Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques

In this paper author is using various machine learning algorithms such as Random Forest, Multilayer Perceptron (MLP), Gradient Boosting, Generalised Linear Model and Feed Forward Neural Network to detect students who may withdraw from course due to interventions (performance or bad grades). Author is using Harvard and OULAD university dataset to train above machine learning algorithms and this dataset contains lots of missing values and contains irrelevant features so author using PCA feature selection algorithm to remove irrelevant features and select only relevant features. From above dataset author identifying 3 different types of students

1. Amotivation: students who withdraw from courses within one week of register
2. Extrinsic: students who perform well in the course
3. Intrinsic: students who underperform in course

After finding above features from dataset author is training all algorithms and then evaluating their performance in terms of ACCURACY, FSCORE, AUC, Sensitivity and Specificity.

Accuracy refers to number of correct prediction from total test data features.

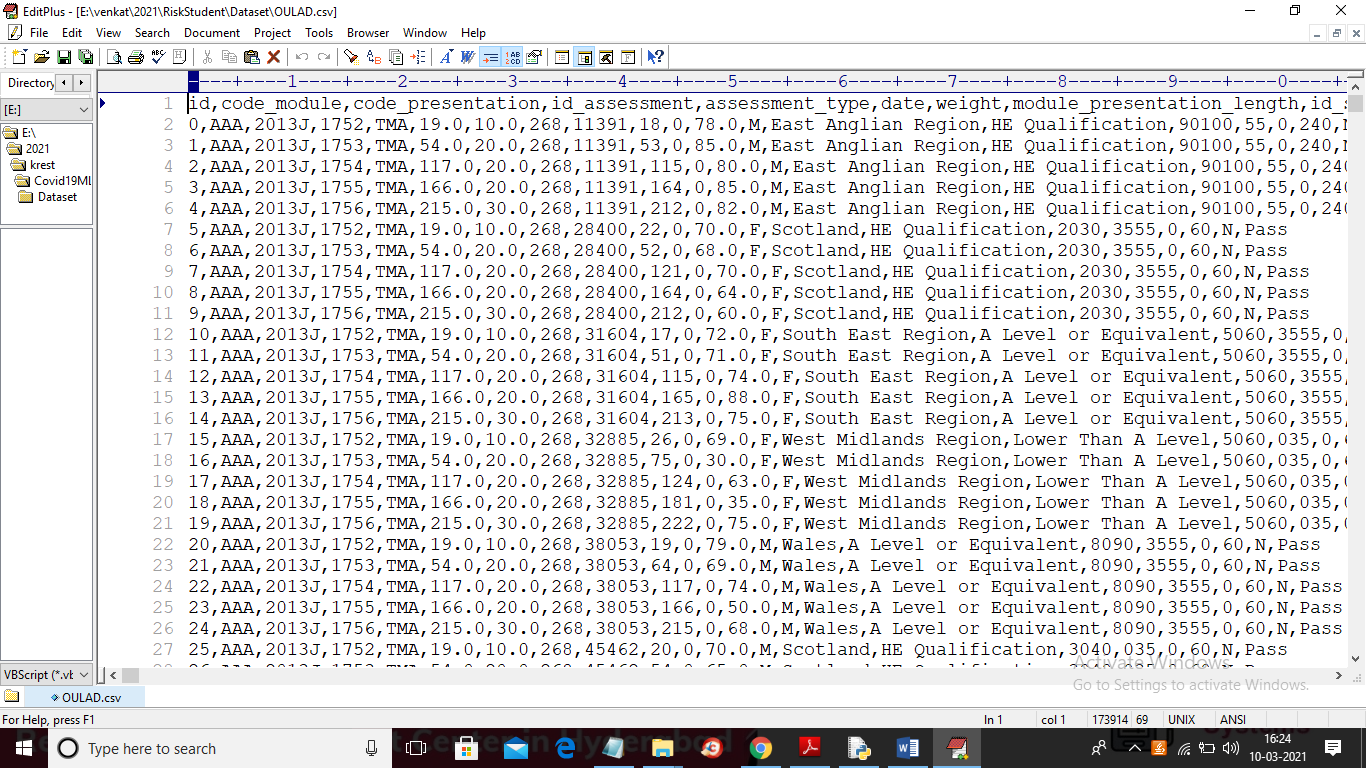
FSCORE: F1 score is the average of precision and recall. Precision is define as number of correct prediction in each class and recall is define as total correct prediction from given data

AUC: AUC – ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with the disease and no disease.

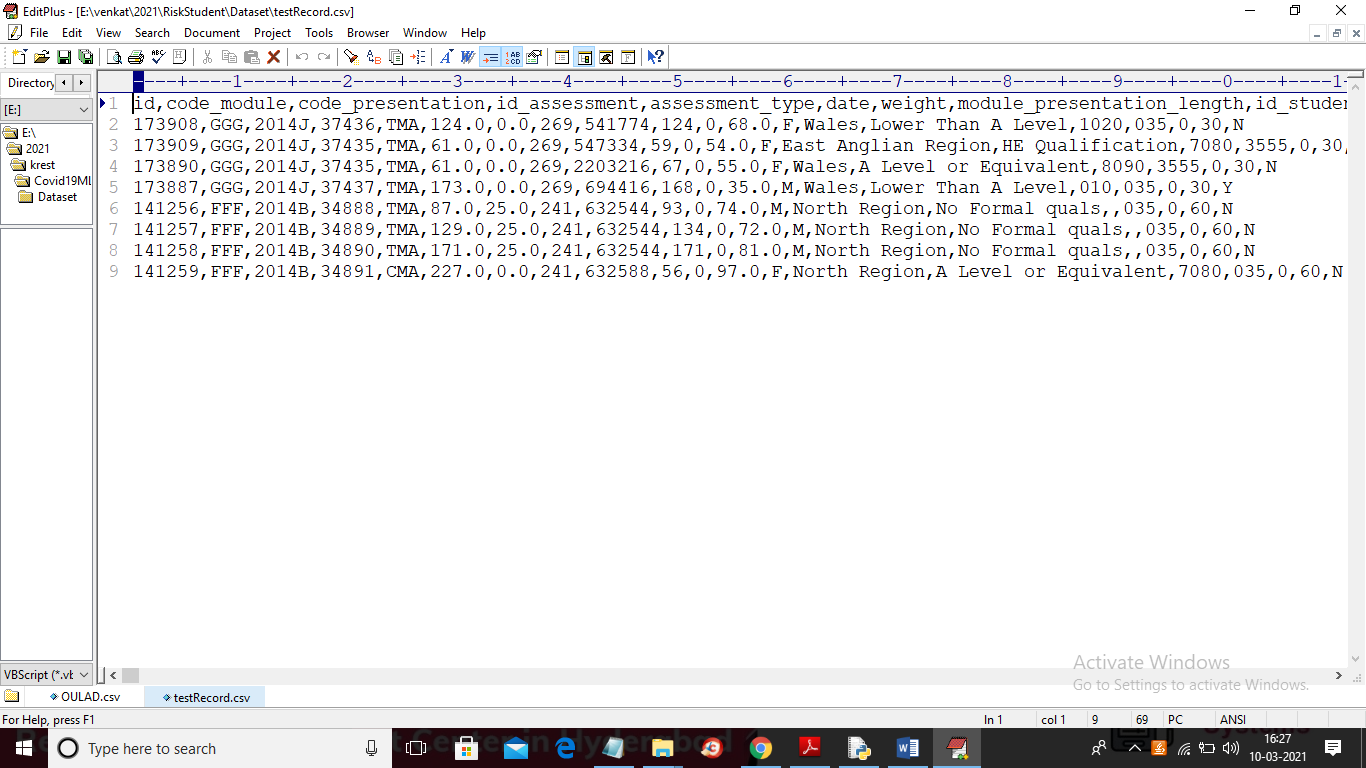
Sensitivity and Specificity: refers to performance of classifier

All algorithms are giving accuracy more than 80% and in rest of the paper author specify about dataset details and dataset processing details such as feature selection and then give brief description on algorithms.

In paper author has used Harvard and OULAD dataset but I don’t find Harvard dataset on internet but OULAD dataset available so I am using OULAD dataset to implement this paper. Below is the OULAD dataset screen shots



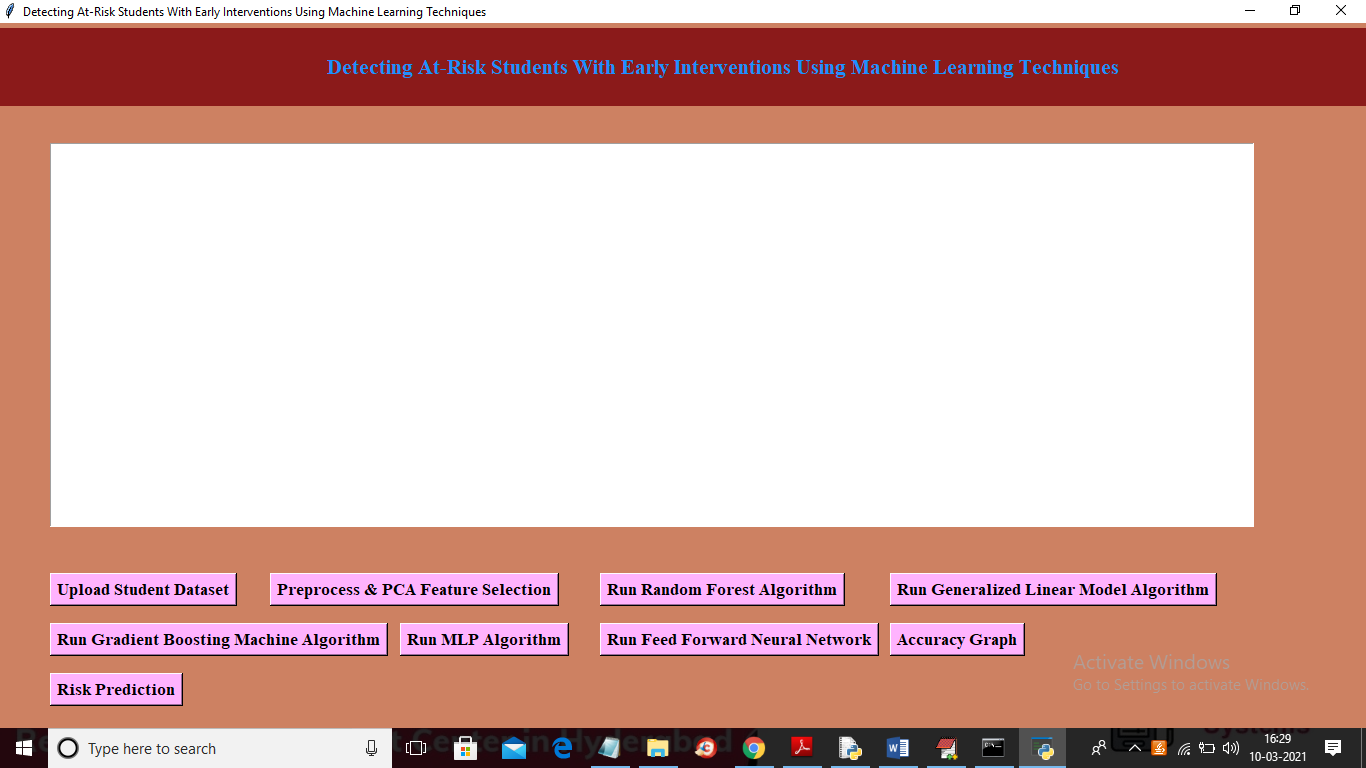
In above dataset we have all details which describe in paper such as code module, gender, number of attempts, exam result etc. In last column we have final result and using this result we can identify whether student is WITHDRWAL or NON-WITHDRAWL. We will use above dataset to train machine learning algorithms and generate model and then this model can be applied on test dataset to predict student result as withdraw or non-withdraw. In below test dataset we don’t have final result column and machine learning model will predict that column by analysing test features



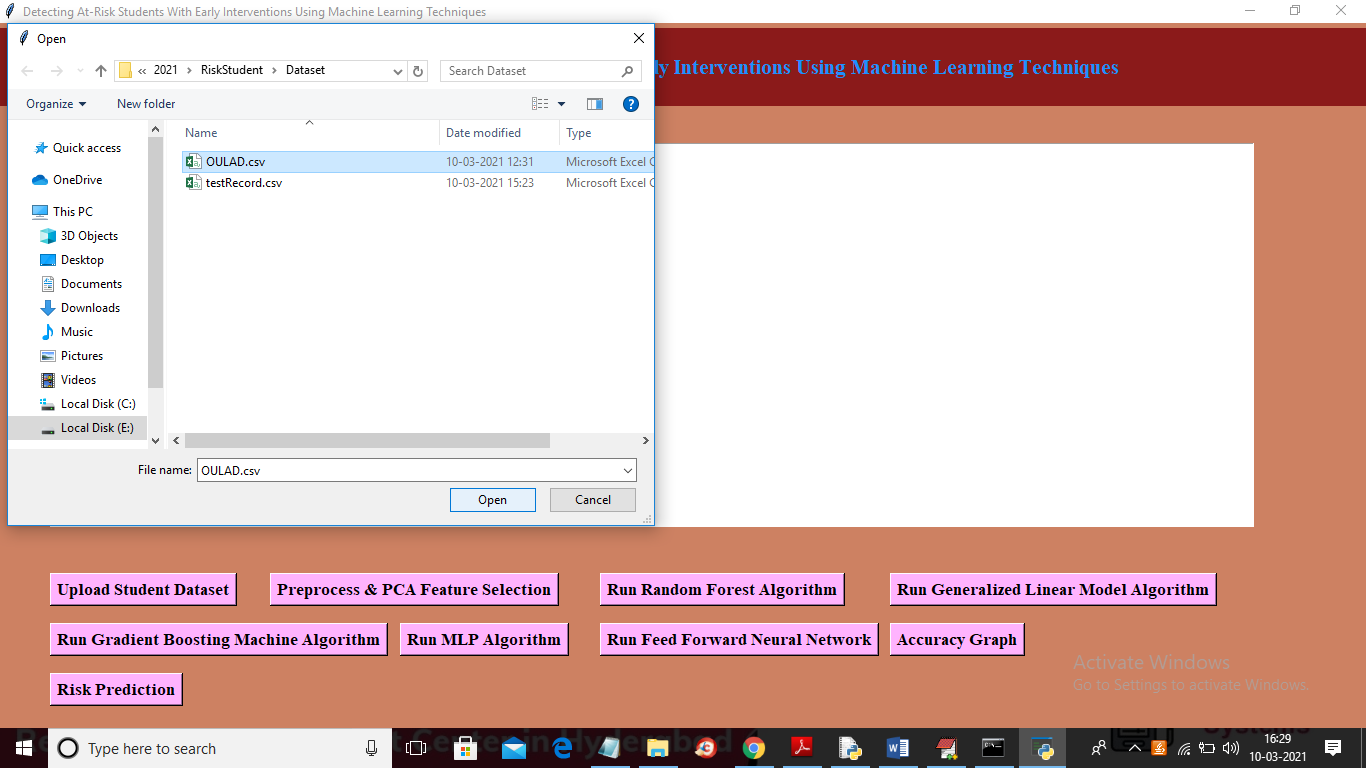
In above screen in last column we don’t have “final result” values and application will predict it by using ML models.

SCREEN SHOTS

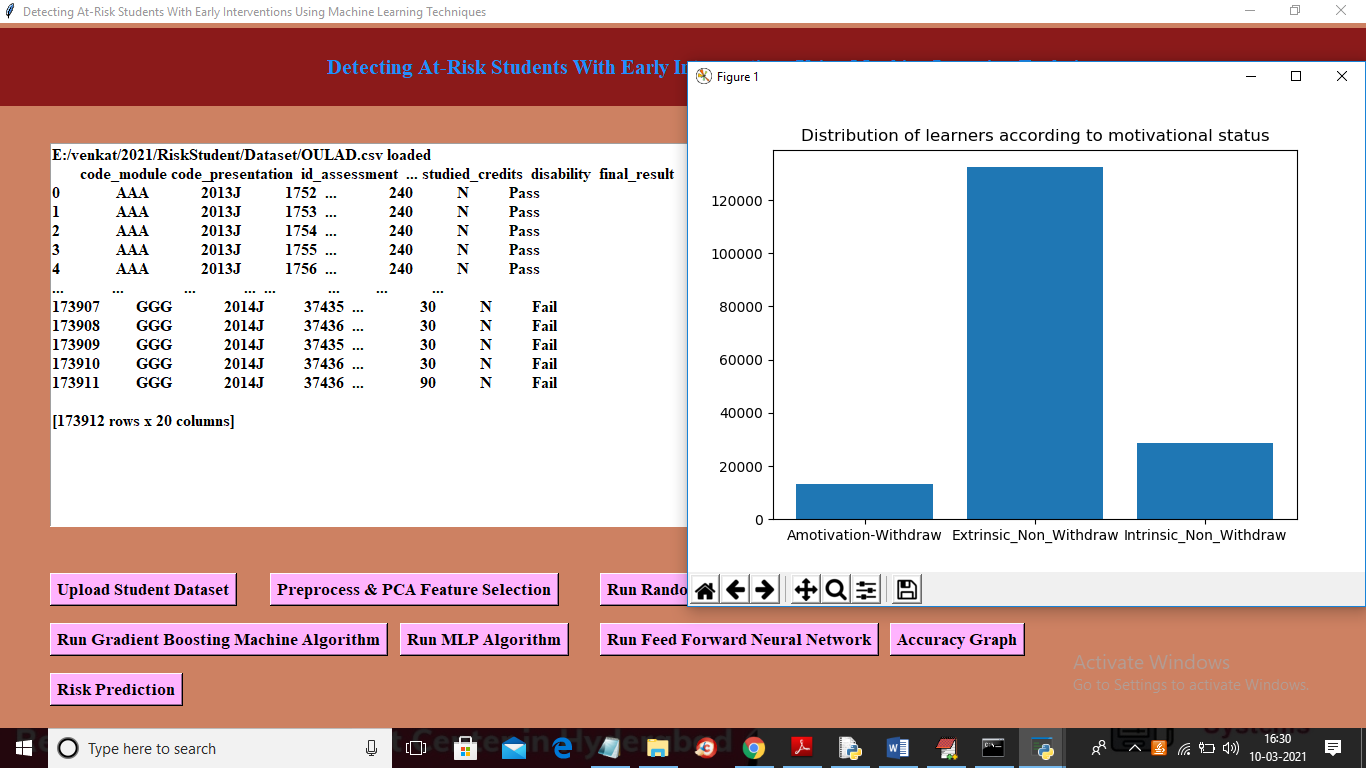
To run project double click on ‘run.bat’ file to get below screen



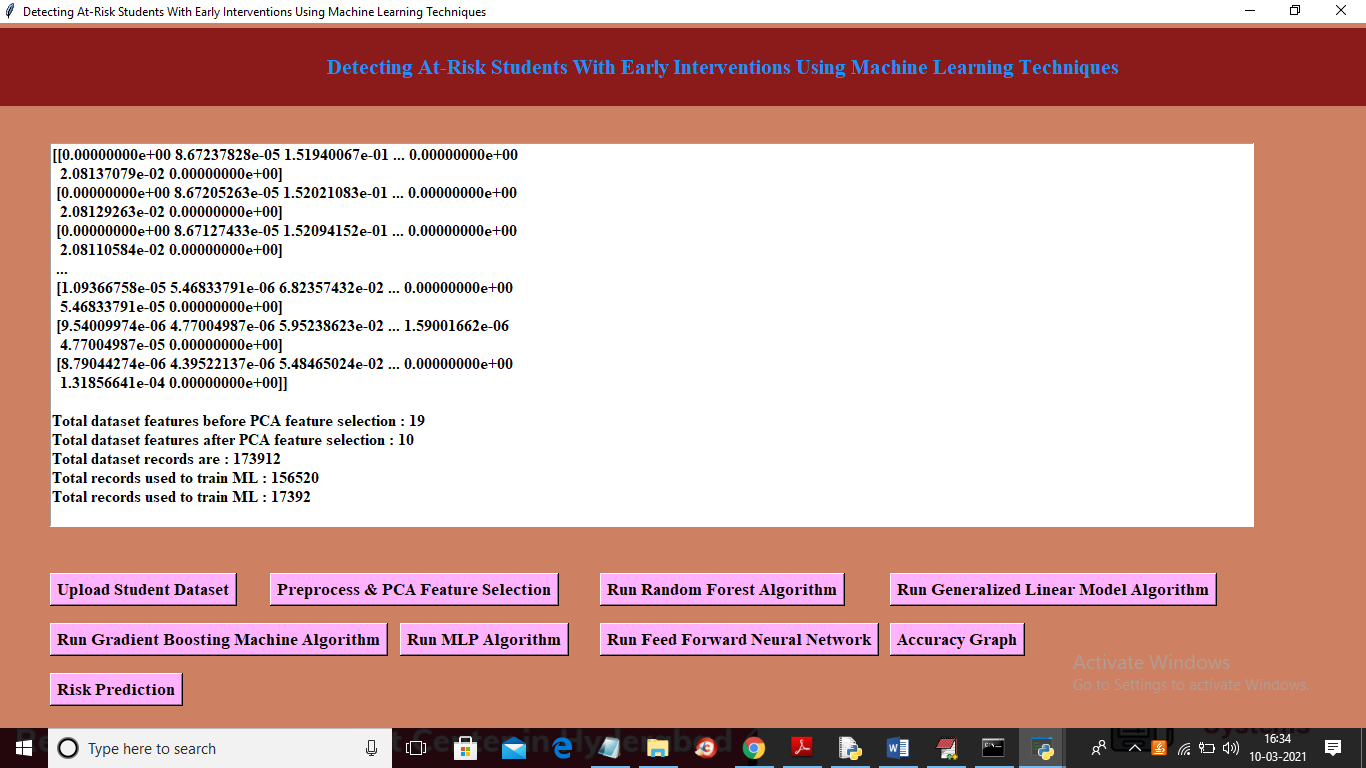
In above screen click on ‘Upload Student Dataset’ button and upload dataset



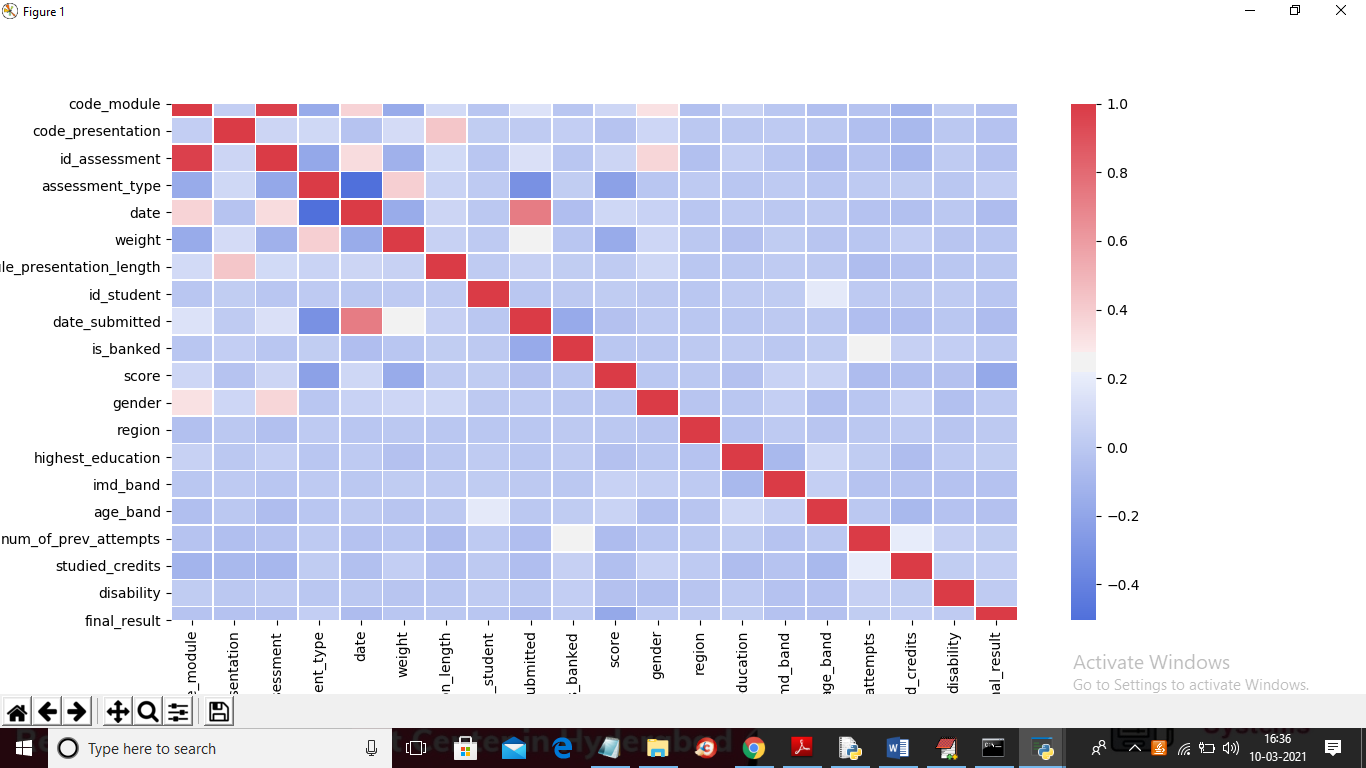
In above screen selecting and uploading ‘OULAD.csv’ dataset and then click on ‘Open’ button to load dataset and to get below screen



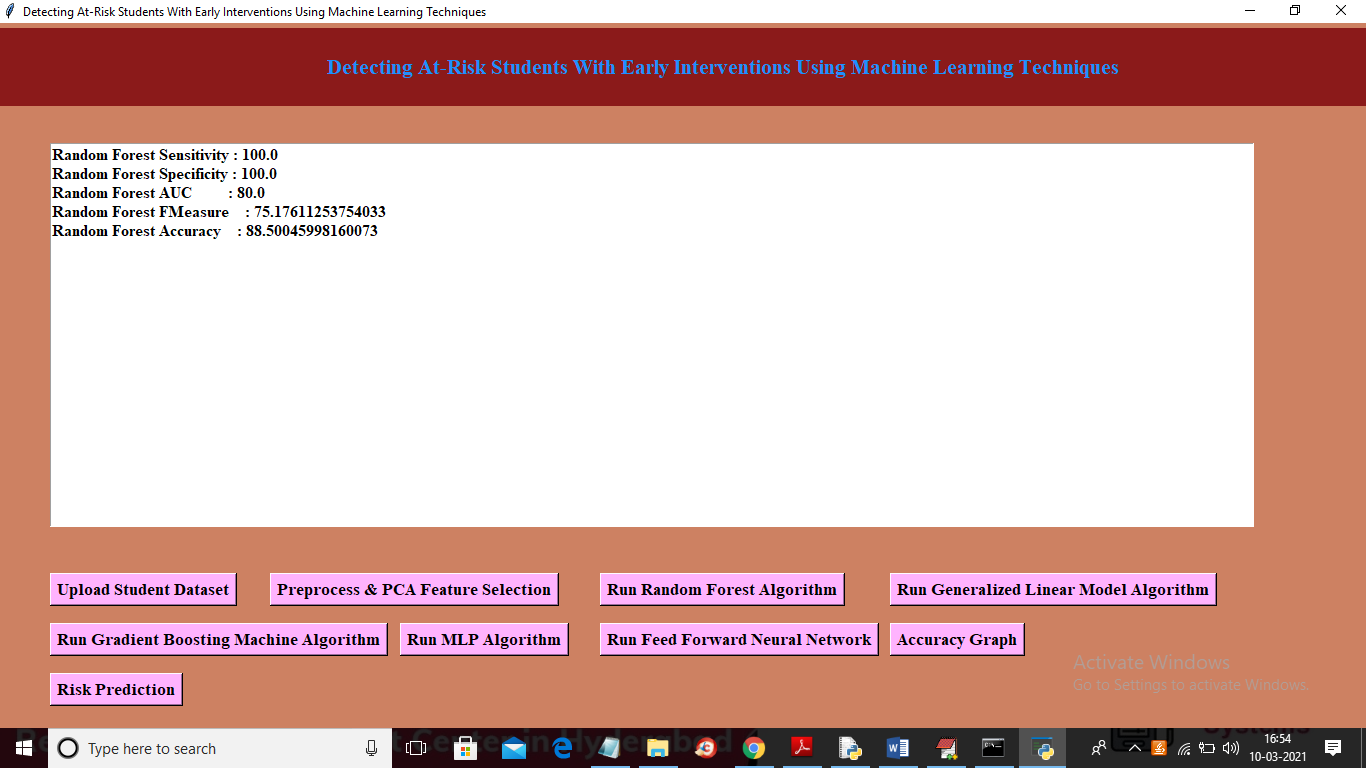
In above screen dataset loaded and in dataset we can see lots of non-numeric values are there and machine learning model will not accept those values so we need to process it to convert to numeric format and in above screen from dataset we found 3 different users such as AMOTIVATION, EXTRINSIC NON-WIITHDRAW and INTRINSIC NON WITHDRAW. In above graph x-axis represents student type and y-axis represents count of that type student. Now close above graph and then click on ‘Preprocess & PCA Feature Selection’ button to convert non-numeric values to numeric format and then apply PCA algorithm to get below screen



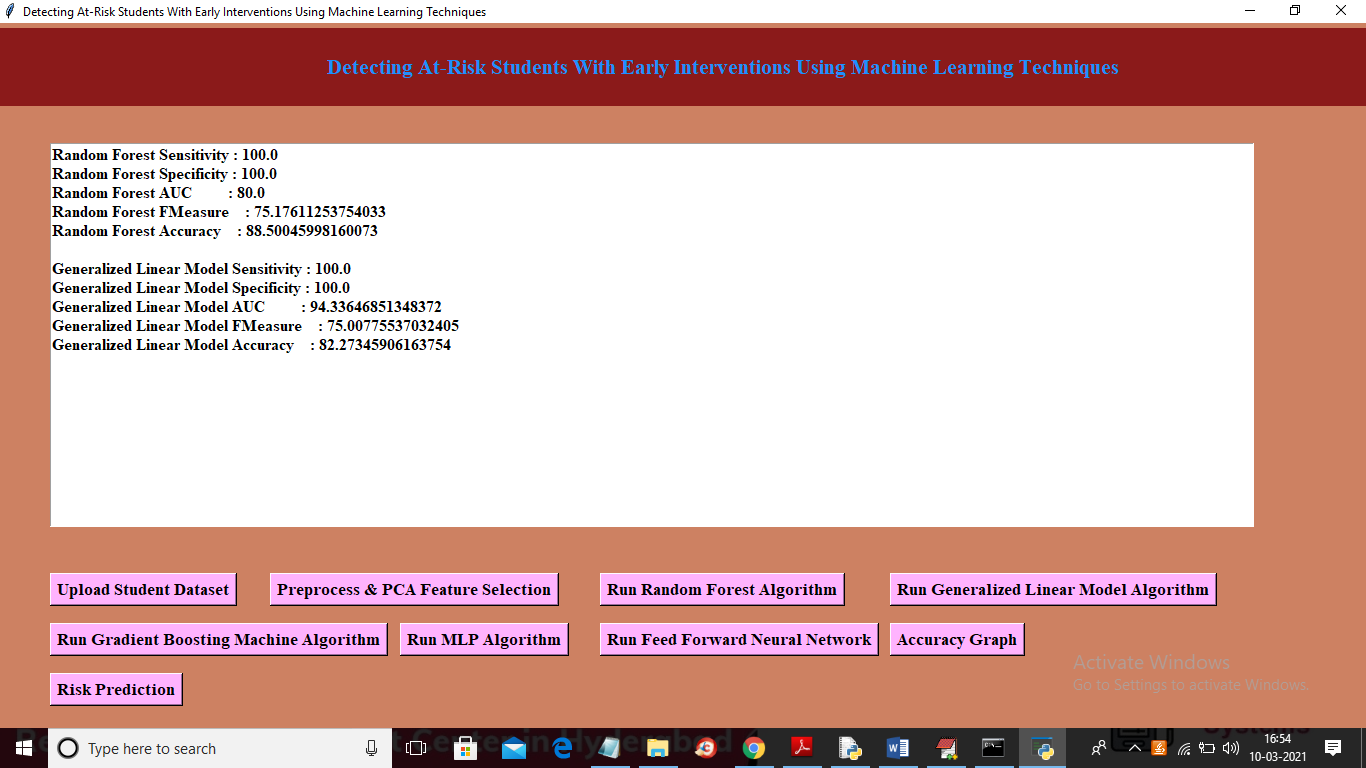
In above screen we can see all non-numeric values are replace with numeric values and then we can see total columns before apply PCA and total columns after applying PCA and PCA will remove irrelevant columns so column size will be reduce with only important features. In above screen we can see total records and records using for training and testing ML algorithms and below graph is the features importance graph



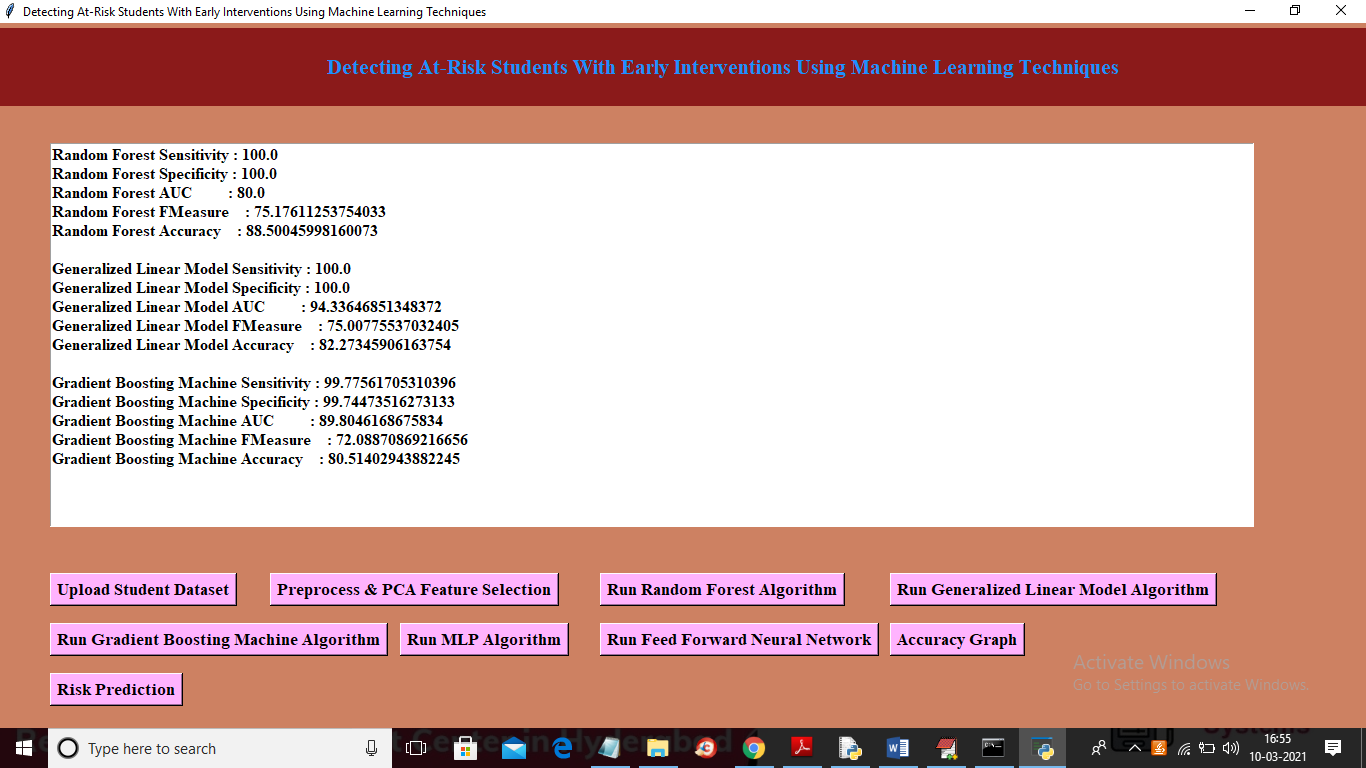
In above graph we can see all columns (features) values from dataset and the features with red colour has probability 1 so only those red colour features are important so PCA will select only those features which has probability 1.0. Now close above graph and both train and test data is ready and now click on ‘Run Random Forest Algorithm’ button to train above dataset with Random forest algorithm



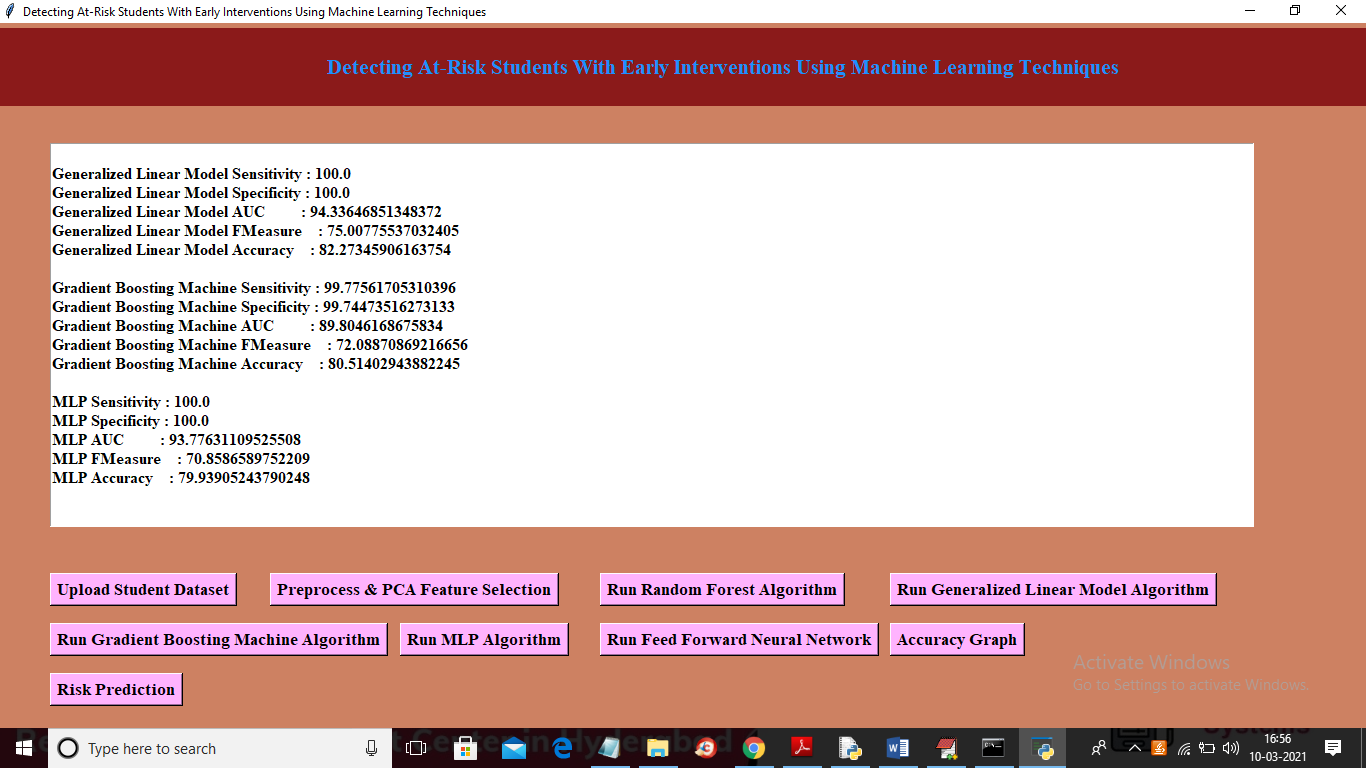
In above screen we got random forest performance details such as accuracy, AUC etc. now click on ‘Run Generalized Linear Model Algorithm’ button to get its performance



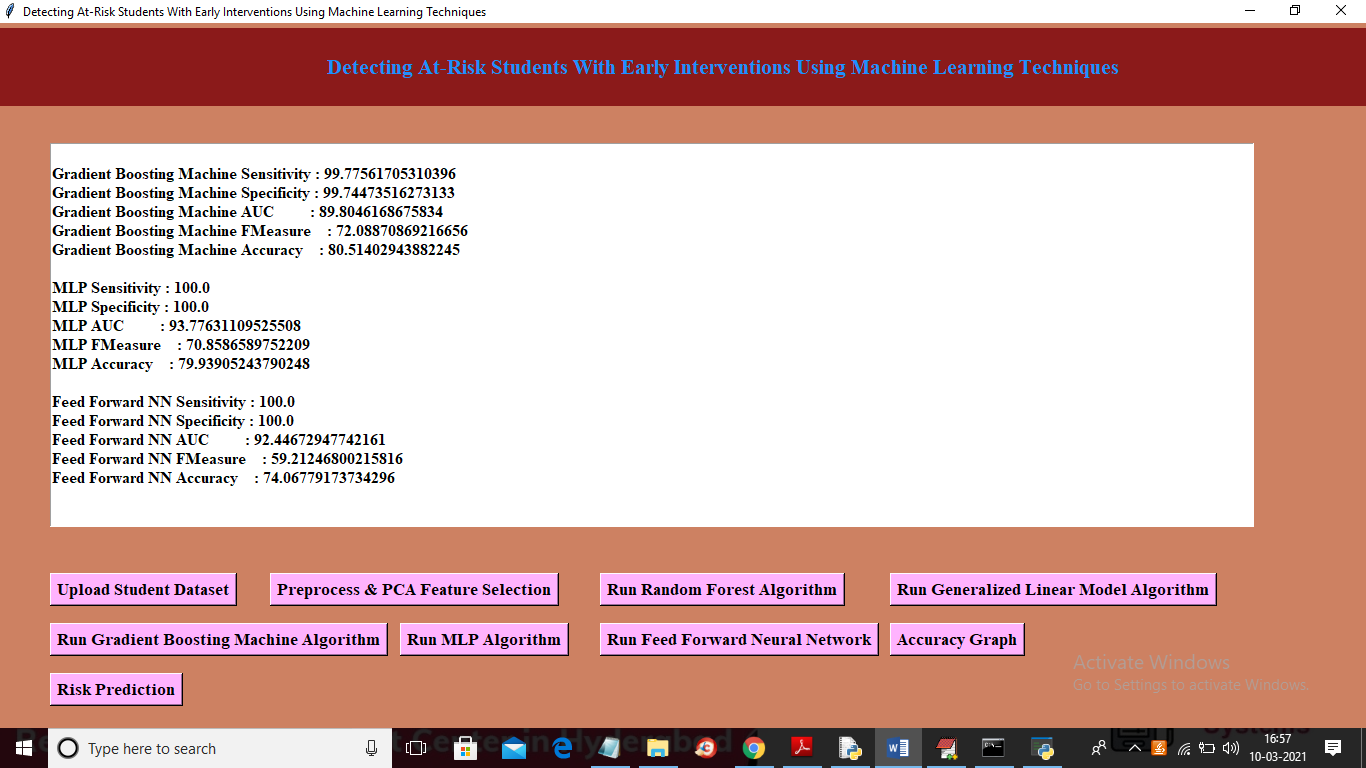
In above screen we got Generalized Linear Model performance and now click on ‘Run Gradient Boosting Machine Algorithm’ button to get its performance



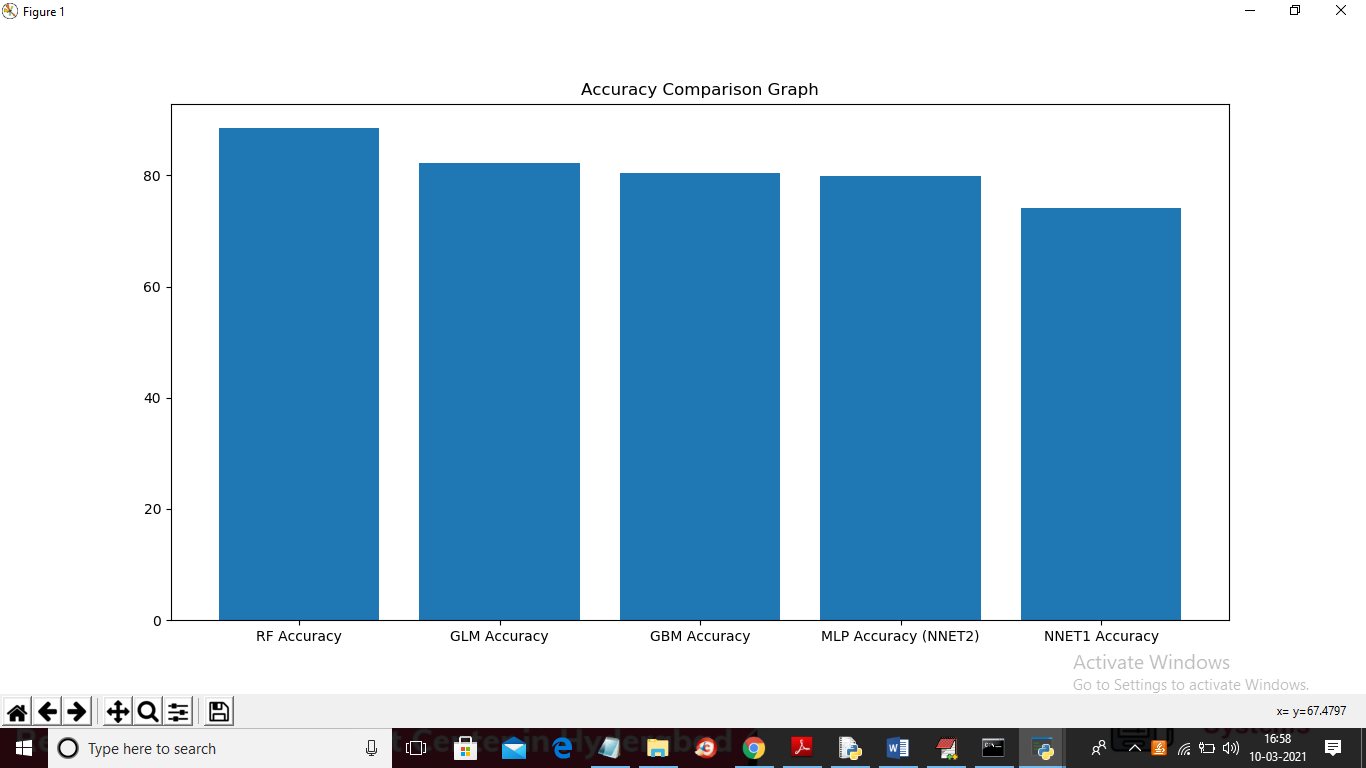
In above screen we got performance of ‘Gradient Boosting Machine’ and now click on ‘Run MLP Algorithm’ button to get its result



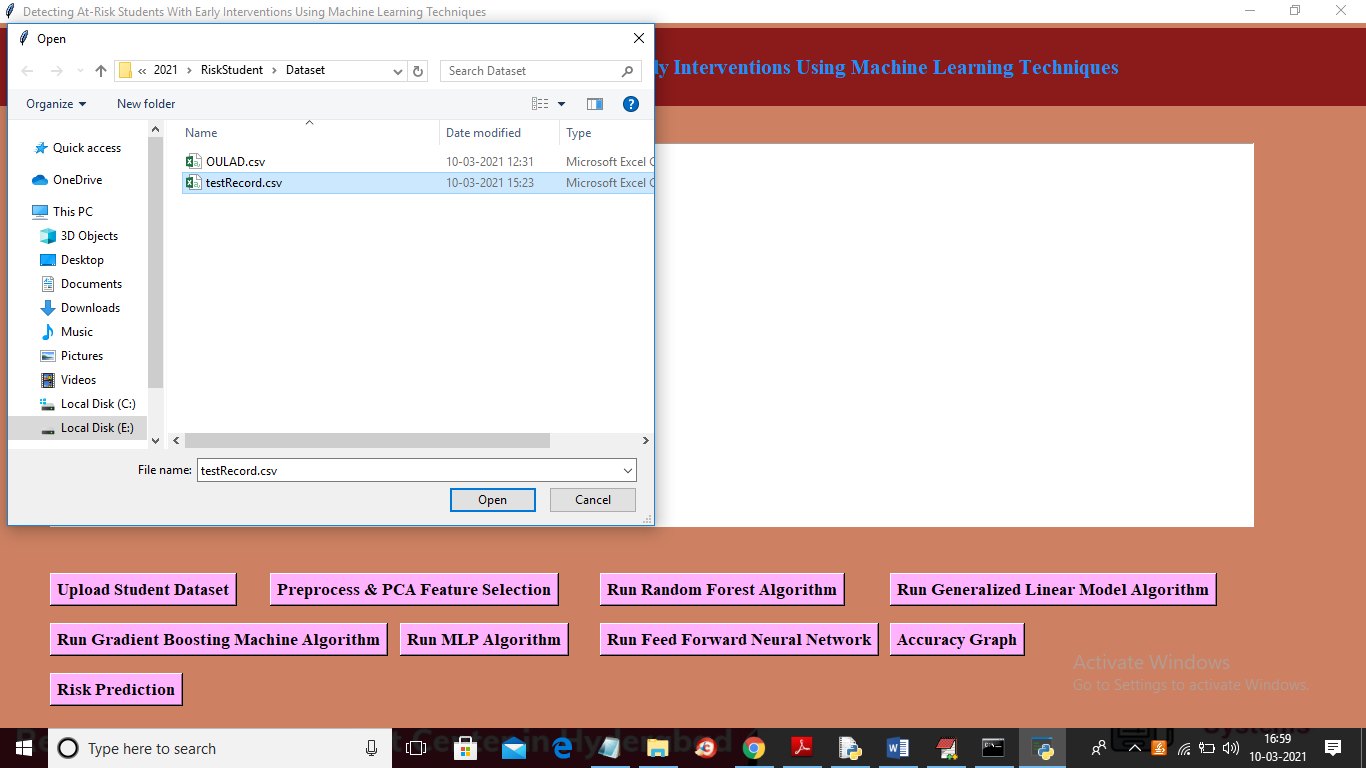
In above screen we got MLP accuracy and other result and now click on ‘Run Feed Forward Neural Network’ button to get its result



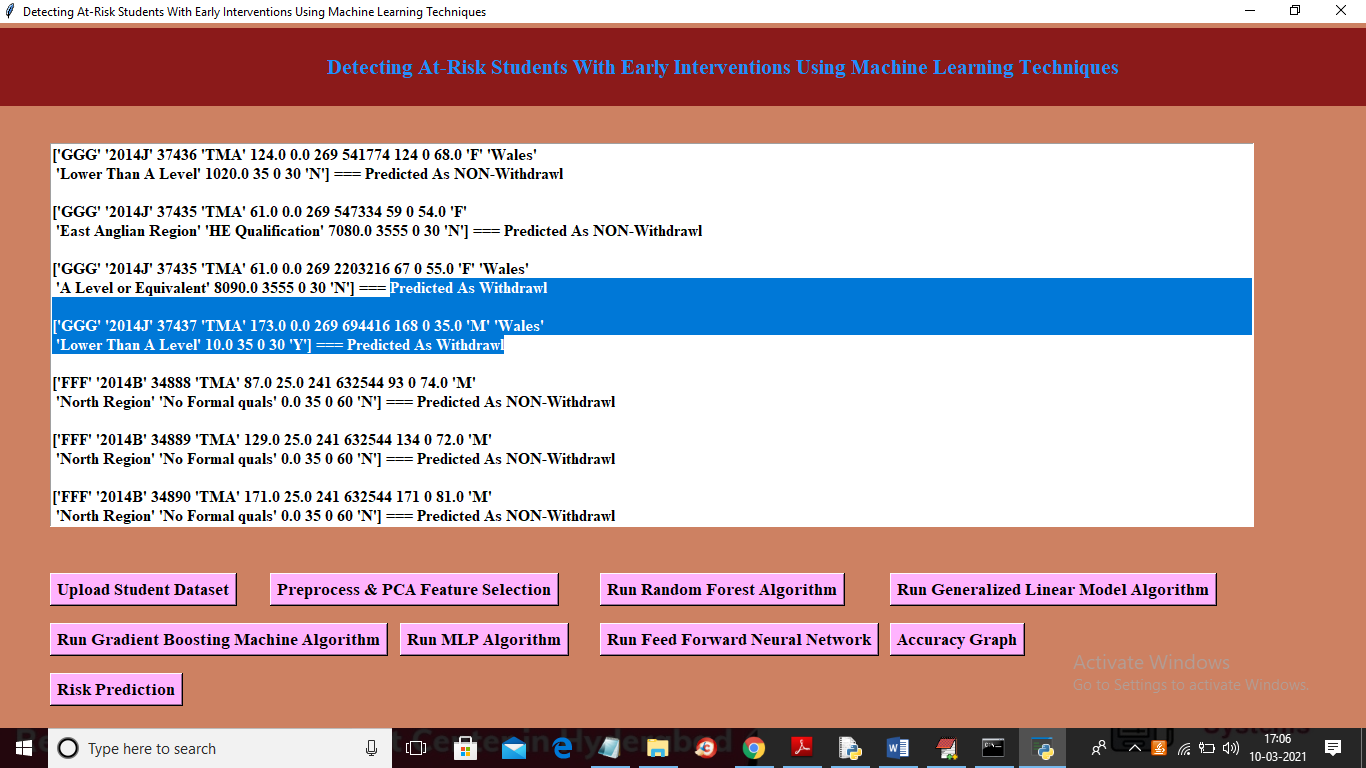
In above screen we got result for ‘Run Feed Forward Neural Network’ algorithm and now click on ‘Accuracy Graph’ button to get below graph



In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and now click on ‘Risk Prediction’ button to upload test dataset



In above screen selecting and uploading ‘testRecord.csv’ file and then click on ‘Open’ button to get below prediction result



In above screen in square bracket we can see student test record values and after dashed symbols we can see predicted result as WITHDRAW or NON-WITHDRAW.

So by using this ML model we can predict risk of withdraw from courses by student just by analysing their past performance features