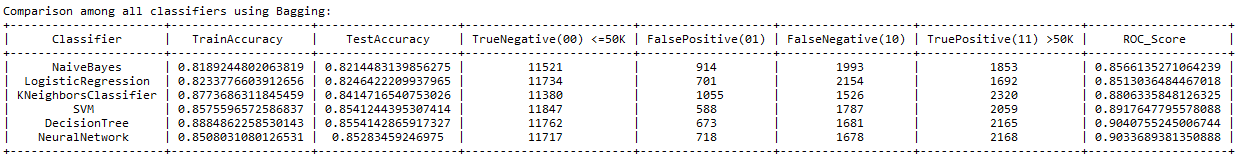
And the corresponding ROC curve comparison,



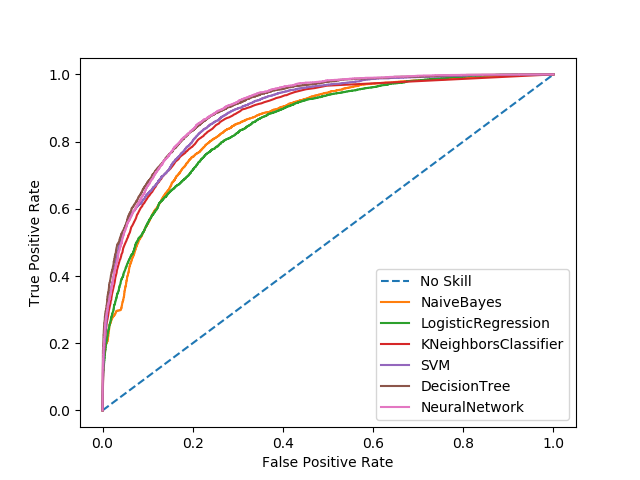
From the table, we see that we have obtained the maximum test accuracy from SVM (85.34%) and the maximum ROC score from NeuralNetwork that covers 90.13% of the plot.

We wanted to check whether we can get better results using Ensemble algorithms. We used Bagging and Boosting techniques,

***The same comparison table using Bagging***,

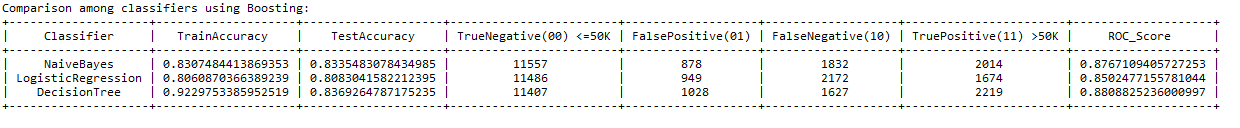


And the ROC plot comparison,

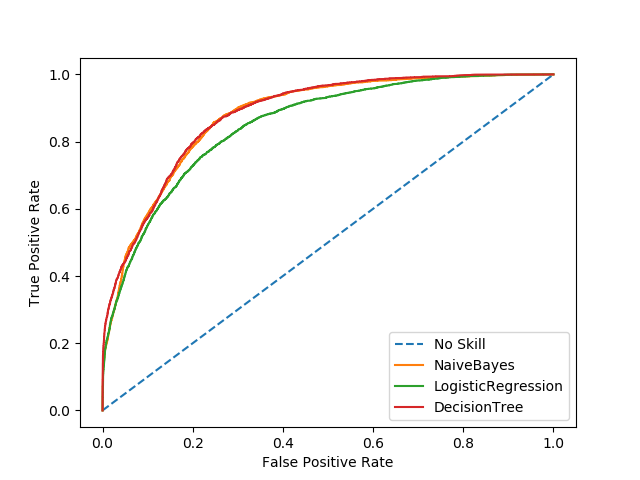


Here, we have got the best result from the decision tree with 85.54% accuracy on the test dataset where the ROC score is 90.49%.

And for the Boosting,



With the corresponding ROC curve,

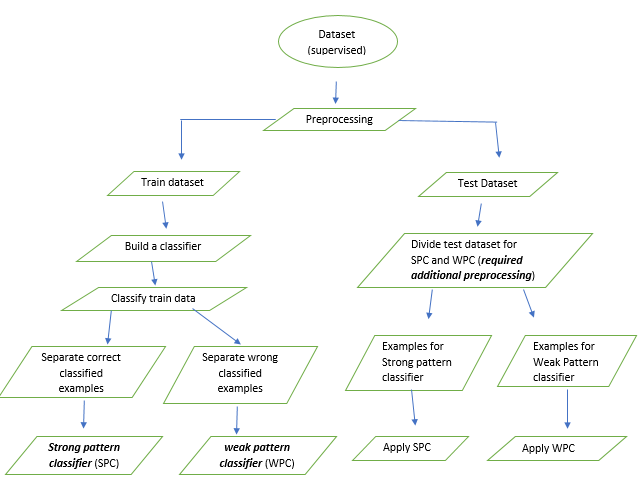


Using Boosting, we have got the best test accuracy (83.69%) and the ROC (88.89%) score from the decision tree, no better than the previous scores.

**A different idea**

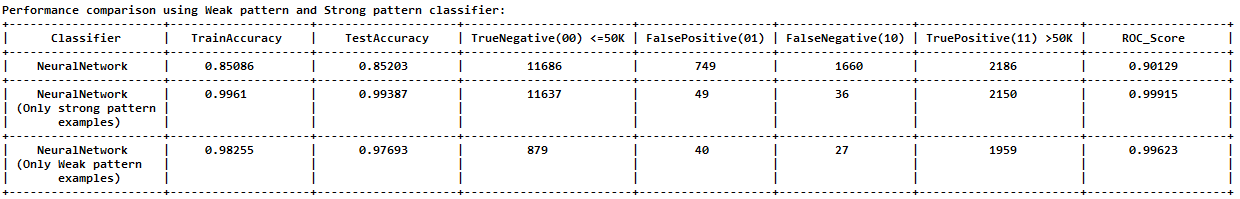
So, we can conclude here that the obtained result is not good enough. To find out the reason, we analyzed separately the correctly classified examples and the examples that are not. We discovered a pattern using the wrong classified examples that is ***very different*** from the pattern using correctly classified examples. So, the pattern from the wrongly classified examples is ***the weak pattern***, and the pattern from the correctly classified examples is ***the strong pattern***. Now, if we can add one more ***preprocessing step*** to separate the weak pattern examples from the strong pattern examples and classify them using different classifier, then it is ***possible to achieve a very high accuracy*** close to perfect.

The flow chart for the weak pattern theory,

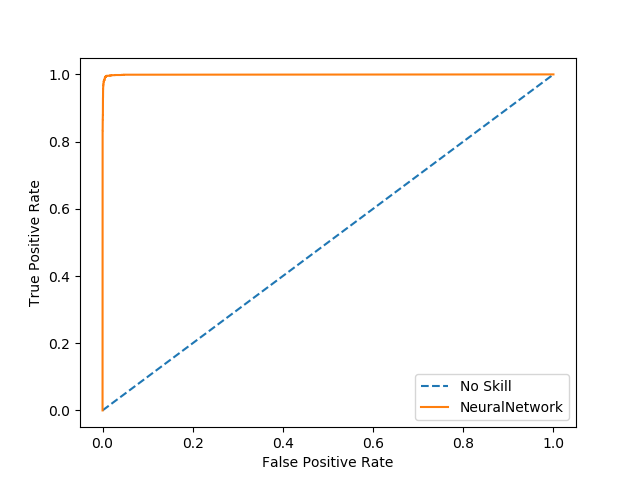


Let’s assume, ***somehow***, we can separate the weak pattern examples from the strong pattern examples so that we can ***verify our weak pattern idea***,

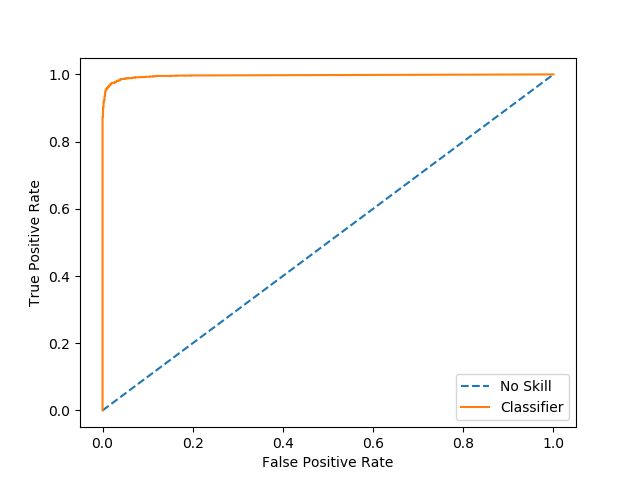
Let’s check the test result using weak pattern and strong pattern,



Corresponding ROC plot for ***strong pattern classifier***,



And the ROC plot for the ***weak pattern classifier***,



From the table, we see that we have **obtained a very high accuracy** on the test data for both strong pattern classifier (99.34%) and the weak pattern classifier (97.69%). The ROC score is also very high, for strong pattern classifier is 99.92%, and the weak pattern classifier is 99.62%.

**References:**

1. The Dataset (<http://archive.ics.uci.edu/ml/datasets/Adult>)