ANALYSIS OF THE FACTORS INFLUENCING STUDENT ACADEMIC PERFORMANCE IN HIGHER EDUCATION USING MACHINE LEARNING TECHNIQUES

Dr. S. Hemalatha

Associate Professor, UG Department of Computer Applications

Nallamuthu Gounder Mahalingam College, Pollachi

ABSTRACT - In higher education institute, many students struggle to complete the courses with good grade because there is no dedicated support offered to students in India who needs special attention in their registered course. Prediction of student's performance became an urgent need in most of educational institutes to improve teaching quality based on the information collected from students through questionnaires and academic details from the institutions where they are studying. The main aim is to identify students who would have difficulty in their learning and to take precautionary measures to help them. Measuring academic performance of the student in current generation is hard since they are affected by several factors like home, college, teacher, health, surroundings, fast growing technologies like social media, Extra-curricular activities etc. In this paper, insights are given into the different Machine Learning algorithms like Support Vector Machine (SVM), Naïve Bayes and K-Nearest Neighbors algorithm (KNN) and advice on the best model to use accordingly. This study might bring edges and impacts to students, educators/lecturers. The factors that are researched in this study are – Demographic Profile, Personal factors, Exam related, Department/Staff related, ECA and Social media related factors.

Keywords: Machine Learning, Academic Performance, Supervised Machine Learning, Statistical Evaluation

.

1. INTRODUCTION

The main goal of each Educational Institution is to deliver quality knowledge and skills to students so that they are competitive in the labor market. The analysis of students' progress during their studies provides the management with information about the probability of success of each student. One way to achieve this goal is to timely predict student performance. Thus, it is easy to identify students who need support and take measures to improve educational outcomes. This would help teachers to provide an effective approach to teaching. It is a general perception that there are some personal characteristics, learning habits, previous academic background and college environmental factors which affect performance of the students at College Level. Educational data mining [7] are broadly used in academic prediction on student performance in the classroom education. Most of the researchers in abroad evaluated student coursework performance against the passing grade of the exam and used CGPA [8] as a Key Performance Index (KPI) to analyze student's academic performance. In some countries, the academic performance was analyzed using statistical tools with an intention of establishing the relationship between students' admission points, social economic status, Educational level of parents, former school background etc. To take any decision, data is analyzed in such a way that it should able to answer certain questions. Based on the answers collected in the analysis phase, the perfect decision can be made by using different Machine Learning techniques and students can be guided in a particular direction without diluting educational standards. The Student Academic Performance was influenced by means of the two different factors namely Independent and Dependent Factors. There are various facts such as financial condition, Parents Education, Stress which comes under Independent factors. Similarly, SGPA/CGPA, Marks obtained in End Semester Exam, Time to graduate the students in their respective colleges or university may come under the dependent factors. The paper aims to identify the key variables that affect educational success and to select the most effective ML algorithm to predict student achievement.

2. LITERATURE REVIEW

In the research paper titled "Factors Affecting The Academic Performance Of College Students", Nisha Arora (2017) [5] exposed that the academic achievement of any student is the result of a complex relationship among the various factors and the variance response like Teaching Effectiveness - 17.985% variance, Distraction Factors - 16.185% variance, Personality Traits - 13.368% variance, Study Habits - 11.382% variance, Family Environment - 10.379% variance. One of the studies titled "A Study Model on the Impact of Various Indicators in the Performance of Students in Higher Education", Jai Ruby, K. David (2014) [3] states that student academic and personal details along with their attendance were collected from the student information system. The collected information was integrated into a distinct table. Student dataset contains various attributes like theory scores, laboratory scores, medium of study, UG course, family income, parental education, first generation learner, stay, extracurricular activities [1] etc. The study conducted by Dr. Johnson (2018) [4], in the paper titled "The Factors Affecting The Academic Performance Of The Indian College Students: A Case Study", it was ensured that the factors like teaching methods in the school past studied, teaching methods followed by the colleges, facilities available in the college, students behavior or mindset, students awareness about their courses, students awareness about their career or opportunities, extracurricular activities of the students, technological impact on students (Internet, Smart phone and so on) influencing academic performance of the Indian college students. Some academicians think that personal characteristics, learning habits are more important whereas others argue that previous academic background and college environmental factors are more related with the students' performance [2]. In [6], Rastrollo-Guerrero J,et.al., stated different techniques and different types of data that are used and analyzed. Based on the data collected in this review, the most widely used technique for predicting students' behavior was supervised learning, as it provides accurate and reliable results. In particular, the SVM algorithm was the most used by the authors and provided the most accurate predictions.

3. METHODOLOGY

Figure 1. displays the various steps for predicting student performance. In the following section, a concise assessment of the machine learning model that will be used to predict the factors influencing the performance of student academic is elaborated. The Steps to build a Machine Learning Model are

- Data Collection
- Data Preparation
- Model Building
- Train the Model
- Evaluate the Model
- Tuning the Parameter
- Make Predictions

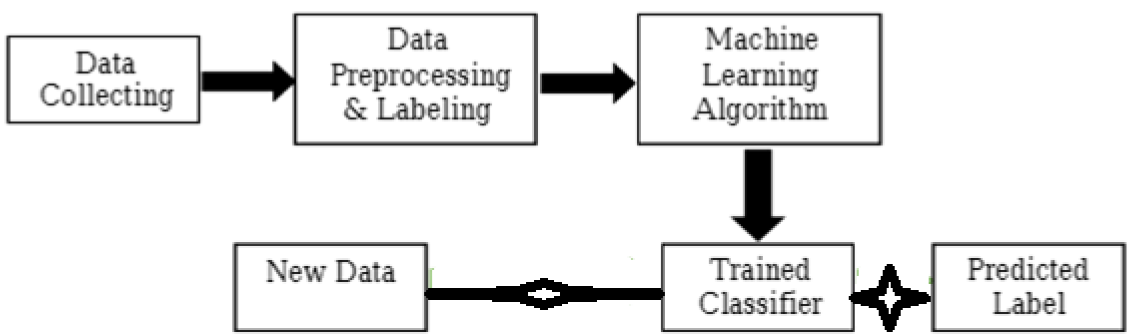


Figure 1. Steps for Predicting Student Performance

3.1 Data Collection

The data collection is the preliminary step of the research work. The methodology is applied to a genuine data containing information about the students of various Department of Computer science streams during the academic years ranging from 2019 to 2022 through questionnaires. The questionnaires that are built on Google Forms are used to conduct a survey on students from Nallamuthu Gounder Mahaligam College, GVG College and STC College, Coimbatore district, Tamilnadu. Around Thirty MCQ type of psychometric response scale in which responders specify their level of agreement to a statement typically in five likert scale points: (1) Always (2) Often (3) Sometimes (4) Rarely (**5**) Never are designed. A total of 230 questionnaires were completed after combining the CSV files from Google Forms. The research sample (230 answers) represents an acceptable sample and the data collected are further processed by Machine Learning tools and Statistical tools.

3.2 Data Preprocessing

It is the process of selecting interested or relevant variable/attribute for making model either manually or automatically. In this work, the real time dataset collected from the students are transformed from raw data into an understandable format. Most of the real time data are often inconsistent, incomplete and lacks in certain behaviors and also contain many errors. Data preprocessing is a proven method of undertaking such issues. The unwanted information is deleted and the texts are converted into numerical values for making it suitable for machine learning model. As mentioned above in 3.1, the options selected from the students such as Always, Often, Sometimes, Rarely and Never are converted into numerical values as 5,4,3,2 and 1 respectively. The below table is a sample shown after preprocessing where the data transformation has been done converting the text into numeric values.

Particulars	**No. of Students**	**Percent**
Area of Residence		
Rural	159	69.10

Semi-Urban	28	12.20
Urban	43	18.70
Gender		
Male	93	40.40
Female	137	59.60
Age		
18 to 20	175	76.10
21 to 23	55	23.90
Marital Status		
Married	9	3.90
Unmarried	221	96.10
Graduate Studying		
UG	179	77.8
PG	51	22.2
Parents Education Qualification		
No Formal Education	4	1.70
Up to School level	201	87.40
UG	16	7.00
PG	5	2.20
Others	4	1.70
Family income per month		
Up to Rs.30000	152	66.10
Rs. 30001 to Rs. 60000	48	20.90
Rs. 600001 to Rs. 90000	26	11.30
Above Rs. 90000	4	1.70
State		
Tamil Nadu	223	97.00
Kerala	7	3.00
Total	**230**	**100.00**

Table 1. Demographic Profile of the Students

The Table 1. depicts the demographic profile of the students analyzed using Simple Percentage Analysis Tool. From this table, it was inferred that 69% of the students are from rural area, 59.60% are female students, 87.40% of parent's educational qualification are up to School level and 66.10% of the family income per month is below 30,000. Hence, from this demographic profile it was analyzed that most of the student's higher education may get influenced by the above mentioned factors.

3.3 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. A model by selecting all the features is not pretty good for a predictive model since it might end with getting less accuracy. After doing some feature selection and feature engineering without doing any logical changes in the model code the accuracy might jump to higher which is quite impressive. Machine Learning offers feature selection techniques like univariate, chi square, gain ratio, info gain, correlation and regression. It was found that the high impact attributes that contribute for the performance of the students are selected. Feature selection technique is useful in reducing the dimensionality of the data which is to be processed by the classifier, reducing execution time and improve the predictive accuracy. Here, the Weighted Average Score is calculated for selecting the specific factor influencing the student academic performance and the results are recorded. Also, the correlation Analysis and Multiple Regression Analysis are used to find out the Significance at five percent and one percent level as shown in Table 2 and Table 3.

** Significant at One Percent Level * Significant at Five Percent Level

Variables	r	r^2
Gender	-0.122**	0.015
Age	0.017	0.000
Marital Status	0.025	0.001
Area of Residence	0.005	0.000
Education Qualification	0.064	0.004
Parent's Education	-0.030	0.001
Family income per month	0.076*	0.006
Sate of Origin	-0.128	0.016

Table 2. Nature of relationship of selected variables with student's academic performance affecting factors- Correlation Analysis

From the above Table 2, it was found that, out of the eight variables selected for the Correlation analysis, two variables have been found to be significant. Of them, Gender is found to be highly significant at one per cent level. The family income per month is found to be significant at five per cent level.

Model	Unstandardized Coefficients		Standardized Coefficients	T (d.f=221)	Sig.
	B	Std. Error	Beta		
Gender	-2.207	1.277	-0.116	-1.729	0.085
Age	-2.953	2.782	-0.135	-1.062	0.290
Marital Status	0.020	3.235	0.000	0.006	0.995
Area of Residence	0.021	0.792	0.002	0.026	0.979
Education qualification	3.733	2.855	0.166	1.307	0.192
Parent's Education	-0.686	1.134	-0.041	-0.605	0.546
Family income per month	0.679*	0.834	0.055	0.814	0.016
Sate of Orgin	-6.657	3.611	-0.123	-1.844	0.067

*Significant at Five per Cent level

Table 3. Factors Affecting Students Academic Performance - Multiple Regression Analysis

Constant	:	85.011	R Square	:	0.044*
Standard Error	:	8.724	Adjusted R Square	:	0.009

The results of regression analysis are consolidated in the above Table 3. Of the eight variables introduced, only one variable is significantly influencing the student academic performance associated with all the six important factors considered such as Personal Factors, Exam Related Factors, Department/Staff Related Factors, Extra-Curricular Related Factors and Social Media Related Factors.

3.4 Model Building

The available Machine Learning (ML) algorithms can be applied to almost any problem. ML learns from the past and tries to capture the possible knowledge from the past and to make accurate decision making for the future. The Student data set are divided into two subsets namely training set and test set. Training set is a subset to train a model and a test set is a subset to test the trained model. In this section, a brief review of the machine learning algorithm and the techniques that are used in this research are introduced.

3.4.1 SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. In Machine Learning, it is mainly used for classification problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

3.4.2 Naïve Bayes

Naïve Bayes classification model is considered as the simplest variation of the Bayesian network. This model assumes that every feature attribute is independent from the other attributes given the target attribute state. Each instance x in the dataset contains attribute values $a1$,$a2$,…,ai. The target function f(x) equals any value from predefined finite set V=($v1$,$v2$,…,vj). Naïve Bayes model uses the following equation.

$$V_{max} = \underset{v_j \in V}{Max}\ P\ (\ v_j\ )\ \prod_i P\ (a_i | v_j)$$

Where v represents the target of the model, P (ai|vj) and P (vj) could be found by calculating their frequencies in the training dataset.

3.4.3 KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

3.5 Train the Model

There are two different phases namely Testing and Training Phase. The data is splitted into training dataset (60%) and test dataset (40%). The suitable classification and regression algorithm which is suitable for the task is examined. Two or more classification algorithm is combined to build Hybrid Model Classifier. The model which is built is tested for the prediction with new test data and the output is produced. The produced output is clubbed with fresh test data and the output is generated. The newly generated training data is used to create an innovative hybrid model using supervised machine learning model. The academic performance of unknown student dataset is predicted using this new model.

3.6 Evaluate the Model

Mean squared error, Root mean squared error, Mean absolute error, Mean absolute percentage error are the metrics used to evaluate the performance of classification and regression model. These are also used to find how the model fit for the dataset which are chosen for the research work.

3.7 Tuning the Parameter

Tuning is the **c**hoosing of parameter that will optimize the model which would be designed. Tuning process is different for different algorithm. Run the model with the data; compare the predicted result with the actual value. The accuracy is evaluated and adjusts the parameters until the best value fits.

3.8 Make Prediction

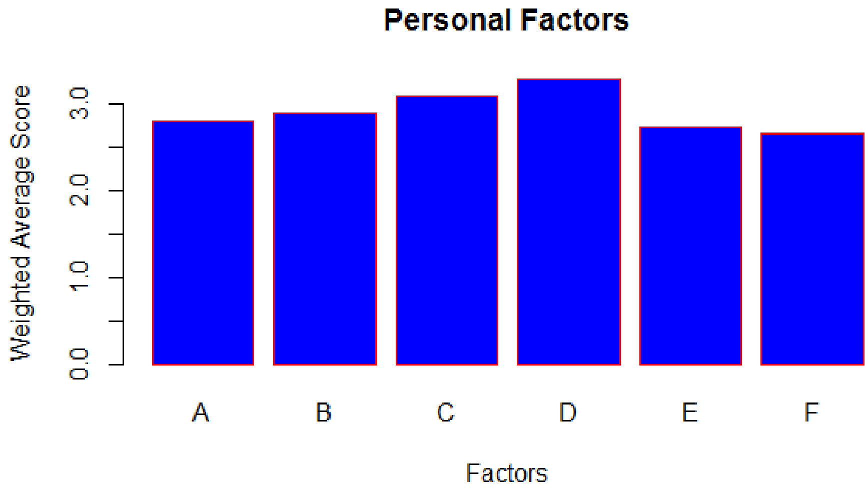
Early prediction of students likely to fail in particular subject and have to take necessary action to improve based on the feature which affects the student performance. The prediction accuracy of algorithms chosen is then compared.

4. RESULTS AND DISCUSSIONS

It is worth discussing these interesting facts revealed for Predicting students' academic performance. It is widely useful to help the educators and learners improving their learning and teaching process. This is an important finding in the understanding of the dependent and independent factors which influence the performance of the students in the higher Education. The results are discussed based on weighted average score and aggregate weight of different questions under five different factors. Based on the analysis, the results are tabulated below for finding out the indicators that affects the students' academic performance.

I. Personal Factors			
S.No	**Factors**	**Weighted Average Score**	**Aggregate Weight**
1.	Family Stress Shows the Negative Impact on the Student's Academic Performance.	2.81	2.91
2.	Language Barrier to Communicate Influences Student's Academic Performance.	2.89	
3.	Socio Economic Factor have impact on myself while Studying.	3.09	
4.	My Family Income Disturbing me to Study.	**3.28**	
5.	My Gender Plays an Important Role to My Study.	2.73	
6.	My Parents are illiterate, so that it Affects My Study.	2.66	

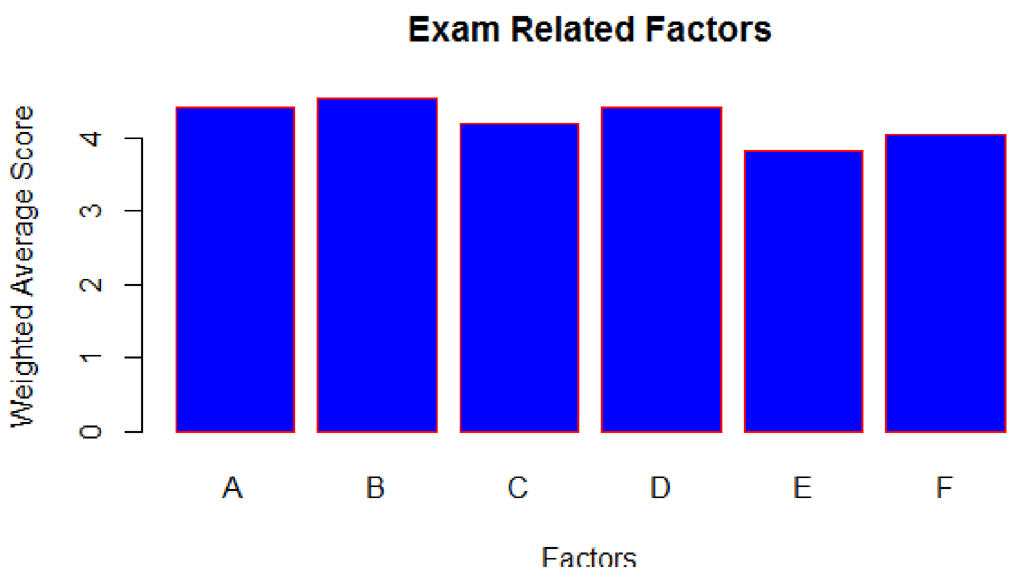
Table 4. Personal Factors



From the above table 4 and the graph, it is proved that among six different personal factors, the family income influences the students acdemic performance sometimes when compared to the other parameters such as family stress, language barrier, socio economic factor, gender and illiteracy of parents.

II. Exam Related Factors			
S.No	**Factors**	**Weighted Average Score**	**Aggregate Weight**
1.	I Make Myself Prepared for the Subject.	4.41	4.24
2.	I Listen to the Lecture of my Teacher	**4.54**	
3.	I Enjoy Home Work and Activities Because they help me improve my Skills in Every Subject.	4.19	
4.	I Exert more Effort When I Prepare for My Exams.	4.43	
5.	I Study the Lessons I Missed if I was Absent for the class.	3.83	
6.	I Study harder to Improve my Performance when I set Low Grade.	4.05	

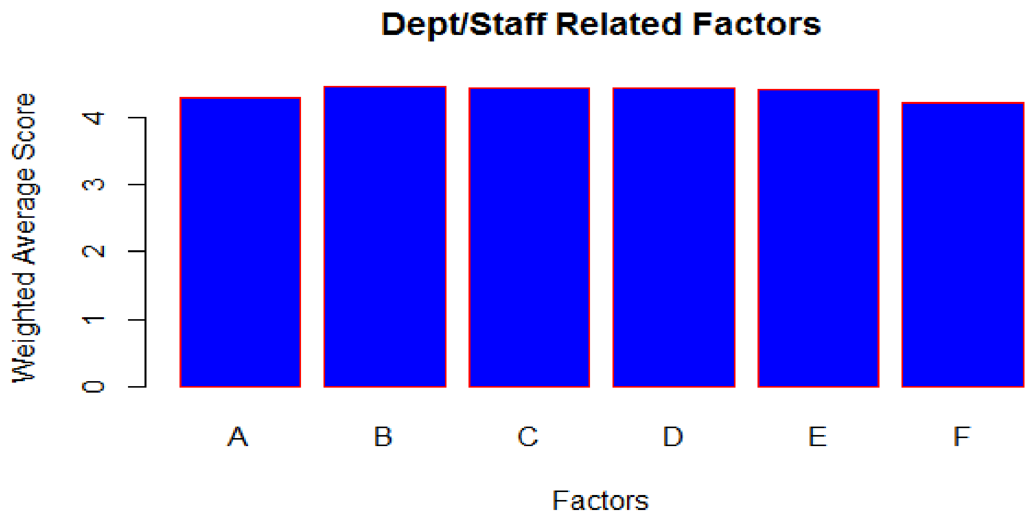
Table 5. Exam Related Factors



From the above table 5 and the graph, it is proved that among the six different Exam Related factors, when the students pay more attention during their leture classes they may get the highest score which improves the students academic performance to the greater level.

III. Department/Staff Related Factors			
S.No	Factors	Weighted Average Score	Aggregate Weight
1.	My Teacher makes me Feel that he/she Cares about me.	4.28	4.37
2.	My Teacher Encourages me to do the Best.	**4.45**	
3.	My Teacher Encourages us to Share Ideas and Options with One Another in Class.	4.42	
4.	My Teacher Explains the Objectives of the Lesson Clearly at the Start of each Period.	4.44	
5.	My Teacher is well Updated With Present Trends Relevant to Subject Matter.	4.40	
6.	My Teacher uses Various Strategies, Teaching Aids / Devices and Techniques in presenting the Lessons.	4.21	

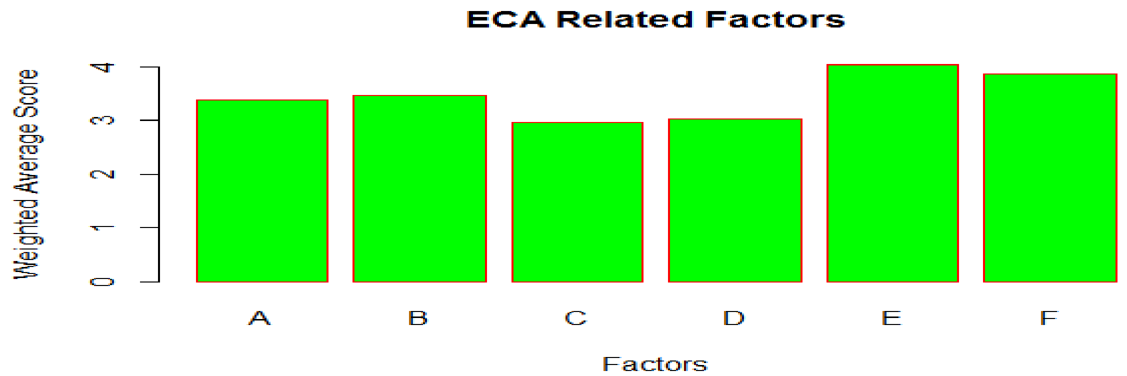
Table 6. Department/Staff Related Factors



From the above table 6 and the graph represented, it is exposed that the department staff members are highly efficient and fulfils the students expectations for their academic performance and no way they are influenced with the factors mentioned above.

IV. Extra-Curricular Related Factors			
S.No	**Factors**	**Weighted Average Score**	**Aggregate Weight**
1.	The Educational Outcomes and Expectations are Associated with Various Forms of Extra-Curricular Activity(ECA).	3.38	3.45
2.	'No Activity At All' has a Greater Negative Impact.	3.47	
3.	Do the Student Engagement to ECA affect student's Grade Point Average (GPA).	2.95	
4.	ECAs are also important for Higher Education Institutions and Form a part of their Public Image, Adding to their Prestige and Reputation.	3.02	
5.	The Students Involved in Extra-Curricular Activities increases Students Success and Persistence to Graduation.	**4.03**	
6.	Extra-Curricular Activity participation will Have a Positive Effect on Student GPA.	3.87	

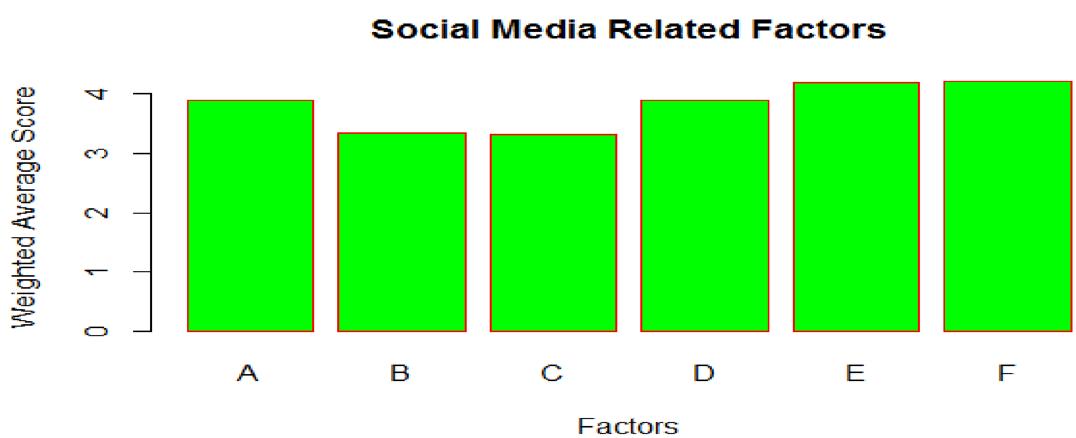
Table 7. Extra-Curricular Related Factors



From the above table 7 and the graph, it is evidenced that the students involved in Extra-Curricular Activities increases students success and persistence to graduation which has positive influence over the academic performance of the students. Also, it is proved that the Student Engagement to ECA rarely affects student's Grade Point Average (GPA).

V. Social Media Factors			
S.No	**Factors**	**Weighted Average Score**	**Aggregate Weight**
1.	Spending More Time on Social Media Sites.	3.90	3.81
2.	Using Social Media to Communicate with Your Teachers.	3.35	
3.	Networking Sites have Created Negative Impact on my Studies.	3.33	
4.	Networking Sites Affect my Study Timings.	3.89	
5.	Social Networking sites can be an effective tool for e-learning.	4.19	
6.	Using Social Media Primarily for Communication with Teachers/Class Fellows.	**4.21**	

Table 8. Social Media Related Factors



From the table 8 and the graph, it is demonstrated that using social media is the primary need for communication with teachers/class fellows. It does not affect the students' academic performance in their higher education.

5. CONCLUSION

The measurement of student achievement across various academic subjects is through the individual's Academic Performance. The art of predicting student performance is exceedingly advantageous to everyone exclusively to the educational administrators and students. In this paper, the previous studies on predicting students' performance with various analytical methods are analyzed. It was stated that, the main reasons for the poor academic performance of college students are outsourced using the machine learning algorithms. The analytical tools are mostly used by the researchers when predicting students' performance as discussed in this paper. When considering the machine learning methods used, the supervised learning method was commonly hired in the research space. Most researchers attempted to predict the performance of students using the students' academic data, social-economic data, and educator competence data as well which was outsourced in this paper. With these analyses, the model will be generated with the training data and further the results will be predicted with the test data using the supervised algorithms mentioned above in this paper to find out the indicators affecting academic performance of the students.

REFERENCES

[1] U. Bin. Mat, N. Buniyamin, P. M. Arsad, R. Kassim, "An Overview of using Academic Analytics to Predict and Improve Students' Achievement: A Proposed Proactive Intelligent Intervention", Engineering Education (ICEED), 2013 IEEE 5th Conference, IEEE, P.No: 126–130, 2013.

[2] B. Cetin, "Approaches to Learning and Age in Predicting College Students' Academic Achievement," Journal of College Teaching & Learning (Online), Vol. 13, No. 1, P.No: 21, 2016.

[3] Jai Ruby, K.David, "A Study Model on The Impact Of Various Indicators In The Performance Of Students In Higher Education", Vol.3, No.5, P.No: 750-755, May 2014.

[4] Dr. T. Johnson, "The Factors Affecting The Academic Performance Of The Indian College Students: A Case Study", Dr.M.G.R PLANETM ***Online*** Journal of Mathematical Sciences, P.No: 1-9, July 2018.

[5] Nisha Arora, Neetu Singh (January 2017), "Factors Affecting the Academic Performance of College Students", i-manager's Journal of Educational Technology, Vol. 14, No.1, P.No: 47-52, April - June 2017.

[6] Rastrollo-Guerrero J, Gómez-Pulido J and Durán-Domínguez, "A 2020 Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review", *Applied Science* **10**, 1042.

[7] G. Siemens, "Learning analytics: The Emergence of a Discipline", American Behavioral Scientist, Vol. 57, No. 10, P.No: 1380–1400, 2013.

[8] S.K. Yadav, B. Bharadwaj, and S. Pal, "Data Mining Applications: A Comparative Study for Predicting Student's Performance," arXiv preprint arXiv: 1202.4815, 2012.

[9] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[10] https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/