Data Mining -Assignment 2
REPORT

1. Explain your methodology: approach and reason clearly in the report.
2. Visualize skewness of data before and after preprocessing.
4. Drive link for the saved top 3 models and corresponding output .csv files.

(note:-- all the models are stored in the following link of the same GOOGLE DRIVE folder----> https://drive.google.com/drive/folders/1bqbA6FWjF5DvxaTbaZtczS4sOhrloEMc?usp=sharing)

Soln for-> 1,2,4
As per the data given in our training dataset its highly binary class imbalance dataset having near about 30,000 data object classifying to class 0 and 20,00,000 data objects classifying to class '1' thus making it difficult for us to do a normal classification .

Approach and reason->
i) First we removed rows having Nan/null values(if any). Then we had 2 choices under sampling or oversampling to balance our dataset.I chose undersampling as we don't have enough knowledge regarding how to create synthetic dataset for oversampling.

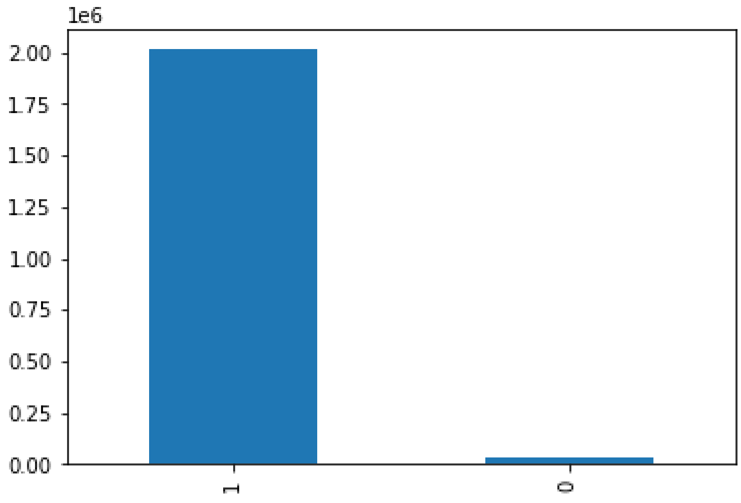
Then we took 30,000 (approx) random data object from class '1' data objects and merged it with 30,000 (approx) data objects of class '0' data objects and with this to make my undersampled dataset of size 60,000 and trained on it and then predicted on behalf of that model.

Same thing i did for 200 times and finally we took the mode value of my predicted test data . Here my base learner were Decision trees.

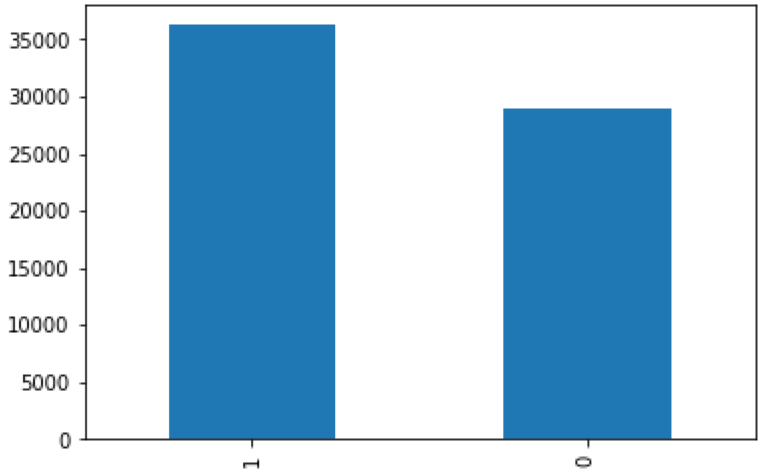
Link for given approach:-
https://colab.research.google.com/drive/19aJe1AOpKm8zFTjoNTK8VaM7ox4TdGXL?usp=sharing

CSV->
https://drive.google.com/file/d/1n3LBVe0krM39_9Ek2aX7mBySmB2XCsUK/view?usp=sharing

Visualization of data before Pre processing



Visualization of data after Pre processing



ii)

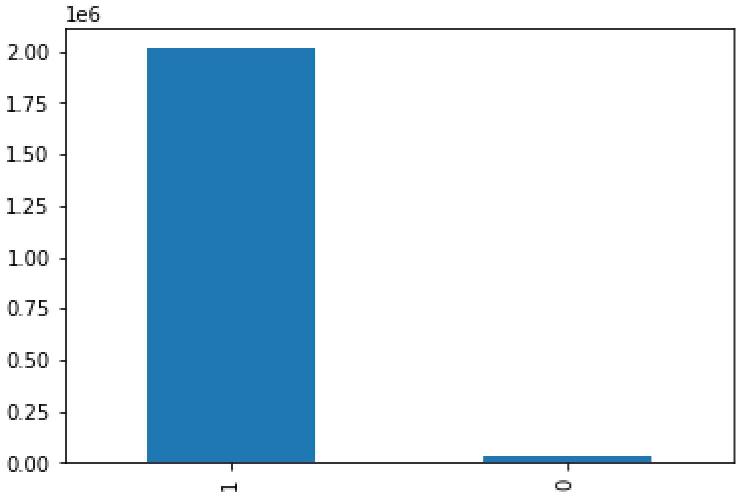
Here we have trained my model on undersampled dataset of near about 60,000 and didn't go for that 200 epochs as already my adaBoost has N_estimmtors of 300 of max_depth=3 each. Base classifier is taken as small decision trees of depth 3 as mentioned.

AdaBoost can be **used** to **boost** the performance of any machine learning algorithm. It is best **used** with weak learners.

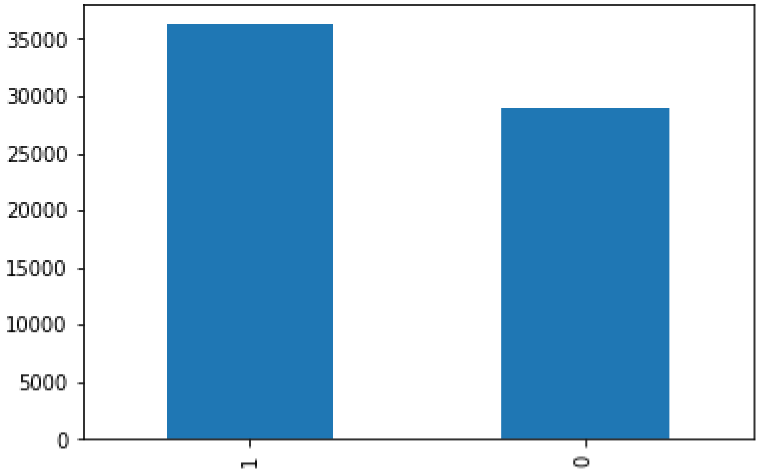
Link for given approach:-
https://colab.research.google.com/drive/1e-TlnzNLif4-DX82zVDejThTcT9I_o0X?usp=sharing

CSV->
https://drive.google.com/file/d/13w2NeWtxD80e6c1A0ziifdr-ahO2AR5v/view?usp=sharing

Visualization of data before Pre processing



Visualization of data after Pre processing



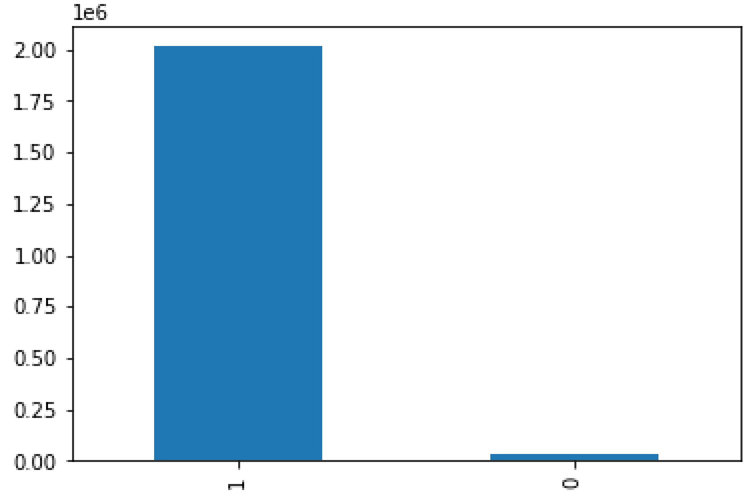
iii)

Finally i tried XG boosting which gave me the best results as of now , we did everything same as approach (i) except the fact that we used XGB classifier and since its boosting, depth of tree is not much deep and no. of epochs on which mode was taken had been reduced to 60 from 200 as already it has 200 n_estimators for each epoch. We used it because it Iis known for its good performance as compared to all other machine learning algorithms.

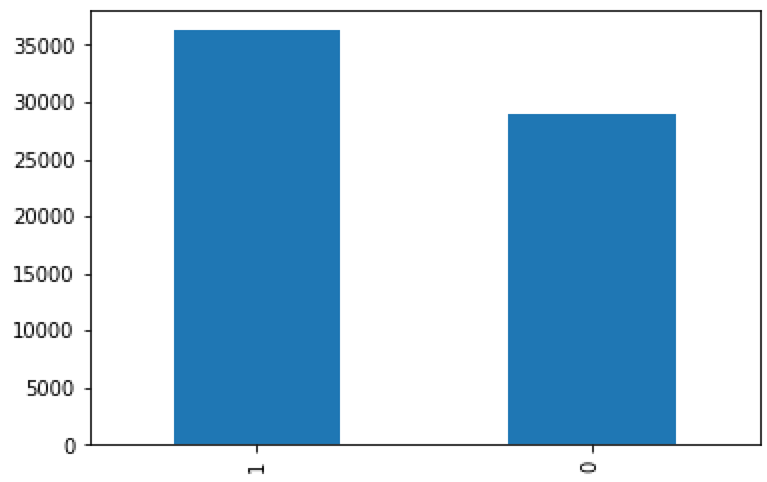
Link for given approach:-
https://colab.research.google.com/drive/1bsmUTEws8nZtkEYvoEVc0AJvM3eB9OuM?usp=sharing

CSV->
https://drive.google.com/file/d/1ppM6XldqMGVf5bZ_WSPQTe5s_CyQgAOQ/view?usp=sharing

Visualization of data before Pre processing



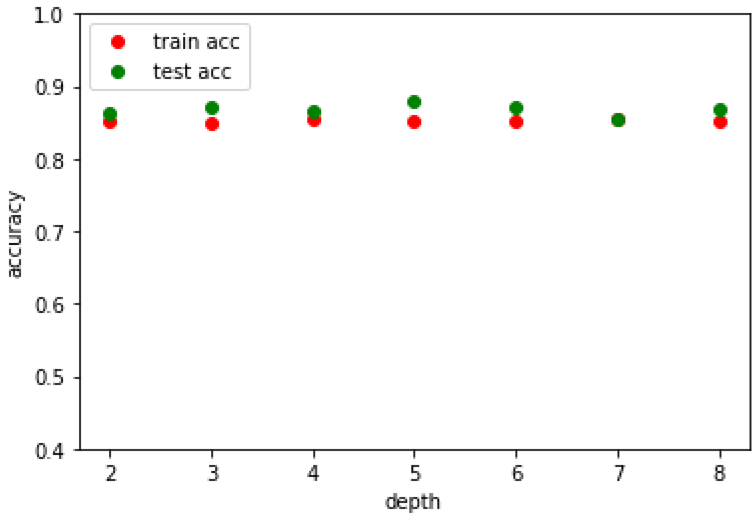
Visualization of data after Pre processing



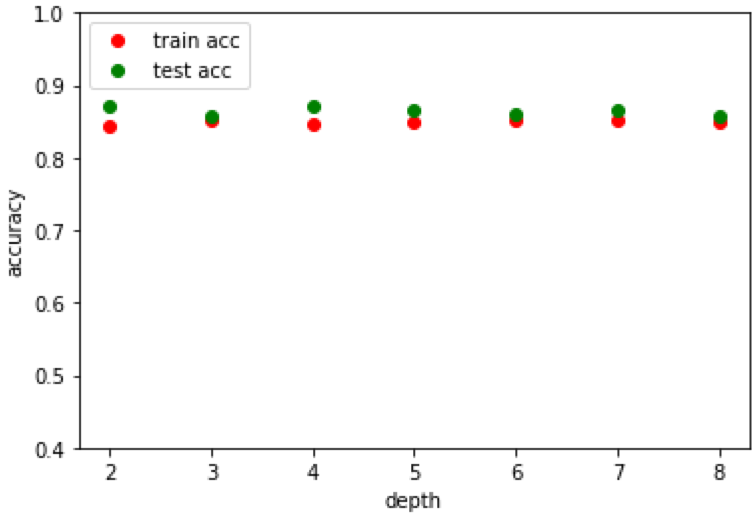
3. Plot Training and Testing accuracy w.r.t. hyperparameters of the model.

Soln->3
Training and testing Accuracy:---

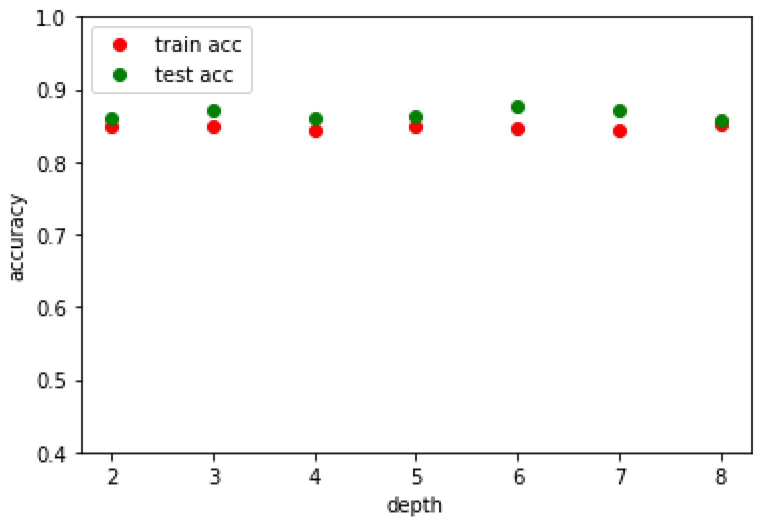
i) For n_estimator=50



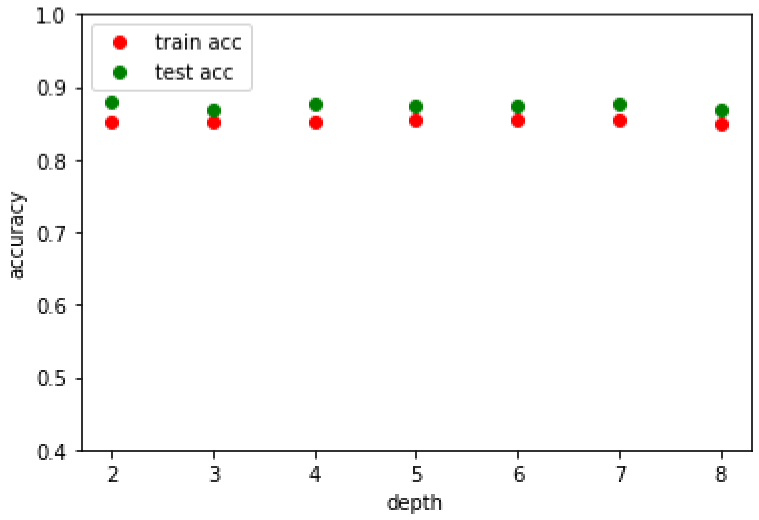
ii) For n_estimator=100



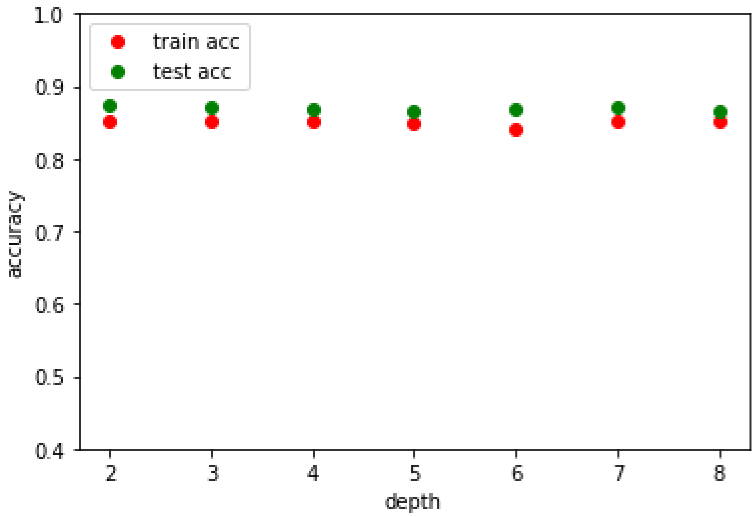
iii) For n_estimator=150



iv) For n_estimator=200



v) For n_estimator=250



For the plotting of train and test i wrote the code separately===> https://colab.research.google.com/drive/1bG8v83avi0YRBt5w_nmoMpMPammwGnMH?usp=sharing

5. Make a section "Learning", which describes your learning in doing this assignment.

soln->5

Learning:-

In this assignment we learned a lot regarding decision trees, ensemble techniques like random forest(bootstrapping & aggregation), wrote random forest from scratch.We also learnt about boosting methods like adaboost,GBDT, XGB and about their parameters and how do we come to low bias low variance model from a low bias and high variance using bagging and from a high bias to a low variance using boosting and most importantly we learnt about how to handle imbalanced dataset which is a very common phenomena in realworld datasets(specially in health disease data, amazon buyer vs visitors data,etc,etc)