Haensel AMS
Advanced Mathematical Solutions

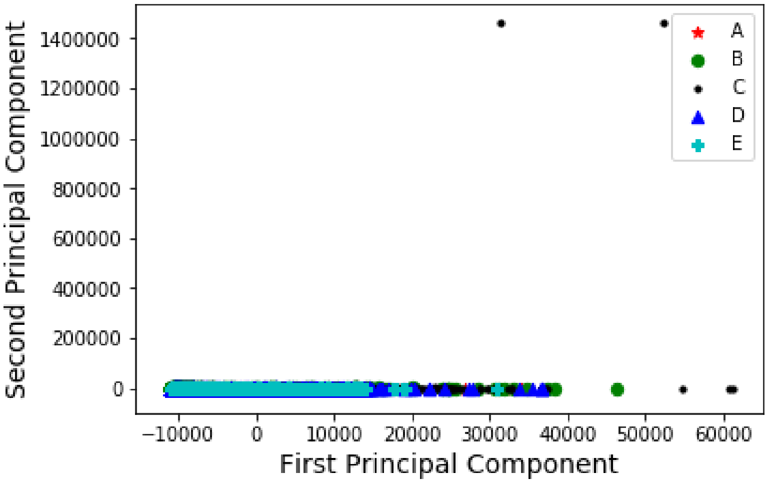
Analysis Report

The data provided consisted around 66k records with 295 features and target classes. There were about 5 classes. The task was to predict the classes for 295 features.

First Look/Data Analysis:

When the .csv file was accessed using pandas package of python , the column names were not present. The column names starting from *column1,.., column296* was named. Hence *column296* is the target column and c*olumn1,.., column295* are features.

The data visualization is done by reducing the features to 2D using PCA. The visualization obtained was something like below.



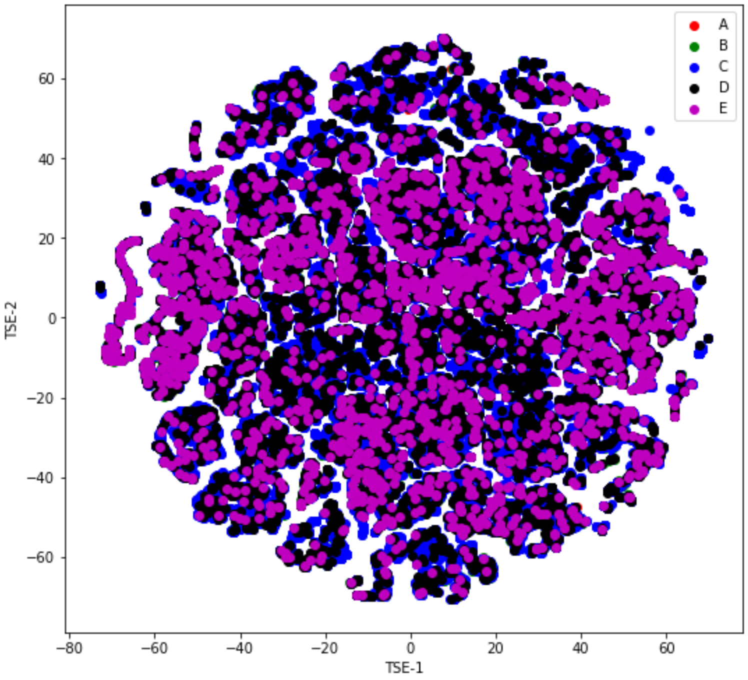
Since large number of non-linear, overlapping classes are present the data visualization using PCA was not succesful.

Further, the data visualization is done using T-SNE:

t-SNE is the algorithm based on Student T Stochastic Neighbor based distribution. Few of the points that can be mentioned are:

- This helps in projecting high dimensional data into low dimensional data. It helps in capturing the non-linear structure.
- Firstly, Gaussian distribution is applied defining the relationships between the points in high-dimensional space. Further, low dimensional space is recreated by applying Student t-distribution.

Conclusion: Non-linear, highly overlapping classification set.



Furthermore, Imbalance in dataset and number of categorical features were analysed.

- The classes appears to be evenly distributed and with threshold of 3, there are about 288 categorical features and remaining appears to be numeric or with more catgorical data.

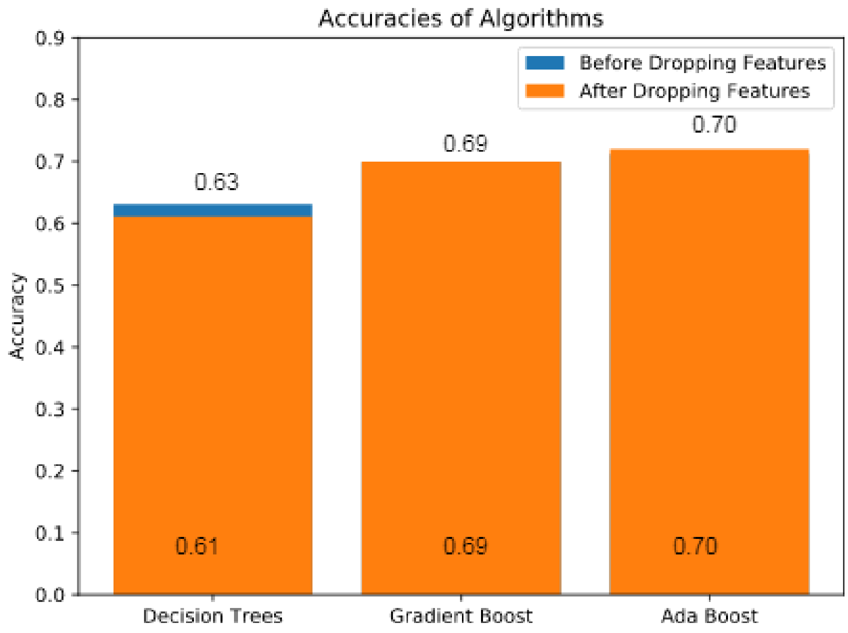
About Algorithm:

Since the data is random and there is correlation with the classes and most of the features appears to be categorical, a **Decision tress** built with importance given to all features helps as prime solution. We do not want to over-fit the model as there is lot of correlation, In order to avoid over-fitting the subset of datasets can be considered for training which also helps in reducing the variance in the model. Or, in other words, ensemble method is can be employed for classification tasks. The variants of these are **Gradient Boost** and **Ada Boost algorithms**

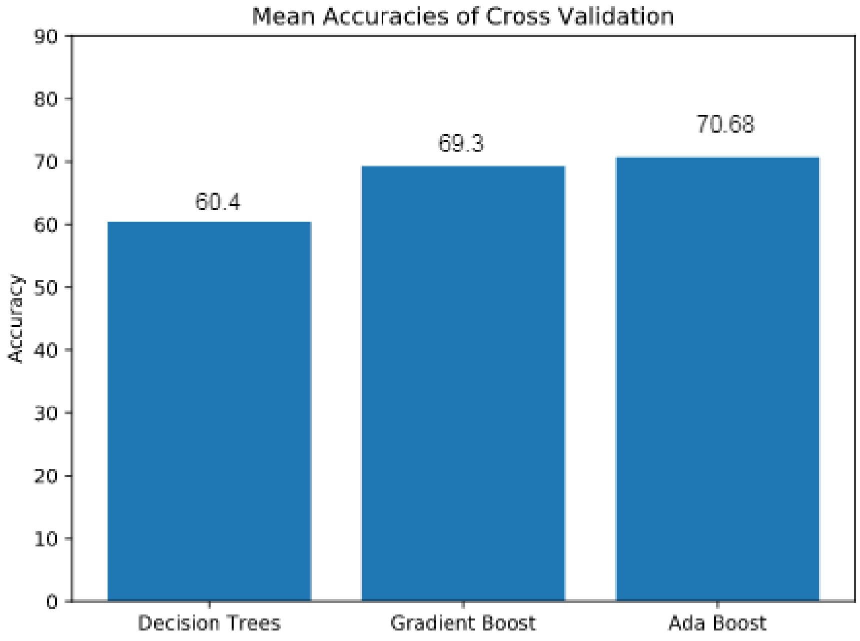
Feature Selection with features accounting 5% only is selected and analysis w.r.t execution time is done as the **evaluation**. **Validation** is done w.r.t different algorithms and cross validation is also selected.

The accuracies of Ada Boost appears to be good and mean cross validation based classification also shows the same results. However, the drop in the execution time is significant in all the algorithms as only 4 out of 295 features are selected for the task by using *feature importance* attribute from used algorithm provided by sklearn.

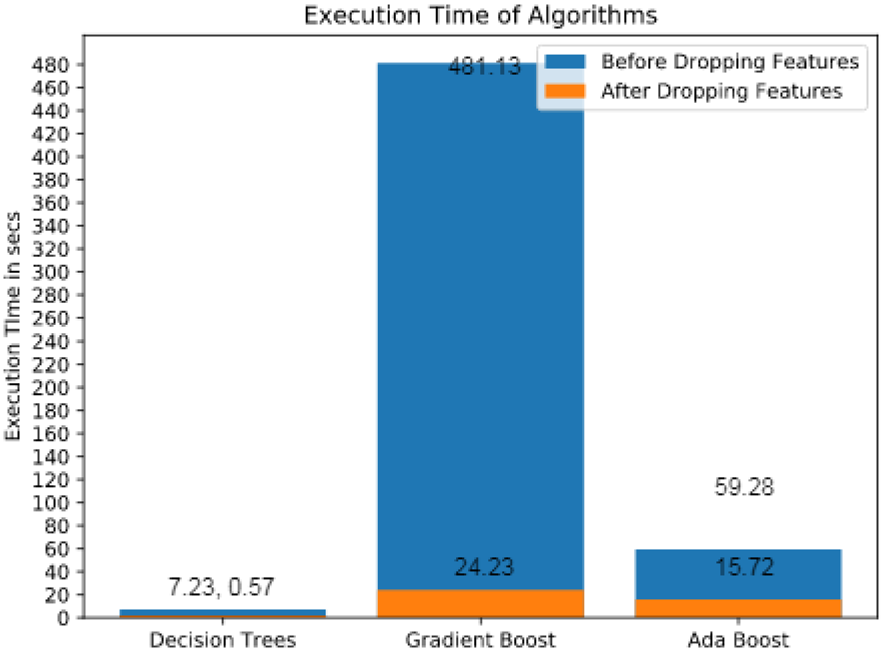
Accuracies of Algorithms used in Classification task:



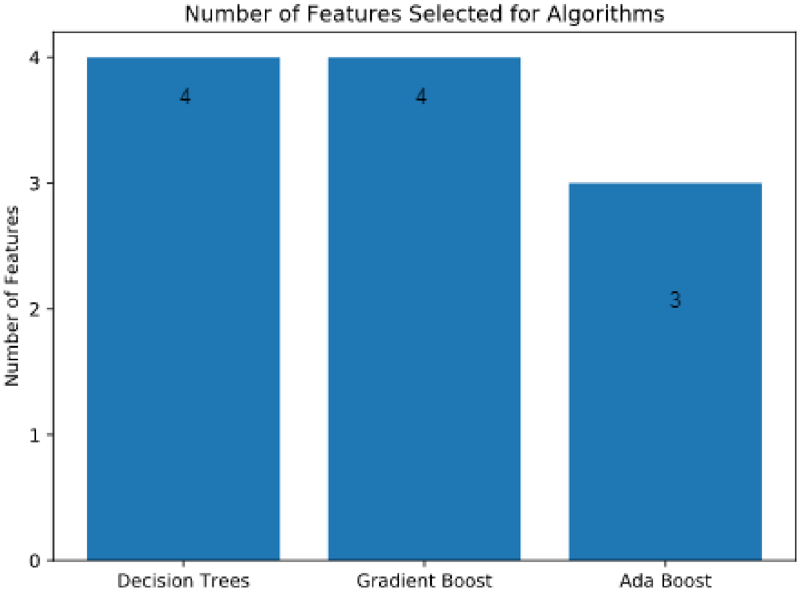
Mean Accuracies of **Cross validation** of different algorithms, after features are dropped:



Execution Time of Classification, before and after dropping the features:



Number of features selected out of 295 features:



Next Steps:

As next steps Artificial Neural Network with only required columns (features with prob 0.0 can be dropped) can be trained.

A usual MLP is what I would go for first to train and test using cross validation technique.