A Project Report

on

Academic Performance Evaluation

carried out as part of the course CS1730

Submitted by

Rashi Singh
169105142
7th Semester B.Tech (CSE)

In partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

Computer Science & Engineering

Department of Computer Science & Engineering, School of Computing and IT

MANIPAL UNIVERSITY
JAIPUR

Manipal University Jaipur,
Rajasthan, India
November, 2019

TABLE OF CONTENTS

Abstract

The purpose of educational institutions is to provide quality education to its students. One way to achieve highest level of quality in these institutions is by discovering knowledge for prediction regarding their scores and overall result. Academic Evaluation using Education Data Mining is vital in today's scenario as a student's marks are not only determined by their performance in classroom activities but also depend majorly on the time they spend on social media, their involvement in cultural/technical clubs, the time they spend with their friends, etc. and it is important to consider those factors in order to ascertain a student's academic performance. The main objective of this project is to use data mining methodologies to study student performance in the courses and predict their end semester results. Data mining provides many tools that could be used to study and analyze the student performance data. We aim to determine the GPA of a student on the basis of factors such as time spent on social media, time spent with friends, MOOC courses, involvement in clubs, attendance, etc. and to provide a feedback to both the administration and the students based on our findings. The data used in this project has been collected through a survey and the technique used for predicting the GPA of the student on the basis of above mentioned factors is random forests. Random Forests was found to be most suited for the data. It has the function for feature importance which was significant for this process. Also, the problem of overfitting was resolved and it gave higher accuracy with hyper tuning the parameters. A result of the findings can be provided to the faculty and students to look into the factors which lead to academic excellence.

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Motivation

The major challenging problem which most of the universities as well as students face is the identification of the factors that affect their results. The data repositories of universities have a vast amount of their previous student's data and with the help of Data Mining, we can transform this data into useful data by identifying the lacking factors of the universities by the responses of the students, predicting the performance of the current students in the university, understanding the learning process of students, etc. This can have a greater impact on educational research. The factors that we have considered in this project are not taken into account while determining the academic performance of a student when in reality; all these factors play a huge role in the overall performance.

1.1.1. This project will be useful for students to assess their progress in the semester and to improvise their schedule according to the predicted scores. This will be beneficial for an overall increase in their final grades.
1.1.2. The teachers working in the university are also benefited by this project, as it will enable them to identify the patterns of the students and to know more about the student's activities outside the classroom which affect his/her marks.

1.2 Project Statement

The aim of this project is to predict the GPA of a student using Random Forests based on factors such as:

1.2.1. Attendance
1.2.2. Study method (Video lectures, Textbook, Group Study, Class notes, Study material)
1.2.3. Learning method (Rote Learning, by making notes, Audio Visual Learning)
1.2.4. MOOC courses taken
1.2.5. Usefulness of MOOC courses during exams
1.2.6. Attendance percentage
1.2.7. Time spent on social media during exams
1.2.8. Time spent with friends
1.2.9. Time spent on watching movies/TV shows
1.2.10. Involvement in technical/cultural clubs
1.2.11. Number of events organized/participated in a semester
1.2.12. Anxiousness during exams
1.2.13. Number of hours of continuous studying
1.2.14. Preparation time before the exams

The findings from this project can be used by students and teachers to analyze student behavior and their impact on the grades they secure.

1.3 Organization of Report

The report consists starts with the background overview of the project in Chapter 2. Section 2.1 explains the concepts used and Section 2.2 mentions the technology used. Then, the methodology of the whole project is elaborated in Chapter 3; Section 3.1 containing detailed methodology and Section 3.2 consisting of the block diagram. After that, Chapter 4 describes the various modules that were implemented. Subsequently, Chapter 5 comprises of the future aspects of the project. Section 5.1 also mentions the timeline or the progress chart.

CHAPTER 2

BACKGROUND OVERVIEW

2.1 Conceptual Overview

Educational data mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings (e.g., universities and intelligent tutoring systems). [1] At a high level, the field seeks to develop and improve methods for exploring this data, which often has multiple levels of meaningful hierarchy, in order to discover new insights about how people learn in the context of such settings. In Educational Data Mining, various computerized methods are implemented to analyze the large collection of student data to reveal the useful patterns out of it. EDM mainly focuses on the study of improving and enhancing the learning process and understanding the learning process. This is the most suitable way for the prediction of end semester scores of a student. [3]

In order to solve the problem statement, visualization of the data is essential as it helps in finding patterns in the data. Tableau and Seaborne were used for visualization of data which gave a clear picture of the trends and factors that affect GPA. The next step was to choose an appropriate machine learning algorithm after going through various research papers on Educational Data Mining and visualizing the dataset. Random Forests is found to be most suited among other algorithms such as Decision Trees, SVM, Multiple Regression, etc. It gave the maximum accuracy and the F1 score. Using Random Forests, we also found out the most significant features. Henceforth, the importance of features was made easy to quantify.

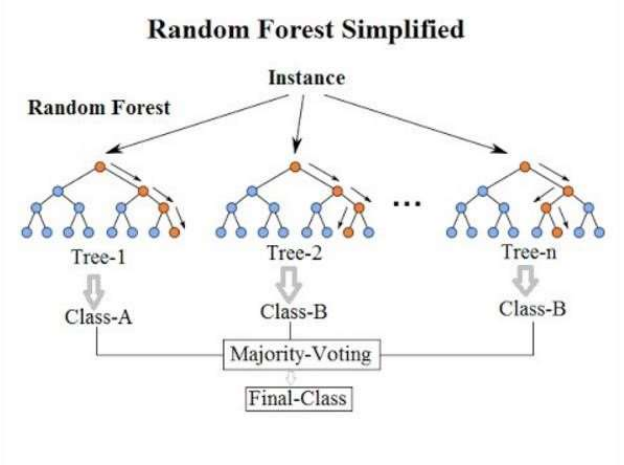


Fig 2.1 Random Forests

2.2 Technologies Used

2.2.1. Spyder(Python)
2.2.2. Jupyter Notebook
2.2.3. Tableau
2.2.4. Seaborne

CHAPTER 3
METHODOLOGY

The algorithm which is used for predicting the GPA is Random Forests. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[1][2] Random decision forests correct for decision trees' habit of overfitting to their set. Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time [2]. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction.
Here, Random Forests was used for classification.

3.1.1 *Steps followed for the implementation.*

Following are the steps which were followed for implementing Random Forest on the given dataset were:

Step1. State the question and determine required data.
The data collection was done in a form of a survey that was circulated among the students of the Engineering Discipline of Manipal University Jaipur. Third year students were the main target. The survey asked the students about their study methods, learning methods, the amount of time they spend with friends, social media, tv series etc. among many other factors.

Step2. Acquire the data in an accessible format. The data was collected in an excel sheet.

Step3. Identify and correct missing data points/anomalies as required. The missing data entries were filled using binning by means.

Step4. Prepare the data for the machine learning model
The categorical data attributes were converted into numeric. The dataset then, only consisted of numeric integer values.

Step5. Establish a baseline model that you aim to exceed
The Random Forest model is then selected and imported from the libraries. The data is divided into train and test data using the split data function.

Step6. Train the model on the training data
The training data is trained based on the model.

Step7. Make predictions on the test data
The predicts are then made on the test data.

Step8. Compare predictions to the known test set targets and calculate performance metrics.
Calculate the performance matrix i.e. precision, recall, F1 score and the support.

Step9. If performance is not satisfactory, adjust the model, acquire more data, or try a different modeling technique.
If the F1 score is not up to the mark, we have to either adjust the model or use a different algorithm.

Step10. Interpret model and report results visually and numerically Show the results in a graphical representation.

3.2. Block Diagram

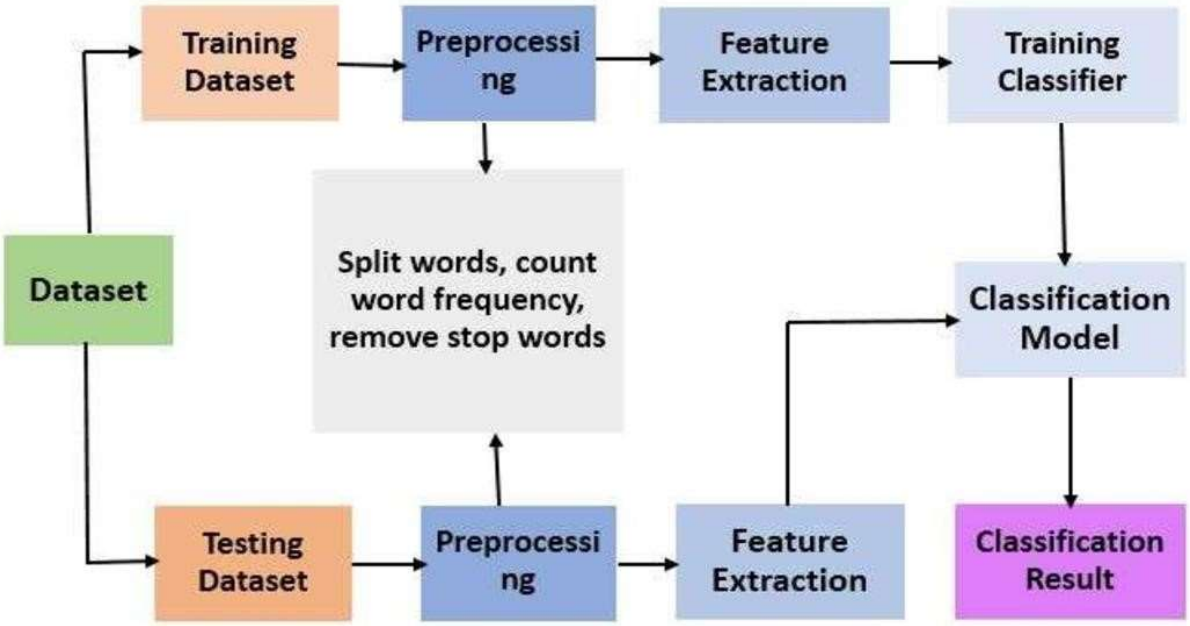


Fig 3.1: Block Diagram

CHAPTER 4
IMPLEMENTATION AND RESULTS

4.1 Modules

The entire project was divided into several modules for an easy and organized flow of work. These modules differ from one to the other making the tasks clear.

4.1.1 Data Collection

Data was collected through a google form that was circulated among the students of Manipal University Jaipur and the target set of students belonged to IT department. Several questions were asked in the survey some of which include the study methods, learning patterns, attendance, CGPA till the current semester, time spent on social media, etc.
The factors were chosen after doing an extensive study on research papers on Educational Data Mining.

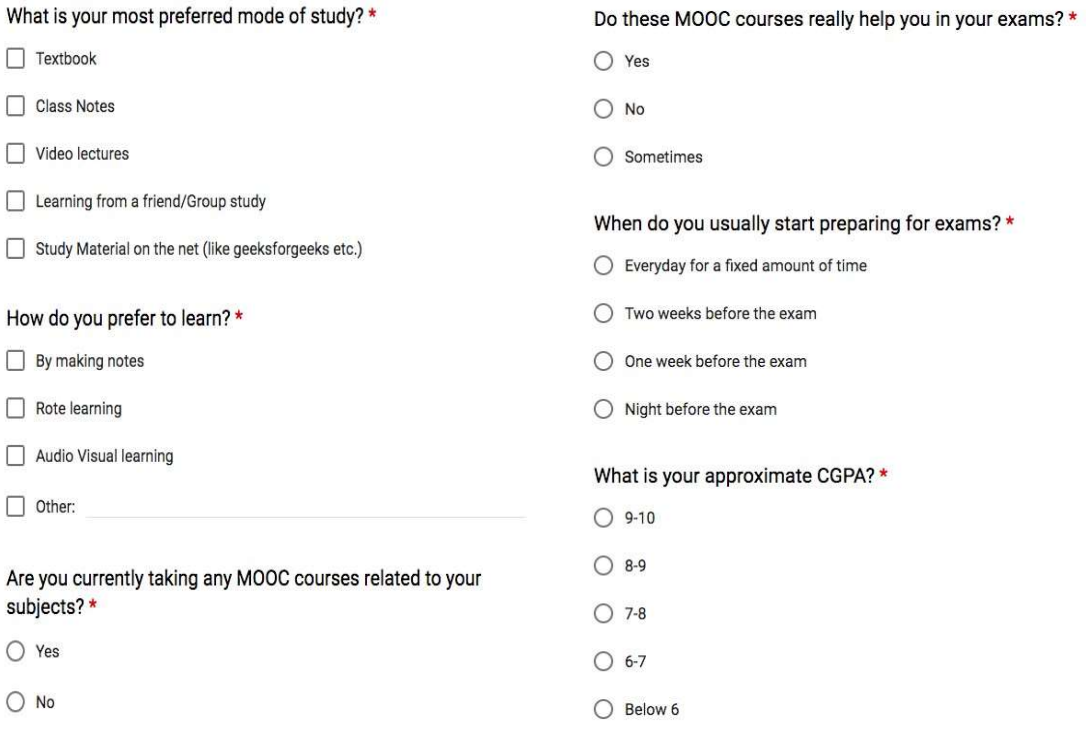


Fig 4.1: Data Collection through google forms.

4.1.2 Data Preprocessing

This module included tasks related to removing fake and irrelevant data and conversion of nominal data into numeric data for easy application of suitable algorithm.
Example: Several students have filled in more than one study methods in the form, a combination of Video lectures and Class notes or a combination of Textbook, Video lectures, Class notes, etc. This kind of data had to be converted into numeric form for better understanding and implementation of the algorithm.

What is your most preferred	How do you prefer to learn?	Are you curr	Do these MOOC	When do you usually start pr	What is your	What is your aver	How much time d	How much tim	Do you get anxious/nervous
Video lectures	Audio Visual learning	Yes	No	Night before the exam	8-9	80-90	3	1	Yes
Study Material on the net (lik	By not learning	No	No	Night before the exam	6-7	70-80	5	5	Yes
Class Notes, Video lectures,	By making notes, Audio Visu	Yes	Yes	Two weeks before the exam	8-9	80-90	3	3	Yes
Video lectures	Audio Visual learning	Yes	No	One week before the exam	6-7	70-80	1	2	Yes
Video lectures, Study Materi	By making notes, Audio Visu	Yes	Yes	Two weeks before the exam	7-8	80-90	1	1	Yes
Class Notes, Video lectures,	By making notes, Audio Visu	Yes	Sometimes	One week before the exam	6-7	Below 70	2	2	No
Textbook, Class Notes, Vide	By making notes	No	No	One week before the exam	8-9	80-90	2	1	Yes
Textbook, Class Notes, Vide	By making notes, Audio Visu	Yes	Sometimes	One week before the exam	7-8	80-90	3	2	No
Study Material on the net (lik	By making notes	No	Sometimes	One week before the exam	7-8	70-80	2	2	No
Class Notes, Learning from a	By making notes	No	Yes	One week before the exam	7-8	70-80	3	1	No
Class Notes	Audio Visual learning	Yes	Yes	Everyday for a fixed amount	8-9	Above 90	3	2	No
Study Material on the net (lik	Audio Visual learning	No	Sometimes	Night before the exam	6-7	80-90	2	1	Yes
Video lectures	By making notes, Audio Visu	Yes	Yes	Night before the exam	6-7	70-80	2	4	No
Textbook, Video lectures, St	By making notes, Audio Visu	Yes	Sometimes	One week before the exam	7-8	80-90	3	2	No
Class Notes, Video lectures,	By making notes, Audio Visu	No	Yes	Everyday for a fixed amount	6-7	70-80	1	2	No
Textbook, Learning from a fri	Audio Visual learning	Yes	Yes	One week before the exam	7-8	80-90	3	1	Yes
Class Notes, Video lectures,	By making notes, Audio Visu	No	Sometimes	One week before the exam	7-8	80-90	5	2	No
Video lectures	By making notes	No	Sometimes	One week before the exam	6-7	80-90	3	2	Yes
Video lectures, Learning fron	Audio Visual learning	No	No	Night before the exam	7-8	Below 70	3	2	No
Video lectures	By making notes	No	No	Two weeks before the exam	7-8	80-90	3	1	No
Class Notes, Video lectures	Rote learning	No	Sometimes	One week before the exam	8-9	80-90	1	1	Yes
Class Notes, Video lectures	By making notes, Rote learn	No	No	One week before the exam	7-8	80-90	3	3	Yes
Textbook, Class Notes, Vide	By making notes, Rote learn	No	Sometimes	One week before the exam	8-9	80-90	2	2	Yes
Video lectures	Save me plis don't wanna le	No	No	Night before the exam	8-9	70-80	5	4	No
Class Notes, Learning from a	By making notes	No	Sometimes	Night before the exam	9-10	70-80	5	1	No
Textbook, Video lectures	By making notes	Yes	Yes	Two weeks before the exam	8-9	Above 90	2	2	Yes
Class Notes, Video lectures	Audio Visual learning	No	Sometimes	One week before the exam	7-8	Above 90	3	1	Yes
Class Notes, Study Material	By making notes, Audio Visu	No	Sometimes	Night before the exam	8-9	80-90	3	1	No
Video lectures, Learning fron	By making notes, Audio Visu	Yes	Sometimes	One week before the exam	6-7	80-90	3	1	No
Class Notes, Study Material	By making notes	No	No	One week before the exam	8-9	70-80	2	1	Yes
Class Notes, Video lectures	By making notes, Audio Visu	Yes	Sometimes	Two weeks before the exam	9-10	Above 90	2	1	Yes
Textbook, Class Notes	Rote learning	Yes	Sometimes	One week before the exam	9-10	70-80	4	2	Yes
Textbook	By making notes	Yes	Yes	Two weeks before the exam	9-10	80-90	3	2	Yes
Class Notes, Video lectures,	By making notes, Audio Visu	Yes	No	One week before the exam	8-9	Above 90	5	1	No
Textbook, Video lectures, St	By making notes, Rote learn	Yes	Yes	Night before the exam	7-8	80-90	3	2	No
Textbook, Class Notes, Lear	By making notes, Audio Visu	Yes	Yes	One week before the exam	8-9	Above 90	1	1	Yes
Video lectures	By making notes	No	No	One week before the exam	7-8	80-90	3	1	Yes
Class Notes, Video lectures,	By making notes, Audio Visu	Yes	Sometimes	Two weeks before the exam	8-9	Above 90	5	4	No
Study Material on the net (lik	Audio Visual learning	Yes	Sometimes	One week before the exam	7-8	70-80	3	2	Yes
Video lectures, Study Materi	By making notes	Yes	Yes	One week before the exam	7-8	70-80	5	3	Yes
Textbook, Class Notes, Vide	By making notes	Yes	Yes	One week before the exam	7-8	Below 70	3	1	Yes
Textbook, Class Notes, Vide	By making notes	No	No	Night before the exam	7-8	70-80	3	2	No
Textbook, Class Notes, Stud	By making notes, Audio Visu	No	No	Night before the exam	7-8	70-80	2	2	Yes

Fig 4.2: Dataset before preprocessing

study_method	code	learn_method	mooc	mooc_help	preptime	cgpa	avg_attendance	social_media	movies	nervous	events	breakstudy	friends time	clubs_involve	clubnum
1	V	1	1	0	1	4	3	3	1	1	4	2	1	1	2
5	S	7	0	0	1	2	2	5	5	1	7	0	5	1	3
25	CVS	4	1	1	3	4	3	3	3	1	4	3	5	1	2
1	V	1	1	0	2	2	2	1	2	1	7	1	1	1	1
28	VS	4	1	1	3	3	3	1	1	1	2	1	2	1	1
21	CVL	4	1	2	2	2	1	2	2	0	3	1	4	1	1
7	TCV	2	0	0	2	4	3	2	1	1	4	1	5	1	2
7	TCV	4	1	2	2	3	3	3	2	0	3	2	3	0	1
5	S	2	0	2	2	3	2	2	2	0	4	2	3	0	1
23	CL	2	0	1	2	3	2	3	1	0	1	1	3	0	1
2	C	1	1	1	4	4	4	3	2	0	2	2	3	1	1
5	S	1	0	2	1	2	3	2	1	1	2	1	3	1	1
1	V	4	1	1	1	2	2	2	4	0	0	2	3	0	1
19	TVS	4	1	2	2	3	3	3	2	0	2	2	4	1	1
25	CVS	4	0	1	4	2	2	1	2	0	0	2	5	0	1
12	TLS	1	1	1	2	3	3	3	1	1	5	2	4	1	3
20	CVLS	4	0	2	2	3	3	5	2	0	4	3	5	1	1
1	V	2	0	2	2	2	3	3	2	1	1	2	3	1	1
28	VL	1	0	0	1	3	1	3	2	0	5	1	2	0	1
1	V	2	0	0	3	3	3	3	1	0	0	3	2	0	1
22	CV	3	0	2	2	4	3	1	1	1	5	2	3	1	2
22	CV	5	0	0	2	3	3	3	3	1	7	2	4	1	1
8	CVL	6	0	2	2	4	3	2	2	1	4	2	3	1	2
1	V	7	0	0	1	4	2	5	4	0	0	0	1	0	1
23	CL	2	0	2	1	5	2	5	1	0	7	0	5	1	2
17	TV	2	1	1	3	4	4	2	2	1	0	2	3	0	1
22	CV	1	0	2	2	3	4	3	1	1	6	1	5	1	2
27	CS	4	0	2	1	4	3	3	1	0	3	2	5	0	1

Fig 4.3: Dataset after Preprocessing

4.1.3 Data Cleaning

Data Cleaning included the job of filling missing values in the data. This was done through binning by mean.

4.1.4 Data Visualization

Visualization of data is very important in any given dataset. It helps in noticing the patterns and trends in data and is a very good tool to analyze the data. This is also useful in order to see if the data is linearly separable or not and to choose the appropriate algorithm to be used. Tableau and Seaborne were used for data visualization.

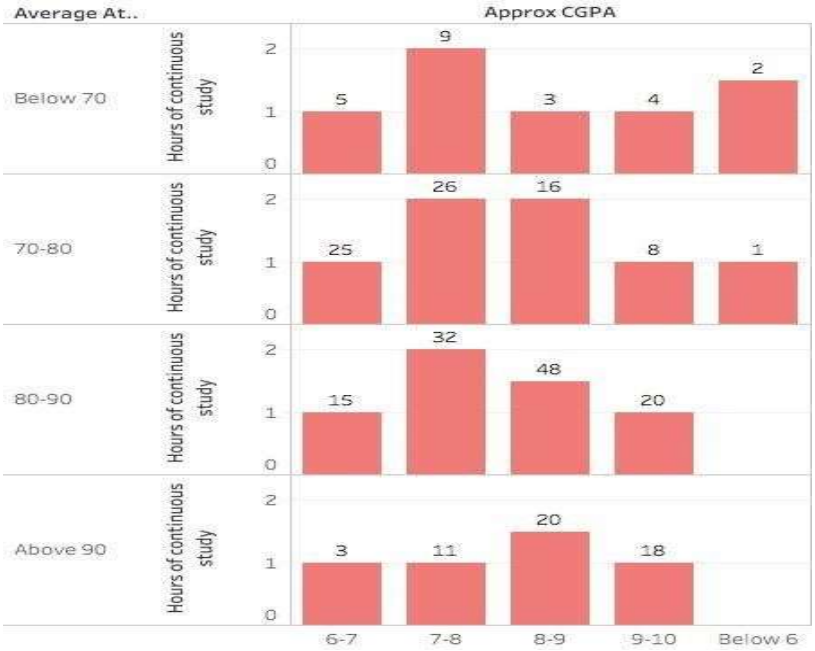


Fig 4.4: Graph of CGPA vs Attendance and Hours of Study

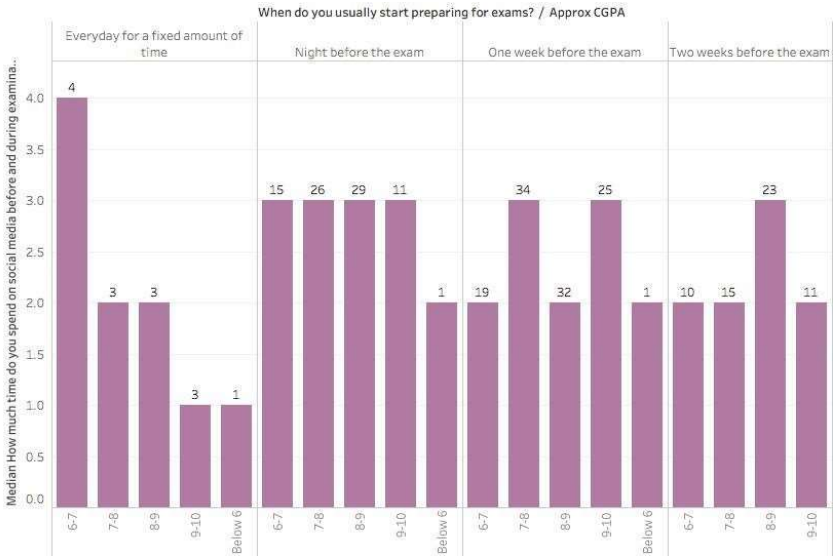


Fig 4.5: Graph of Social media usage and Preparation time vs CGPA

4.1.5 Application of Model

Random Forests has been used in order to predict the GPA of the student. Most of the work was done on Jupyter Notebook.

```
In [43]: from sklearn.cross_validation import train_test_split

In [44]: X=survey.drop('cgpa',axis=1)
         y = survey['cgpa']

In [45]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)

In [46]: from sklearn.ensemble import RandomForestClassifier
         rfc = RandomForestClassifier(n_estimators=100)
         rfc.fit(X_train, y_train)

Out[46]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                     max_depth=None, max_features='auto', max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
                     oob_score=False, random_state=None, verbose=0,
                     warm_start=False)

In [47]: rfc_pred = rfc.predict(X_test)

In [48]: from sklearn.metrics import classification_report,confusion_matrix

In [49]: print(confusion_matrix(y_test,rfc_pred))

         [[ 0  0  1  1  0]
          [ 0  2  6  6  2]
          [ 0  4  8 10  1]
          [ 0  4  5 11  1]
          [ 0  1  2 12  3]]
```

Fig 4.6: Code Snippet of Random Forests

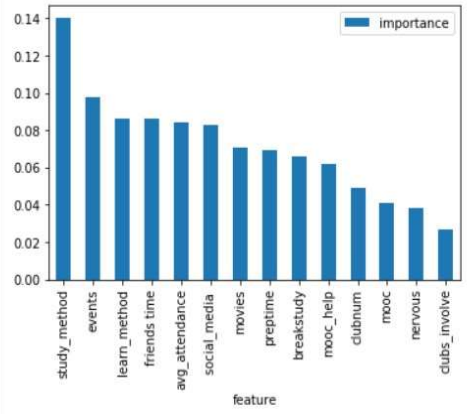
4.2 Results

After successfully completing the tasks in the above mentioned modules, there are a few findings:

4.2.1. The feature that affects the GPA the most is the Study Method a student chooses to follow.

```
In [54]: importances.plot.bar()

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x1b204998c18>
```



feature	importance
study_method	0.140
events	0.098
learn_method	0.086
friends time	0.086
avg_attendance	0.084
social_media	0.083
movies	0.071
preptime	0.069
breakstudy	0.066
mooc_help	0.062
clubnum	0.049
mooc	0.041
nervous	0.038
clubs_involve	0.027

Fig 4.7 Most important factors

4.2.2. The F1 score of the model is 0.28+0.5(To normalize the data) = 0.8

```
In [50]: print(classification_report(y_test,rfc_pred))
```

	precision	recall	f1-score	support
1	0.00	0.00	0.00	2
2	0.18	0.12	0.15	16
3	0.36	0.35	0.36	23
4	0.28	0.52	0.36	21
5	0.43	0.17	0.24	18
avg / total	0.31	0.30	0.28	80

Fig 4.8 Performance Analyzers

```
In [155]: rfc.predict([[17, 2, 0, 0, 1, 1, 3, 3, 0, 7, 0, 3, 1, 3 ]])
Out[155]: array([5])
```

Fig 4.9 Prediction for GPA

REFERENCES

Journal / Conference Papers

[1] Brijesh Kumar Baradwaj, Saurabh Pal. "Mining Educational Data to Analyze Students Performance, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[2] Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, 61, 133-145

[3] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

Reference / Hand Books

[1] Jiawei Han, Micheline Kamber, "Data Mining - Concepts and Techniques", Elsevier publications, Third Edition.

Web

[1] https://en.wikipedia.org/wiki/Educational_data_mining

[2]https://www.researchgate.net/publication/304293217_Educational_Data_Mining_techniques_and_their_applications