A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs

In this paper using Machine Learning algorithms author is suggesting concept to predict future courses performances of students by using students previous terms result data as feature vectors. Every year in all universities only 50 % students completing graduation courses successfully and remaining students are failed to complete course so by using this paper machine learning algorithms college peoples can predict future performance of students by giving his past performance GPA as input to the machine learning algorithms.

To implements this paper author has face 3 problem and gave solution to all 3 problems

1. All students will be differ in selected courses as different students may take different courses then how we can build machine learning model to predict performance of particular student in particular course. To solve this issue author is using Clustering concept using Matrix Factorization. In matrix factorization we will form cluster of related courses and then in matrix we will put value 1 (or marks obtained by students in term) if user selected course otherwise we will put value 0. So by forming this cluster we will have matrix feature vector which contains similar courses student’s data. This matrix features will be passed to machine learning to train model and to predict future performances. A single course consists of many subjects and all those subjects marks will be used to calculate GPA in all terms. Past course GPA and marks will be input to machine learning algorithms to predict future course performance.
2. Courses are not equally informative for making accurate predictions and by forming feature vector we will get prediction for selected course only.
3. All existing algorithms were concentrating on past data to predict future performance but in this paper we will used past data as well as on going course performance data to predict future course GPA. Students’ evolving progress needs to be incorporated into the prediction.

To implement this paper author using base classifiers such as Random Forest, SVM, Logistic Regression or KNN. The prediction results of base classifier will be pass to ensemble classifier to predict better results for ongoing courses and future courses.

To implement this project author using student performance dataset from UCLA University and this dataset saved inside dataset folder. Below are some example of dataset records.

**Math33A,Math33B,Math31B,MATH32A,MAE105A,CHEM20B,MAE103,MATSCI104,MAE105D,MAE94,PHYS1A,PHYS1B,GPA**

-0.13,-1.58,0,1.93,0,0,371.08,435.04,717.61,819.58,1.34,1.21,0

0,0,-5.16,1.01,-4.04,-0.49,336.5,481.63,672.11,850.33,4.59,1.68,0

0.84,2.85,1.85,5.32,1.47,1.8,379.72,411.92,586.72,717.06,5.87,1.95,1

0,0,-5.23,0,-4.21,-0.81,403.64,621.58,780.255,768.79,3.26,1.44,1

Above dataset is from UCLA mechanical students and in above dataset all bold names are the subject names of mechanical course and all decimal values are the marks obtained by students in those subjects. If student not taken course then marks will be 0 and in last column GPA is defines as 0 and 1. 0 means low GPA score and 1 means high GPA score, so using above dataset we will build base learning classifier and then pass result to propose ensemble algorithm called Ensemble-based Progressive Prediction (EPP). EPP algorithm will predict GPA value as LOW or HIGH for given test data or new student on going performance data. Below is the test data using for new students and in test data there will be no GPA value and EPP algorithm will predict it.

Math33A,Math33B,Math31B,MATH32A,MAE105A,CHEM20B,MAE103,MATSCI104,MAE105D,MAE94,PHYS1A,PHYS1B0

0.27,1.45,1.03,1.69,1.04,0.08,399.67,441.83,704.345,0,0,0

1.31,0.38,-1.78,1.13,-0.64,-1.23,0,477.21,0,630.17,3.78,0

0.51,1.45,-0.24,2,0.44,-0.49,330.89,469.75,688.65,1103.04,3.26,3.13

1.06,1.45,1.03,1.2,2.77,0,0,490.38,592.565,0,2.85,0

In above test dataset only subjects names and its marks are there but no GPA is available and to predict GPA we will apply EPP algorithm. In above dataset if student not taken subject or marks yet to expect then we will put 0.

In code we are using below statements to pass base classifier result to ensemble classifier

base = RandomForestClassifier() //creating base classifier object

epp = BaggingClassifier(base\_estimator=base)// passing base classifier to ensemble

epp.fit(X\_train1, y\_train) //now training ensemble with past data as training

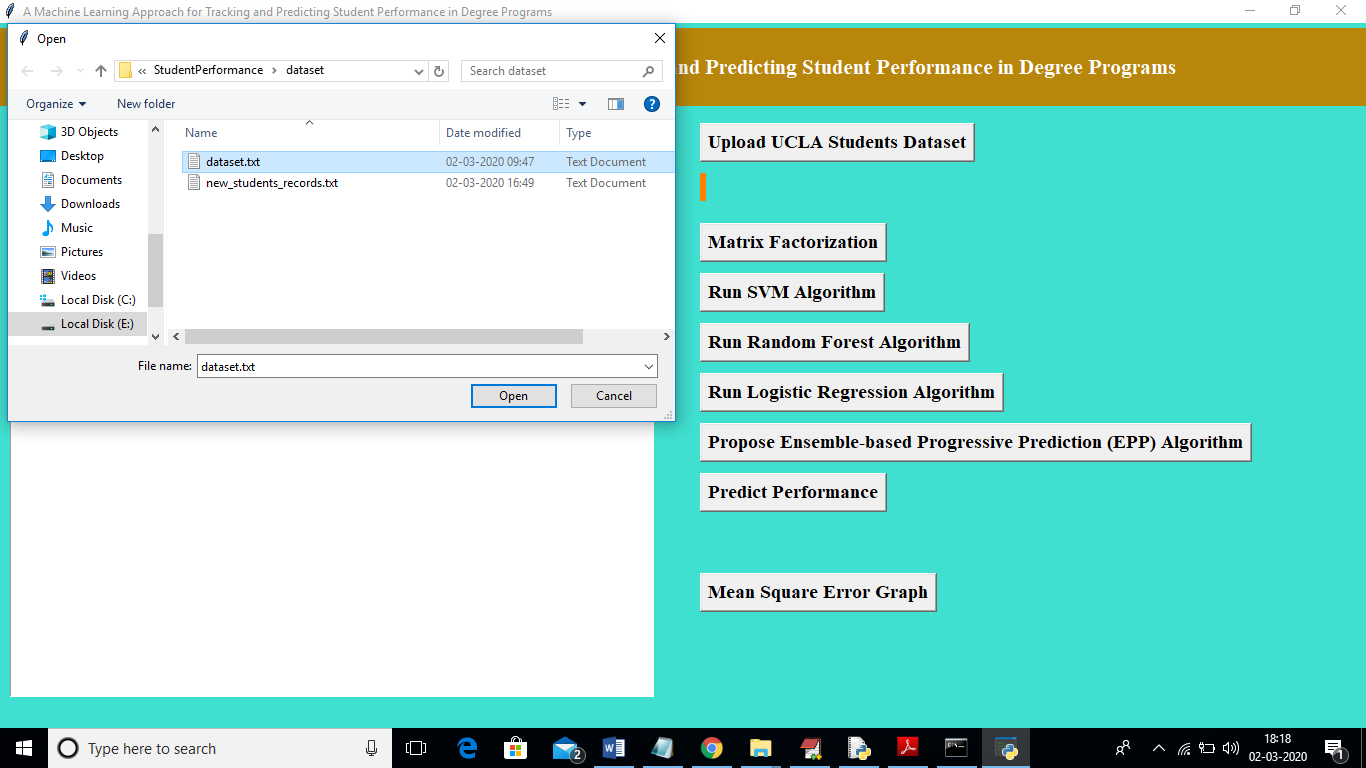
prediction\_data = prediction(X\_test1, epp) //now calling prediction function to predict future course GPA by passing test data as student on going course result. Prediction will be done using propose EPP algorithm.

Screen shots

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



In above screen click on ‘Upload UCLA Students Dataset’ button to upload dataset



In above screen I am uploading ‘dataset.txt’’ as student dataset. After uploading will get below screen



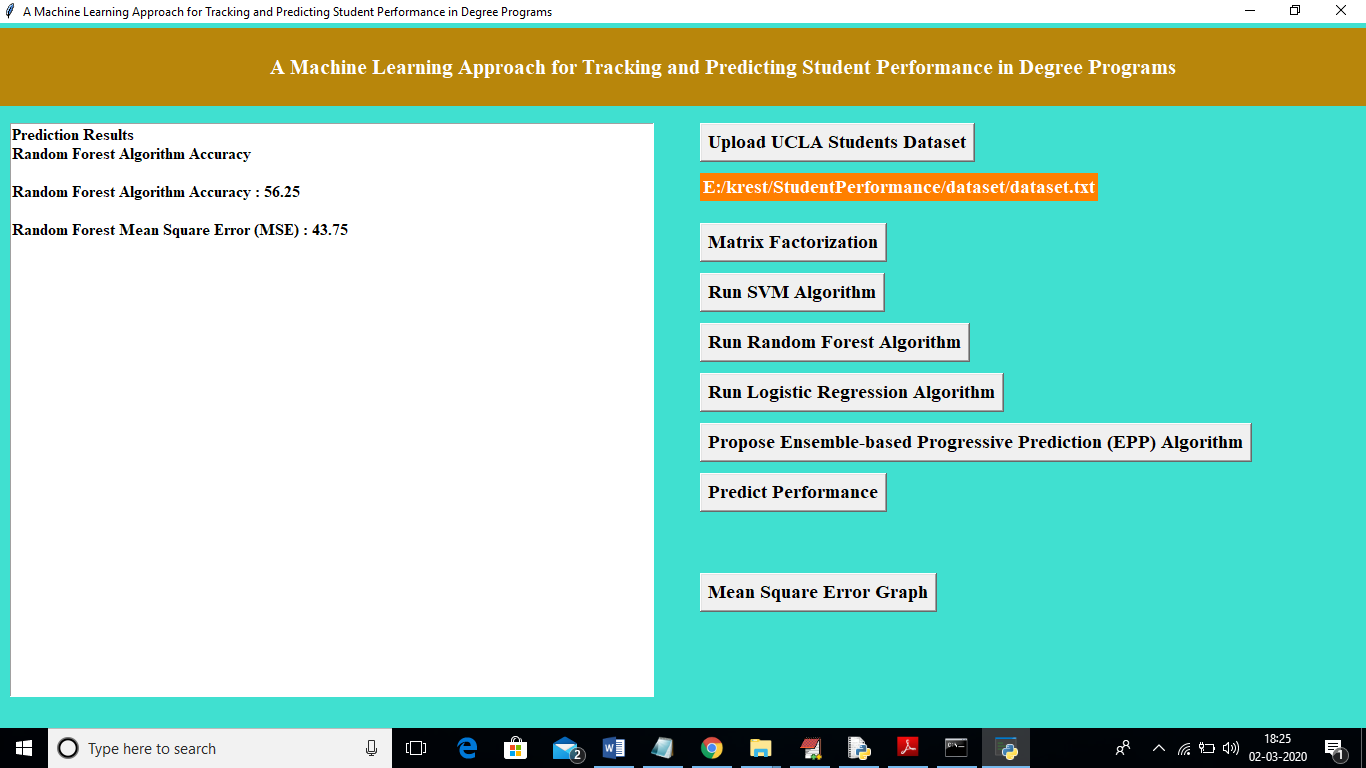
In above screen we can see dataset contains total 77 student’s records. Now click on ‘Matrix Factorization’ to build feature vector from dataset. In this matrix we will have all related course data and if student taken course then matrix contains marks otherwise 0.



In above screen we can see all records converted to feature vector and in above screen in first 3 lines we can see from above matrix application using 61 records to train machine learning model and 16 records to test accuracy or to calculate Mean Square Error of classifier. If algorithm prediction result is high then accuracy will be more and Mean Square Error (MSE) will be less. Now we got matrix and data to train and test classifier. Now click on ‘Run SVM Algorithm’ to train SVM classifier and to get it accuracy and MSE value



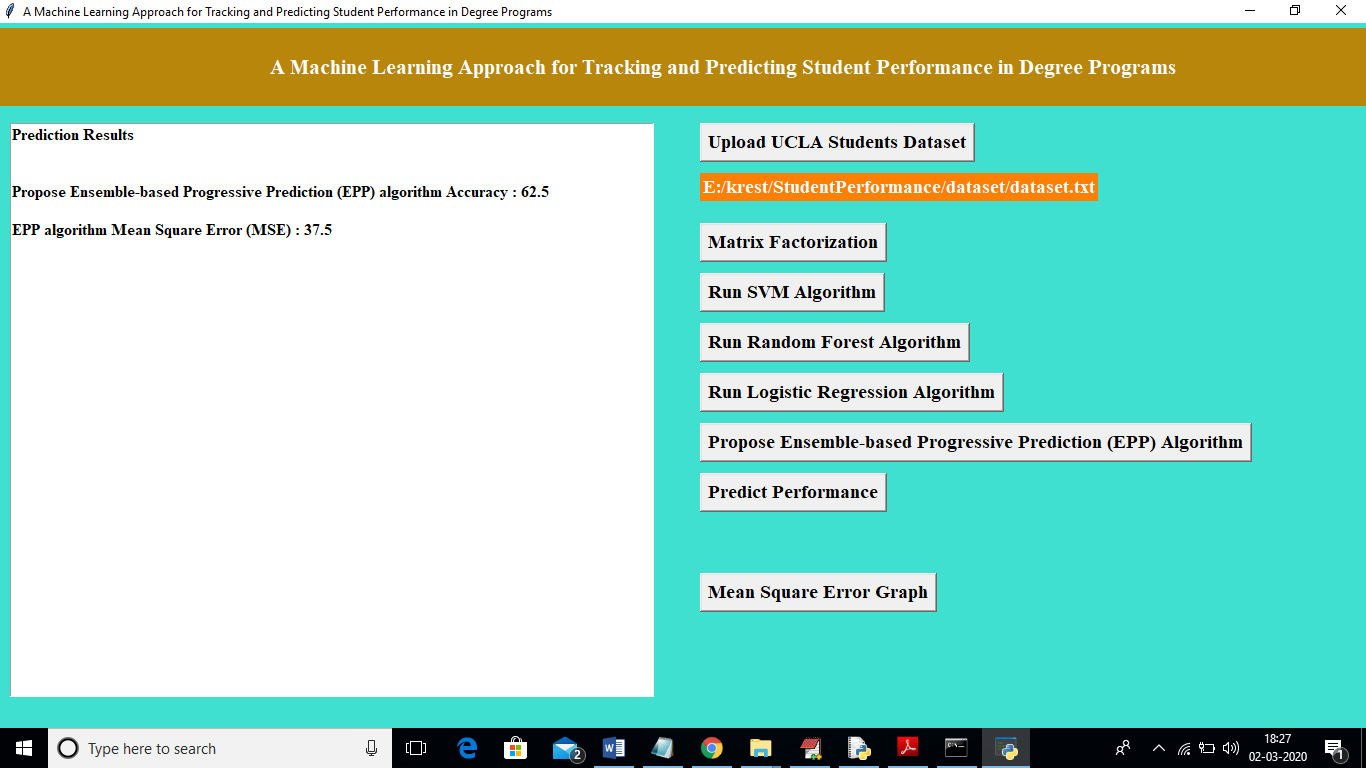
In above screen SVM MSE is 56%. Now click on ‘Run Random Forest Algorithm’ to generate training model using Random Forest and to get it accuracy and MSE



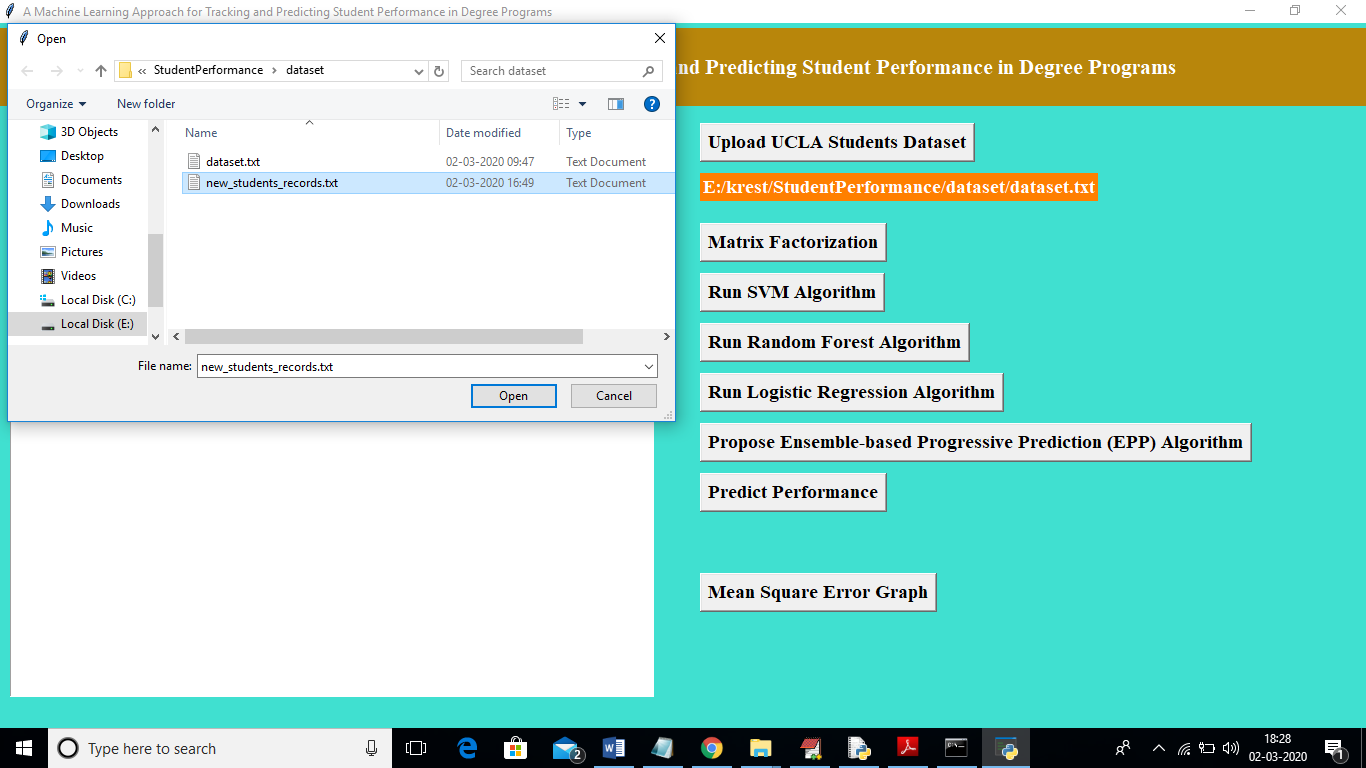
In above screen random forest got 43% MSE and now click on ‘Run Logistic Regression Algorithm’ button to get it accuracy and MSE



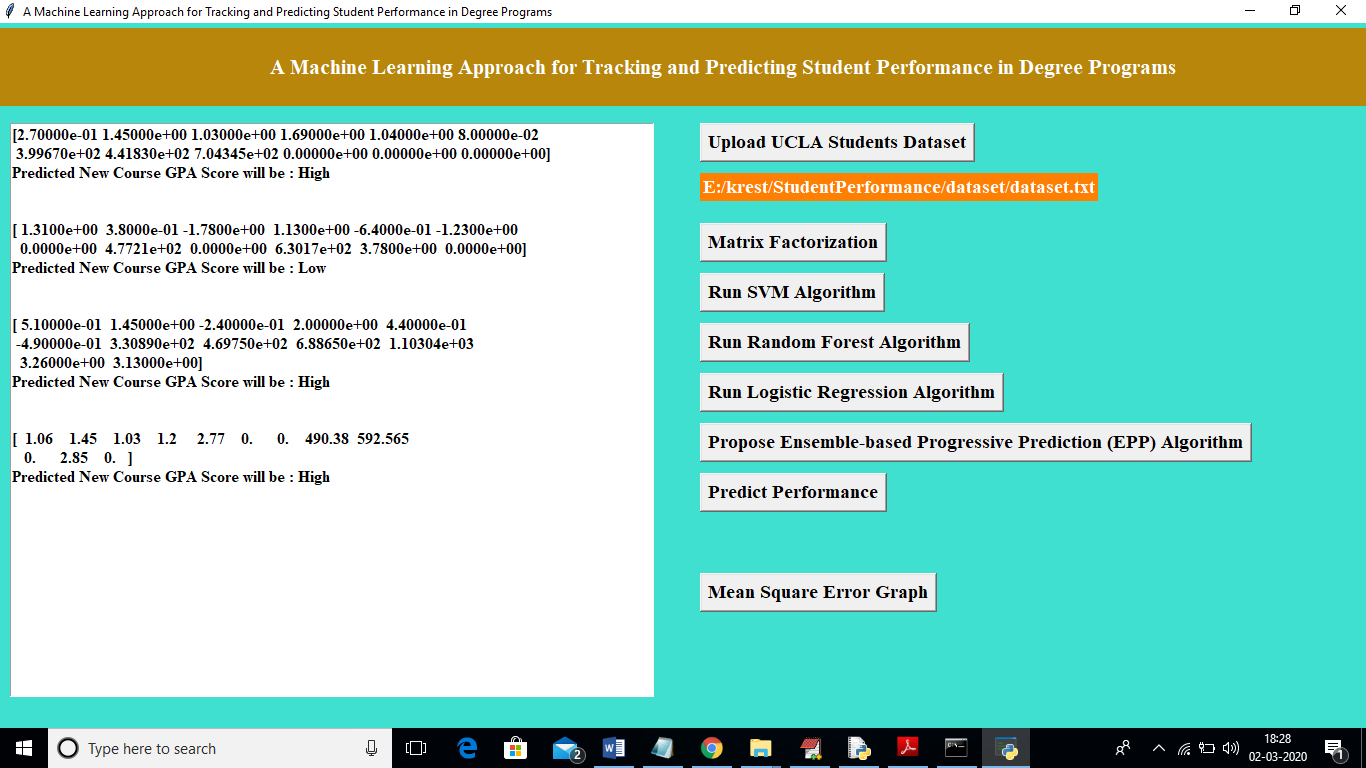
In above screen logistic regression got 43% MSE and now click on ‘Propose Ensemble-based Progressive Prediction (EPP) Algorithm’ button to generate model using propose EPP algorithm and to get it accuracy and MSE



In above screen EPP propose algorithm got 37% MSE and now click on ‘Predict Performance’ button to upload student on going test marks and to predict GPA for future course



In above screen uploading new student records as test file and below are the prediction results



In above screen in square brackets are the student marks of ongoing subjects and this marks are converted to matrix factorization and then applied on EPP train model to predict GPA as LOW or HIGH. In above screen after each test record I am displaying predicted result value. Now click on “Mean Square Error Graph’ button to get below graph



In above graph x-axis represents algorithm name and y-axis represents MSE (mean square error). From above graph we can see propose algorithm got less MSE error and has high accuracy compare to other algorithms. From above graph we can conclude that propose EPP is better in prediction compare to other algorithms