**Feature extraction for classifying students based on their academic performance**

In this project author is describing concept to predict or classify student performance based on their previous academic performance. Using this paper we will concentrate more on poor performance students by extracting grade features from their past performance records. In this project we are using university dataset which contains record from A to W and we are extracting 4 features from this dataset to classify poor performing students.

1. Taking those records from dataset which has features D and F and consider as failing student and we will assign features values as 0 (Fgr) for such students.
2. Taking those records from dataset which contains dropout students and assign feature value as 1 (Wgr)
3. Taking those records from dataset which has grades lower than expected and assign feature value as 2 (RelF)
4. Taking those records from dataset which has grade value lower than expectation and he is having difficulty in study course and assign value as 3 (RelCF)
5. Rest student we are marking with feature value as 4 which indicate student is performing well.

By using ‘University of Minnesota’ grade dataset we are extracting above features and assign those values as the target or class label for this dataset. After extracting features we are applying 4 machine learning algorithms on this dataset to generate training model, later new student record will be applied on this dataset to classify that student records as good performer or poor performer and we can know the reason of poor performance such as Fgr (indicate as failing student), Wgr (indicate as dropout), ReIF (lower than expected grade) or RelCF (lower than expected grade and having course difficulty).

4 algorithms used in this paper

**SVM Algorithm**: Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

**Random Forest Algorithm:** its an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

**Decision Tree Algorithm:** This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

**Gradient Boosting Algorithm**:Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.

The Python machine learning library, Scikit-Learn, supports different implementations of gradient boosting classifiers, including XGBoost. By using multiple algorithms a single accurate train model will be generated. In all this algorithms Gradient Boosting give better performance.

Dataset samples

**Subject, Course, CRN, CourseTitle, A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, W, Extracted\_Features**

AAS,100,41758,Intro Asian American Studies,10,15,3,1,0,0,1,0,0,0,0,0,0,0,1.89,1

AAS,100,47100,Intro Asian American Studies,7,12,1,5,1,0,0,1,1,0,0,0,0,0,1.68,1

AAS,100,47102,Intro Asian American Studies,6,4,6,3,1,1,0,0,1,0,0,0,0,0,1.61,1

AAS,100,51248,Intro Asian American Studies,9,6,7,3,1,0,2,0,0,1,0,0,0,0,3.61,4

ABE,100,31263,Intro Agric & Biological Engrg,0,30,0,0,11,0,0,1,0,0,0,0,0,0,3.69,4

ABE,223,58927,ABE Principles: Machine Syst,1,28,2,1,6,2,0,0,0,0,0,0,0,0,2.75,4

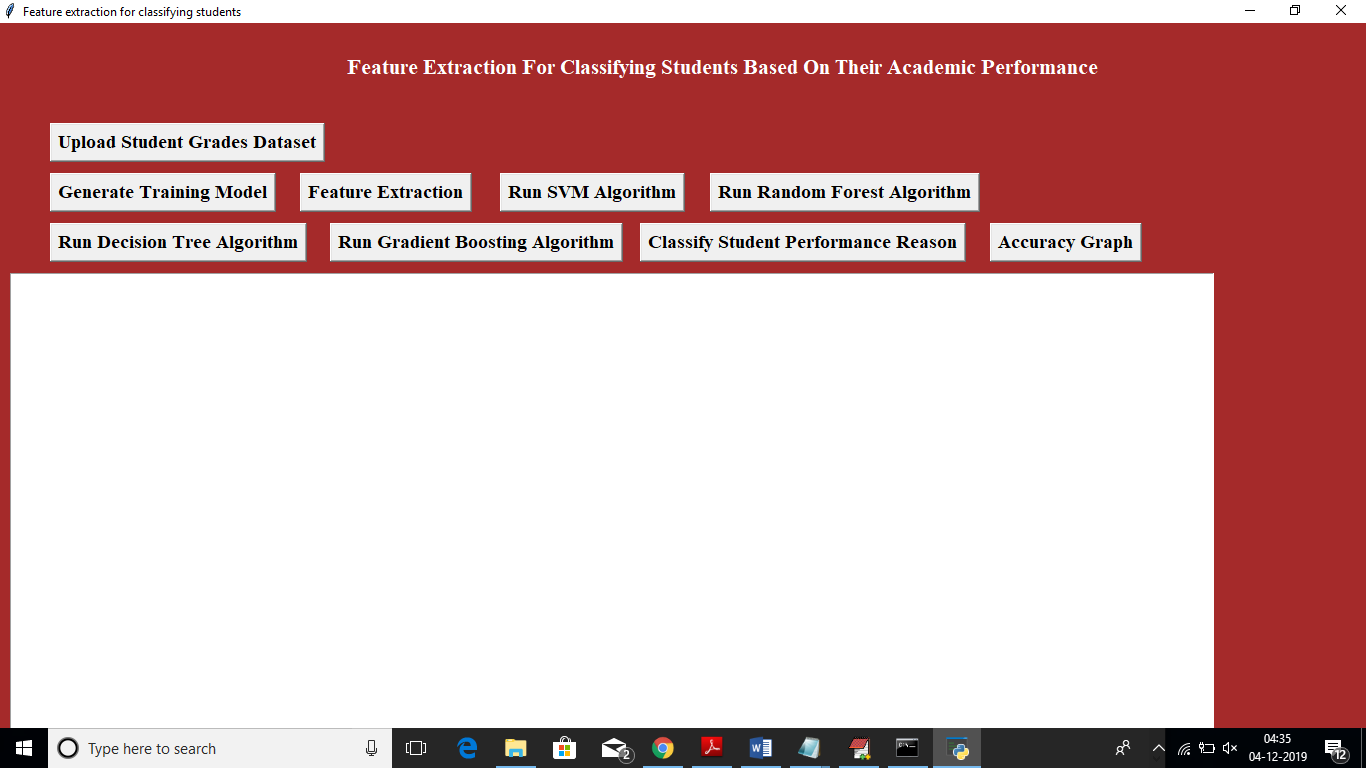
ABE,224,58931,ABE Principles: Soil & Water,1,33,9,3,2,0,0,0,0,0,0,0,1,0,0.78,0

In above dataset all bold names are columns of this dataset and rest are the dataset values and in this dataset last column is the extracted features such as 0, 1, 2, 3 and 4 and about this values I already explain in 4 features extraction in 1 page of this document.

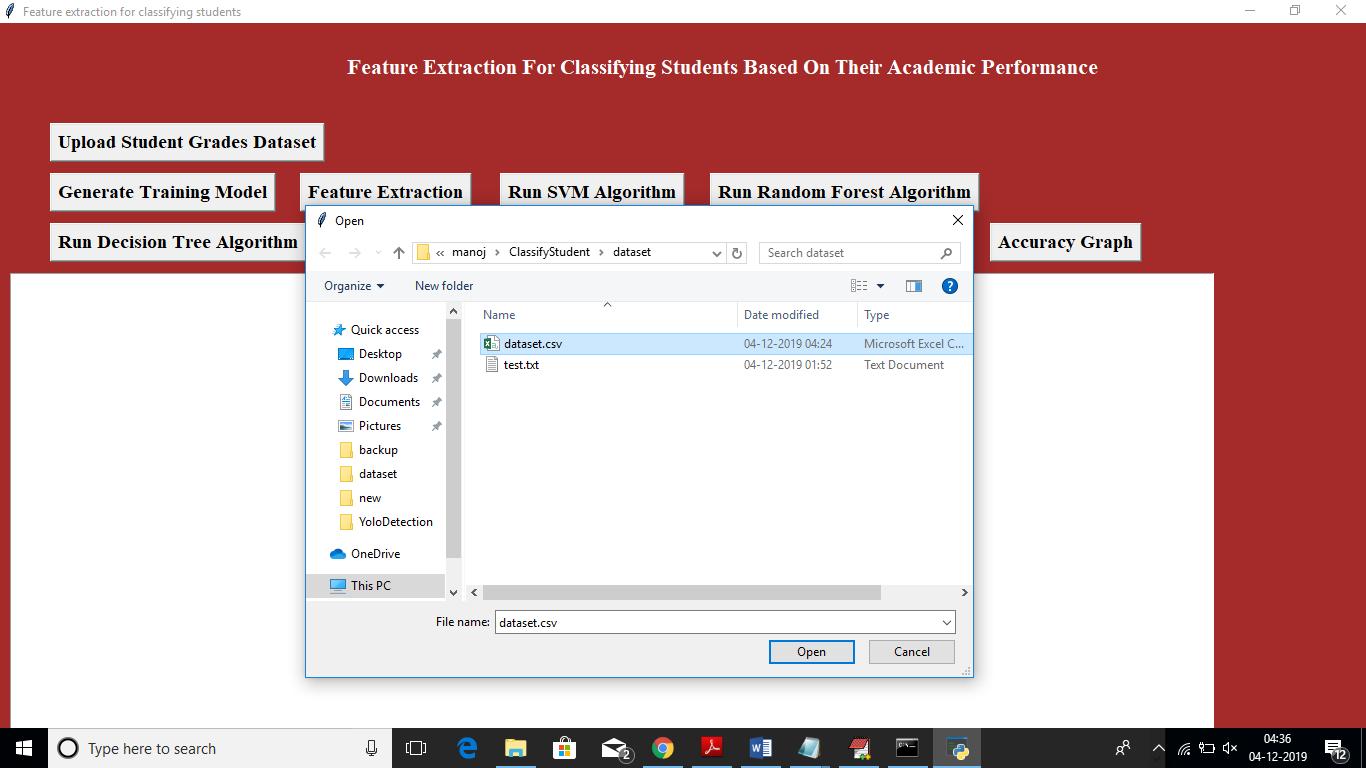
You can find this dataset inside ‘dataset’ folder.

Screen shots

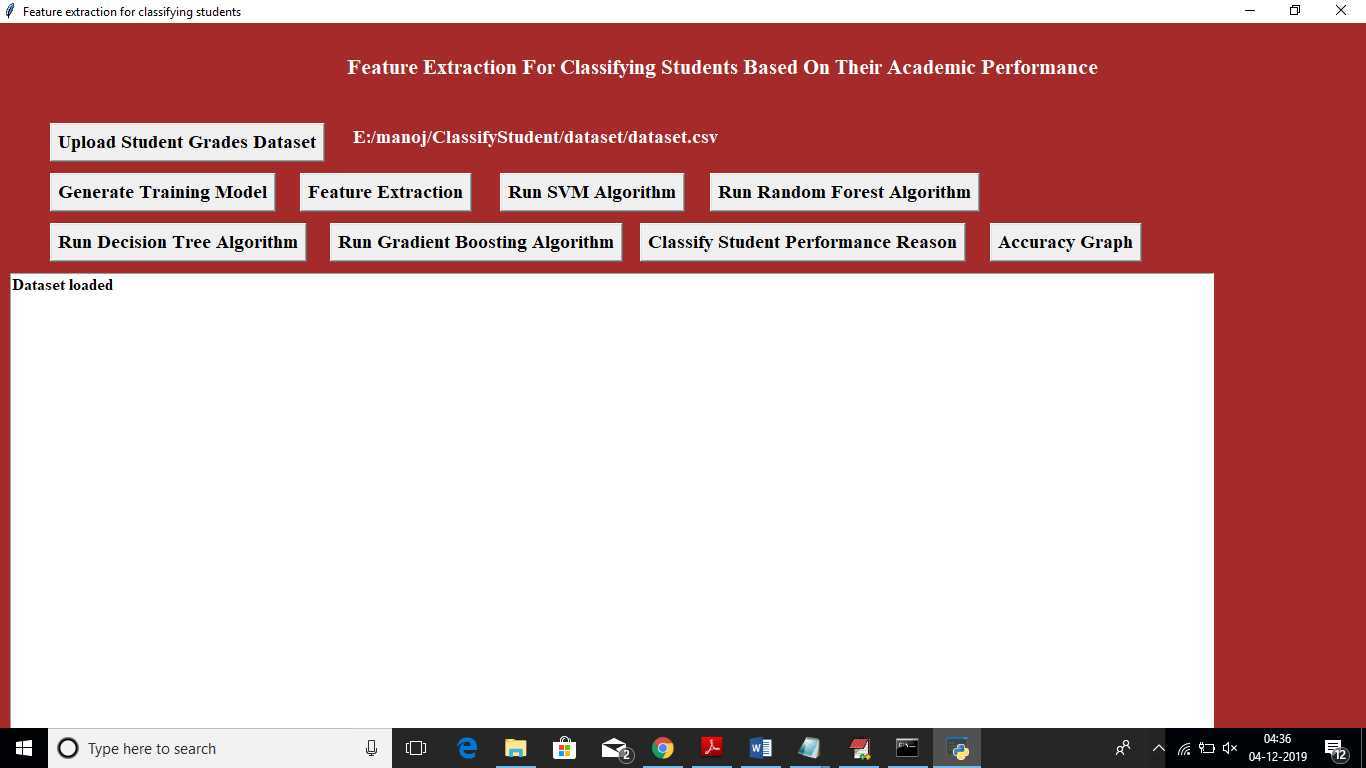
To run this project double click on ‘run.bat’ to get below screen



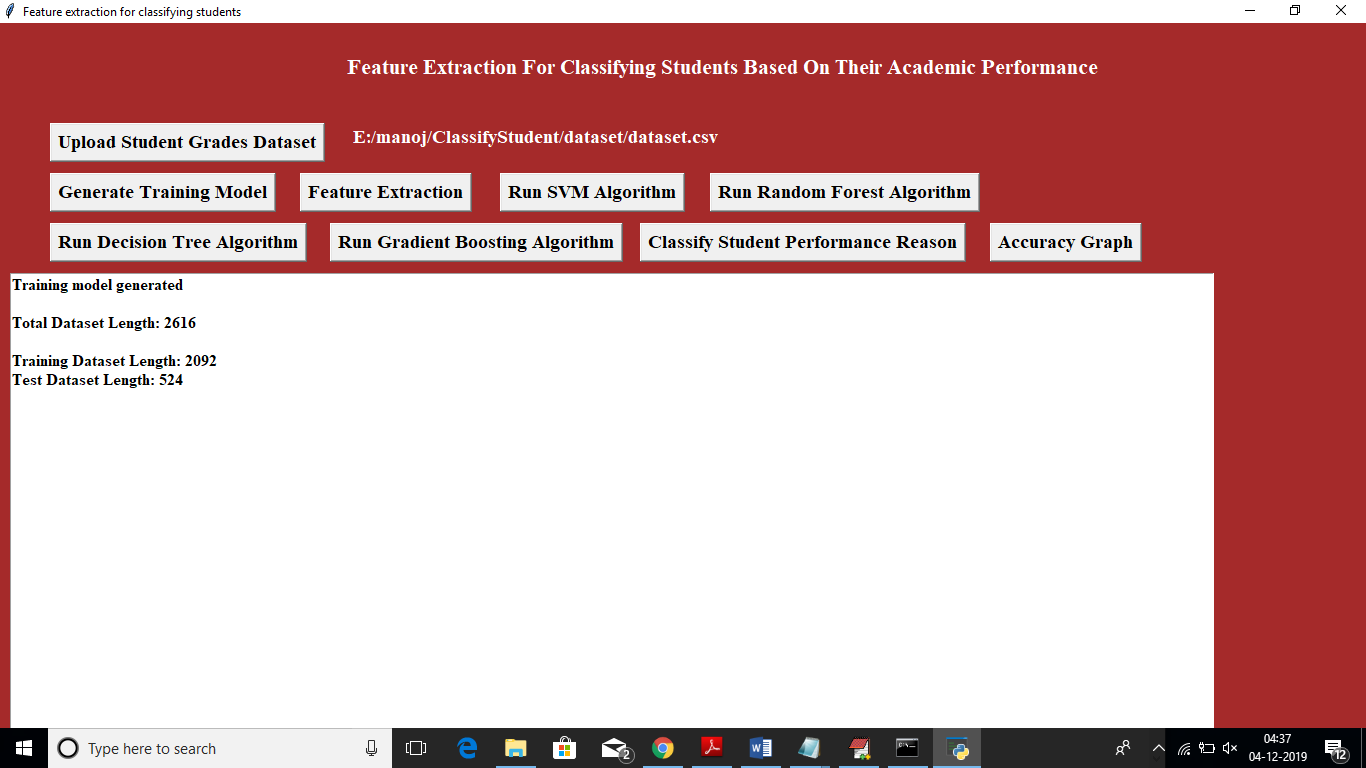
In above screen click on ‘Upload Student Grades Dataset’ button to upload dataset



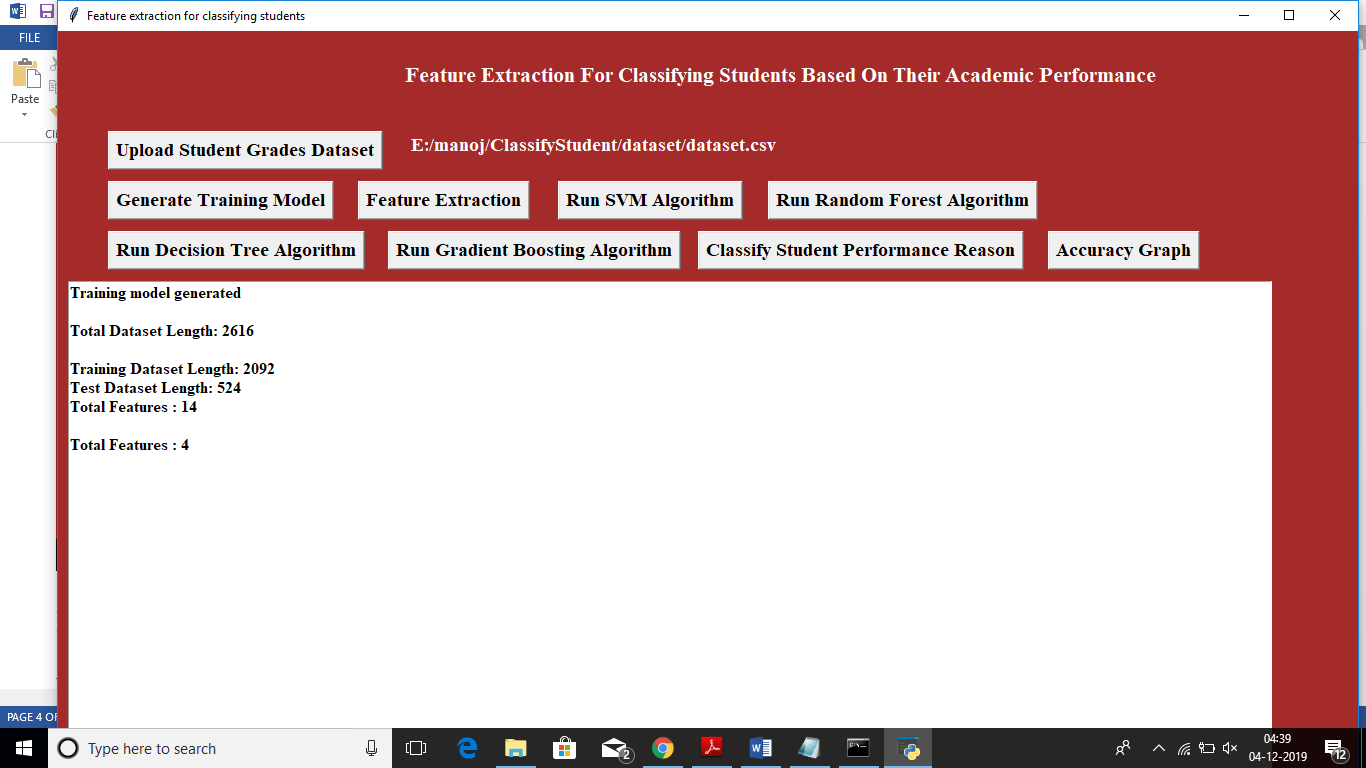
After uploading dataset will get below screen



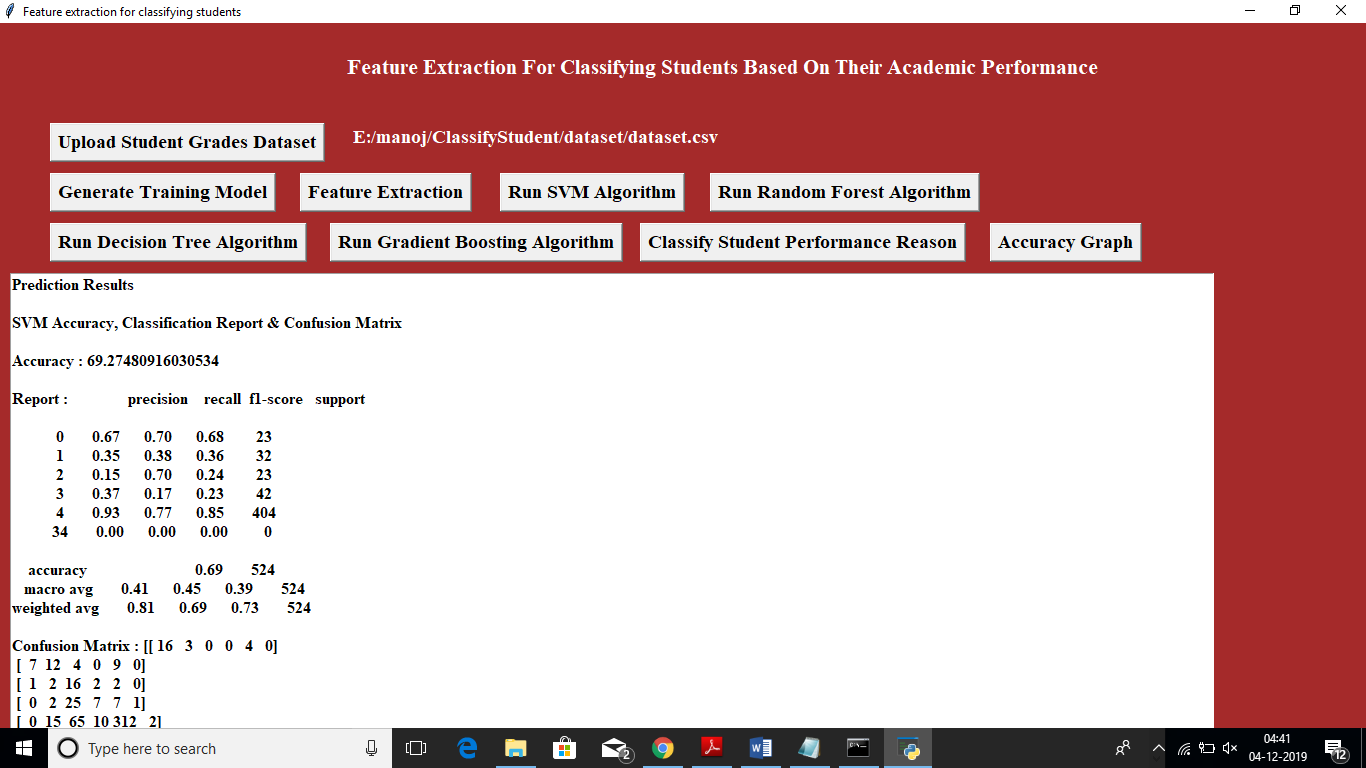
Now click on ‘Generate Training Model’ button to read dataset and build array for training purpose



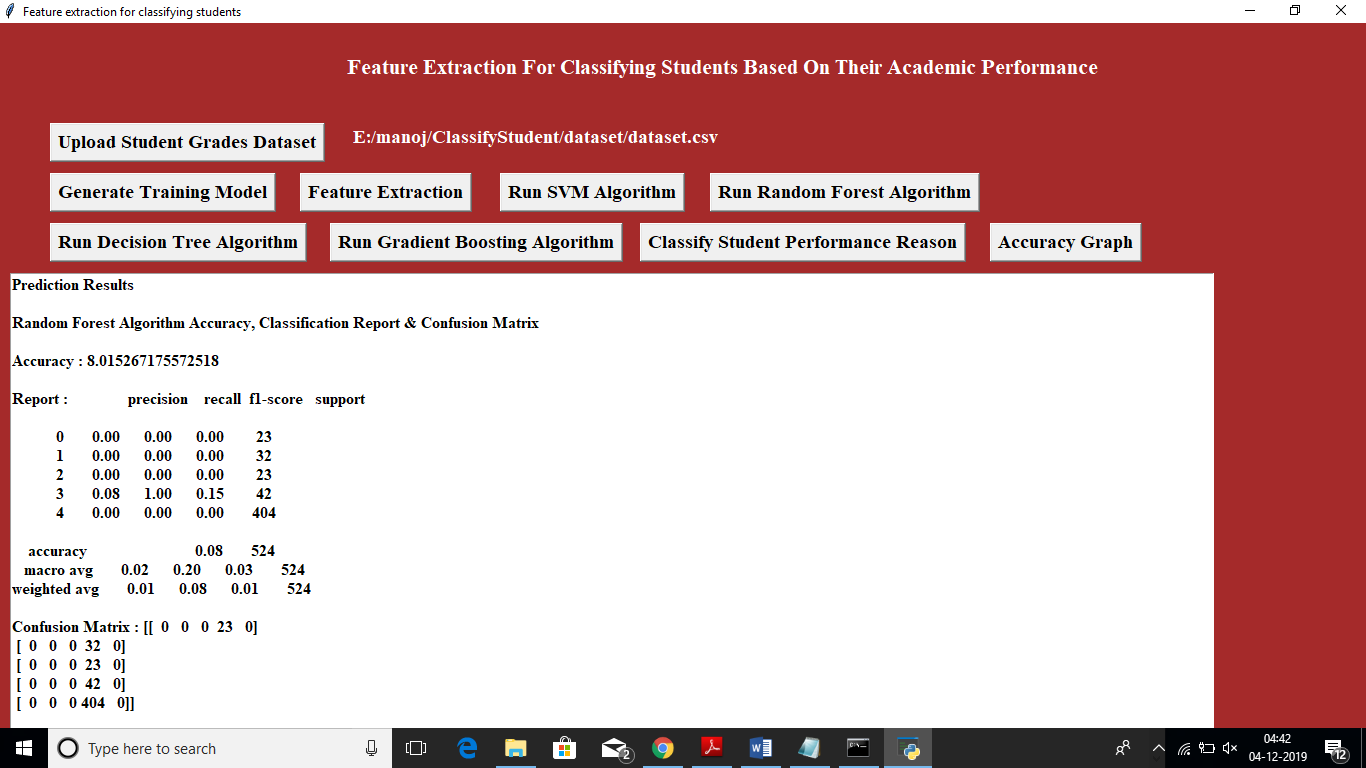
In above screen we can see total number of records in dataset and then displaying algorithm chooses how many records for training and testing purpose. Now click on ‘Features Extraction’ button to extract features and assign as class label to the classifier algorithms.



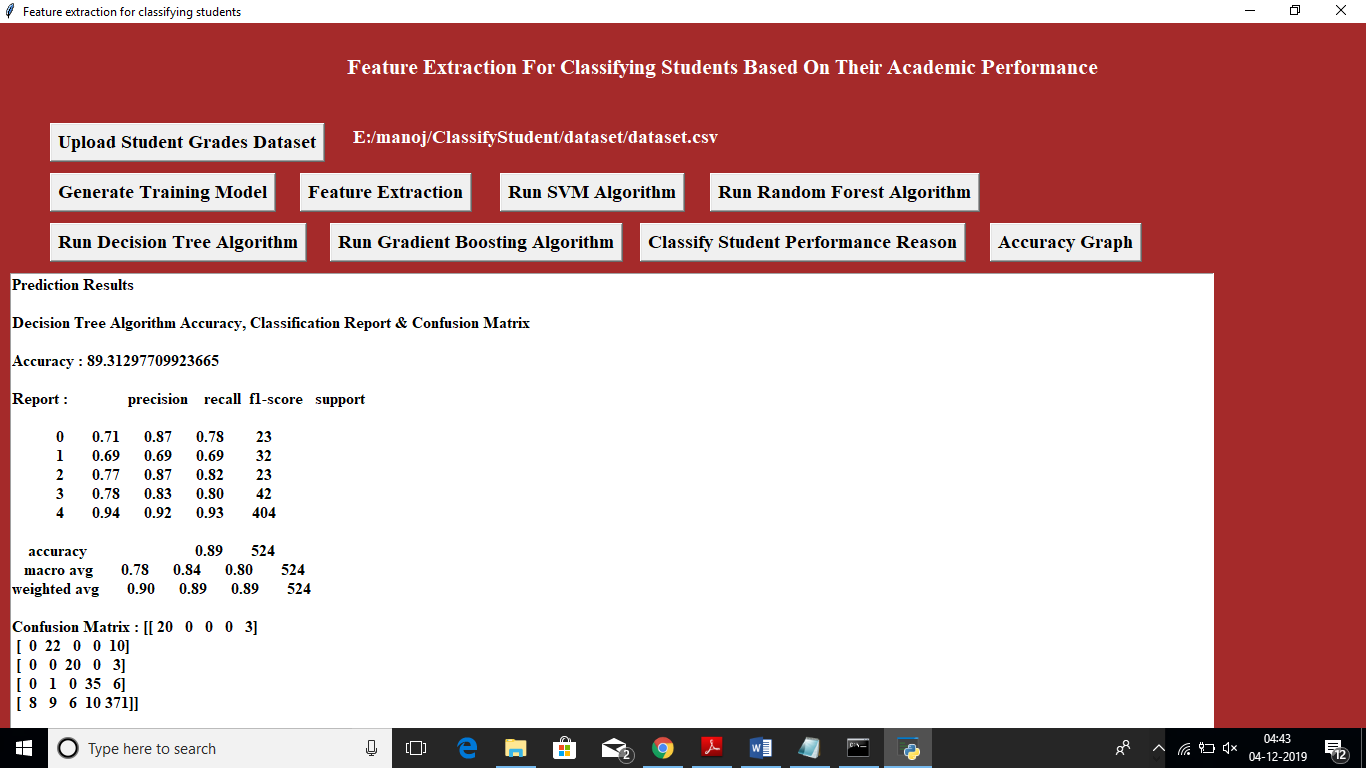
In above screen we can see dataset contains total 14 numeric features and extracted features are 4. After feature extraction click on ‘Run SVM Algorithm’ to build SVM train model and to get classifier accuracy and FSCORE value



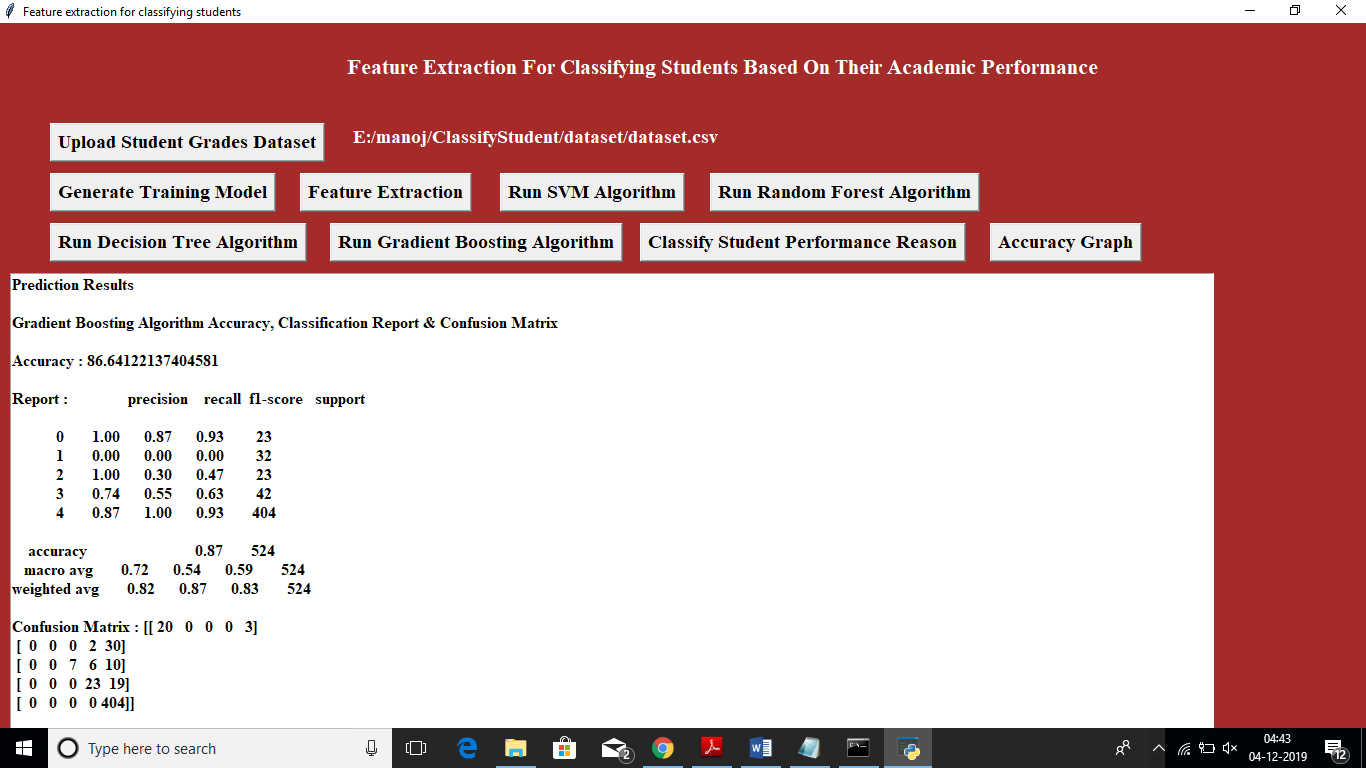
In above screen we can see SVM accuracy is 69% and we can see FSCORE value also. Now click on ‘Run Random Forest Algorithm’ to build its model



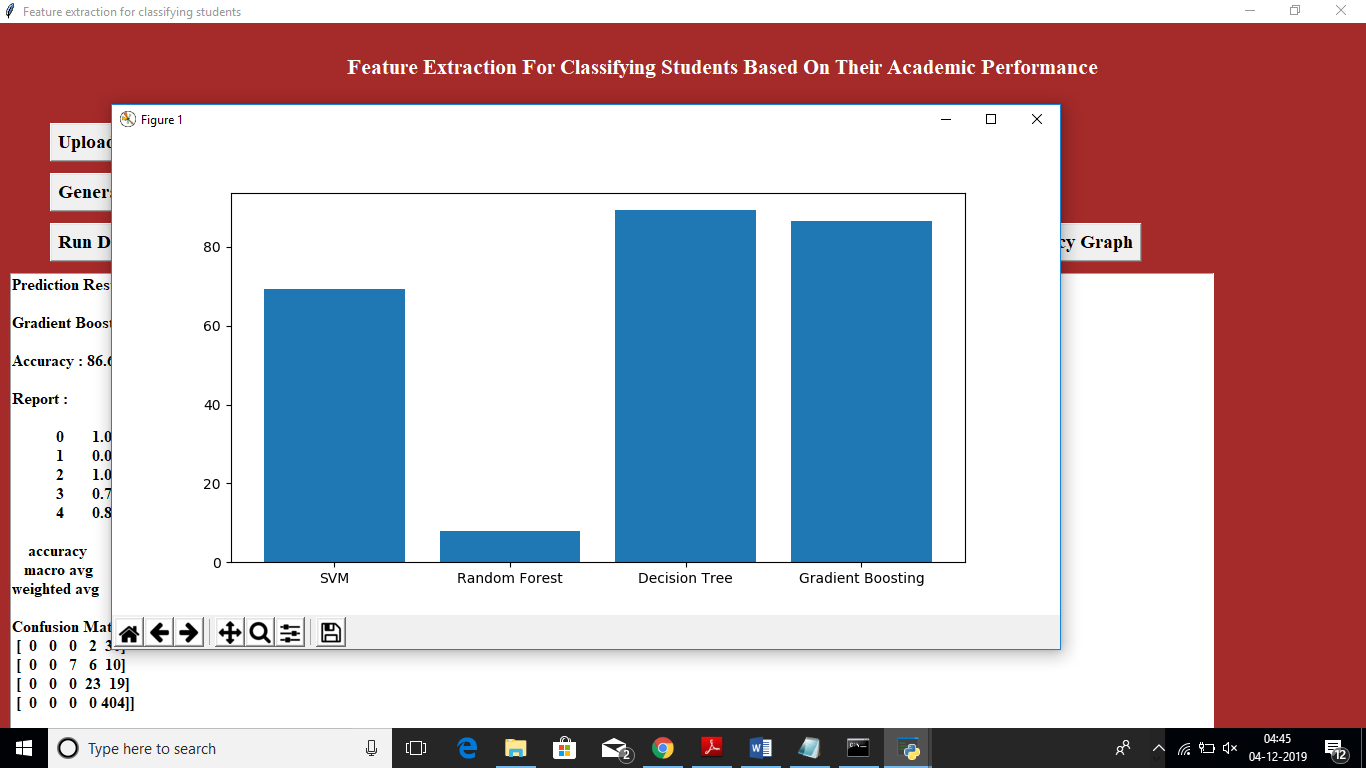
In above screen random forest got only 8% of accuracy. Now click on ‘Run Decision Tree Algorithm’ button to build tree model



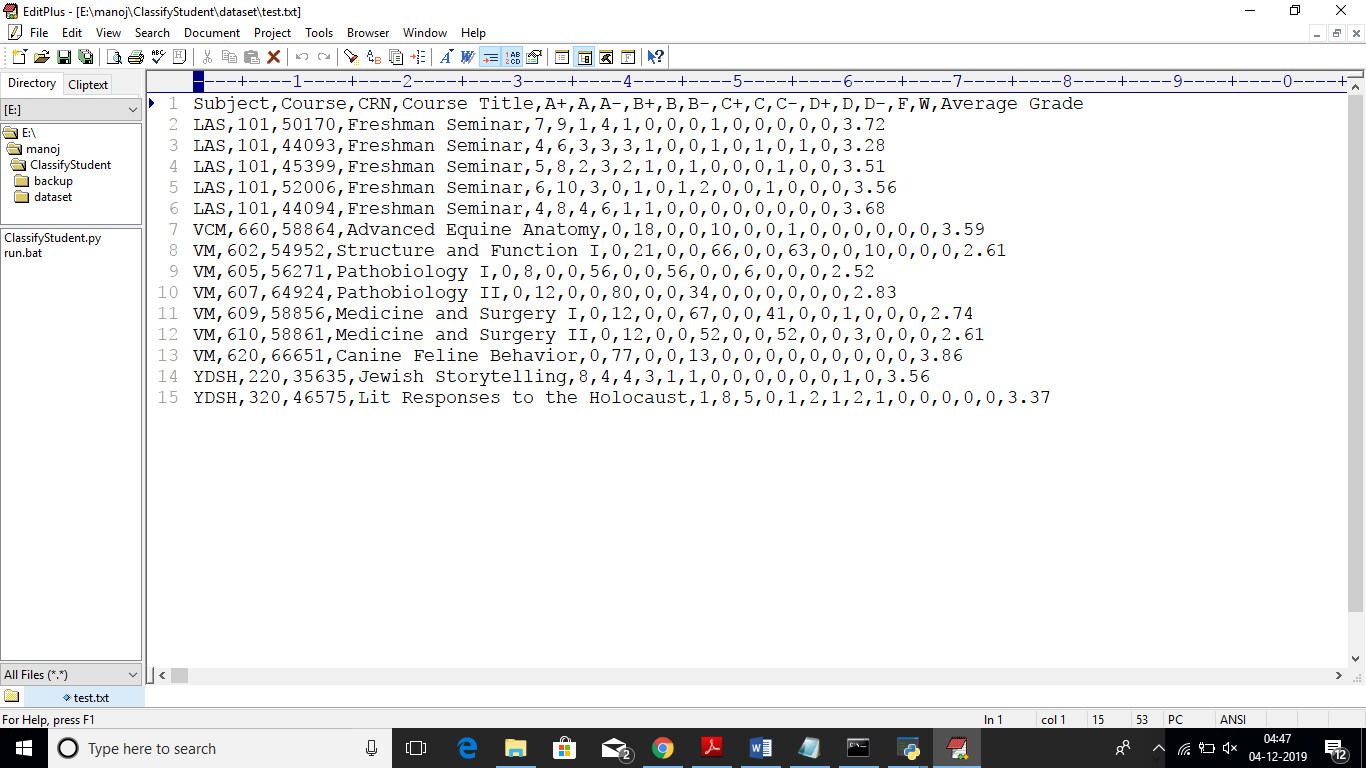
In above screen decision tree got 89% accuracy, now click on ‘Run Gradient Boosting Algorithm’.



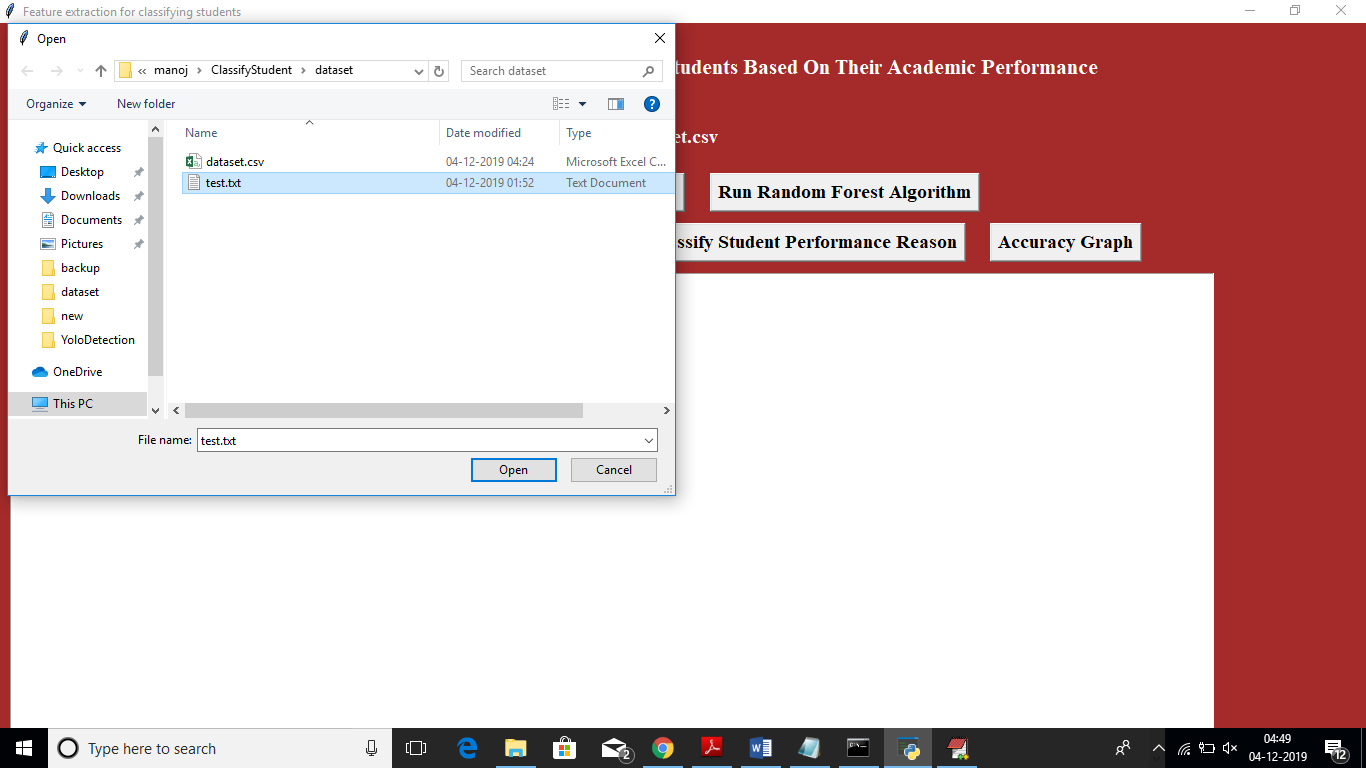
In above screen gradient boosting got 87% accuracy and decision tree got high accuracy but its FSCORE is less compare to gradient boosting. Now click on ‘Accuracy Graph’ button to get below accuracy graph



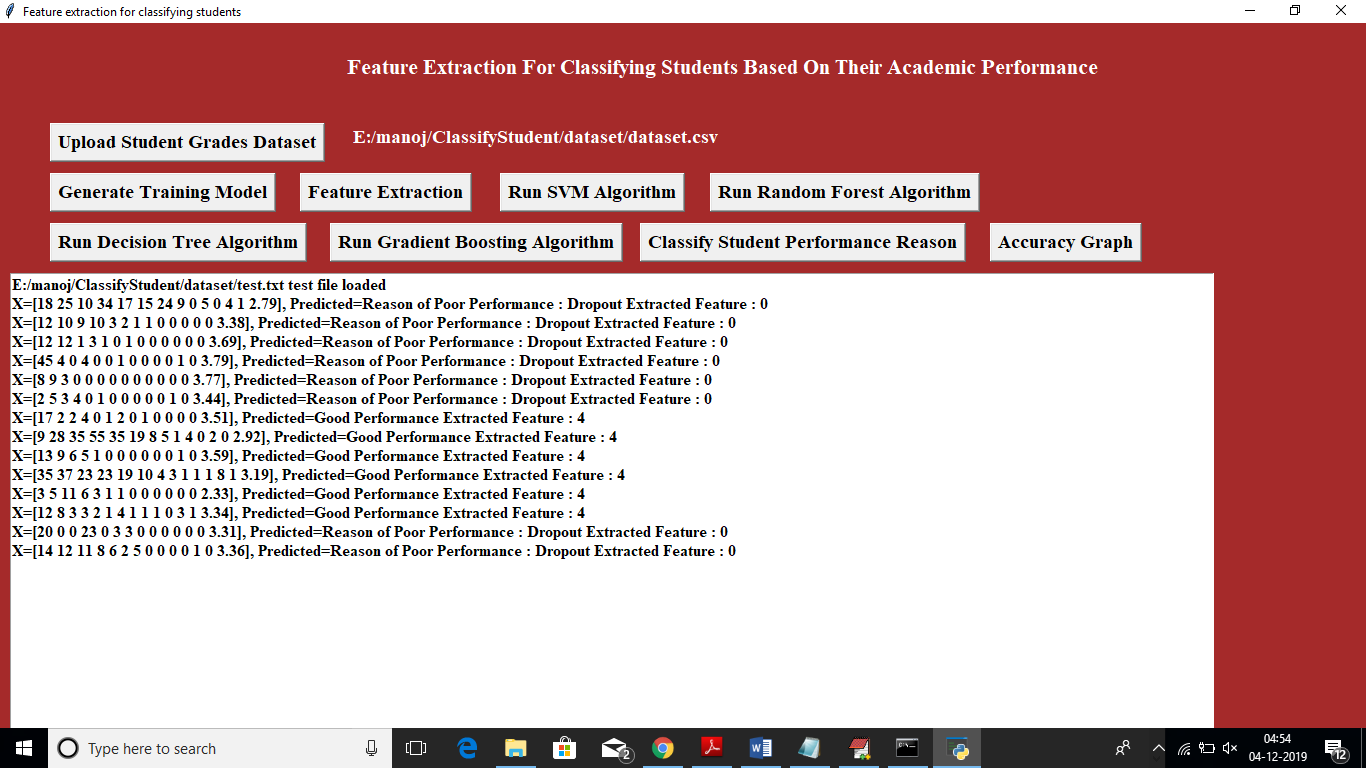
In above screen x-axis represents algorithm name and y-axis represents accuracy of that algorithm. Now we can test new student records on this train model to predict or classify new student performance. To check new student we need to upload ‘text.txt’ test dataset from dataset folder and this dataset contains below data



In above test data we don’t have extracted feature values such as 0,1,2,3 or 4 and this value will be predicted or classify by this machine learning algorithms. Just we need to click on ‘Classify Student Performance Reason’ button and upload test dataset then will get below result



After uploading test data will get below classification result



In above screen based on grades values application has given result as poor performance due to drop out or good performance