Kurdish Studies
May 2024
Volume: 12, No:4, pp. 820-833
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: https://doi.org/ 10.53555/ks.v12i4.3063

Educational Data Mining for Predicting Students' Academic Performance

Irsa Alvi[1*], Prof. Anjum Bano Kazimi[2], Muhammad Asif Alvi[3], Dr. Stephen Swift[4], Saad Ahmed[5]

[1*]IQRA University, Karachi, Pakistan, irsa.alvi@iuk.edu.pk
[2]IQRA University, Karachi, Pakistan, anjum.bano@iuk.edu.pk
[3]Brunel University London, asif.alvi@brunel.ac.uk
[4]Brunel University London, stephen.swift@brunel.ac.uk
[5]NED University, Karachi, Pakistan, saadahmed@iqra.edu.pk

Abstract

Educational Data Mining is an emerging trend that is used for automatically extracting purposeful information from collected data especially in the field of education for predicting future educational achievement or problems. Worldwide higher education institutes have huge number of students' data; Data Mining has a potential to use this data for many purposes. Today's impact of using ICT especially Social Media is taken as an advantage in educational process. This study focuses on identification of various factors which impact the performance of students' learning during academic process, as well as how using Social Media can impact their future achievement. We also want to see how far we can predict about future performance on the bases of this information. This research is quantitative in nature, through purposive sampling students (N=200, size of cohort) of undergraduate's program were selected as respondent from IQRA University of Karachi. The data was collected through structured Questionnaires. For data analysis and prediction Naive Bayes algorithms were implemented through Rapid Miner. On the basis of the responses received through questionnaire and their current Grade Point Average (GPA) prediction were made for their future performance. Results indicated that Social Media use for educational purpose has positive impact on students' academic performance. Several other factors such as mother's qualification, father's qualification and family income also had an impact on students' performance. This research will provide a guide line for stakeholders for predicting future trends and implication on achievement of students and developing proper planning and guidelines for them. This study can spot out those students who require special attention and will help in minimizing the failure by facilitating them to perform well.

Keywords: Data Mining, Social Media, Future Prediction, Rapid Miner

1. INTRODUCTION

Education is an area in which scholars and researchers always try to analyze students' academic and try to find out factors that influence their academic performance. The factors that had been studied were the information technology, social media, students' interest, parental involvement, teachers teaching etc. (Samad, et al., 2019; Ismail, Mohd Erfy, et al., 2019). These studies enhance the guiding and decision-making process, and try to help stakeholders to produce students with excellent academic performance.

In last decade the social media consider to be one of the major factors that highly used by students and impact their educational performance. Social networks have become an integral part of student social life as they serve as platforms for users to interact and relate with their peers (Mingle, Jeffrey, Musah Adams, and E. A. Adjei, 2016). Social networks are now been seen as learning platforms or communities that could be utilized to enhance student engagement and performance. The uses of social media are increasing rapidly all over the world. In recent years, people are shifting from watching television and listening to radios to using social media. Thus, the social media is impacting on human being living styles and on society especially on the students (Mim, Faijun Nahar, Mohammad Ashraful Islam, and Gowranga Kumar, 2018).

In the last two decades, we have witnessed a vast technological advancement in the area of computers and information systems that helps educational institutions to digitalize most of its educational information (Saa, Amjad Abu, Mostafa Al-Emran, and Khaled Shaalan, 2019). Data Mining is a process that sort large data set and help in identifying different patterns and establish relationship between these patterns through data analysis (Al-Saleem, Mona, et al., 2015). In general, this process has an ability to uncover hidden patterns and relations in data that can be used to make predictions and allow different enterprises to predict their future trends. To extract knowledge from this educational data a new discipline known as Educational Data Mining (EDM) emerged; EDM is a one of the application of data mining that is conducted in educational environment (Berhanu, Fiseha, and Addisalem Abera, 2015). Education data mining is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems (Saa, Amjad Abu, 2016).

In Education data mining, classification is the most effective technique to classify and predict values. Therefore, this technique was used in this paper to analyze students' collected information through survey and predict and classify academic performance of students' in their upcoming semester. The objective of the proposed study is:

1. To examine the use of social media for different purpose (Social/Education) and its impact on Students' academic performance.
2. To find out the types of social media used by the students' and its impact on their educational performance.
3. To ascertain how the use of social media has influence the academic performance of computer student.
4. To predict the future educational performance of students on the basis of their previous results.

There are several different classification methods that are used in Knowledge Discovery and data mining. Thus, this paper uses Naive Bayesian mining technique for the extraction of useful information. Hence, result proves that Naive Bayesian algorithm provides more accuracy over other methods like Decision Tree, Regression, Neural networks etc., for comparison and prediction. Moreover, this model also found that variable like family income, father and mother qualification are also correlated with academic achievement of students. This research will provide a guide line for teachers, parents as well as for educational organization for predicting future achievement of students and also highlight the impact of social media on students' academic life which helps in developing proper planning and guidelines for students to control future results. This study is not only beneficial for underperformers but also helps well performers to do more effort and improves their academic achievement.

2. Literature Review

Al-Saleem et.al. (2015) conducted research where they used two of the most recognized decision tree classification algorithms: ID3 and J48 in order to build a performance prediction model based on the grade of previous students in core and elective courses. This model help students in their course selection by predicting student performance in future courses.

Similarly, Devasia, Vinushree T P, & Hegde (2016) conducted analysis on student data and proposed a system by making use of Naive Bayesian mining technique. In Amrita Vishwa Vidyapeetham, Mysuru an experiment was conducted on dataset of 700 students. Data of student's background and previous results were collected and applied to Naive Bayes method to predict student's performance at the start of their semester. This prediction lifts lecturers and students' enthusiasm and helps them to take appropriate actions for their upcoming semesters in order to improve their academic performance.

Saa, Amjad Abu (2016) conducted a research in which they examine multiple factors that affect students' performance in higher education and finds a model which best classifies and predicts the students' performance based on related personal and social factors. They used methods of Decision Tree and Naive Bayes and predict the grades of students from collected data set in order to help both university as well as the students in many ways.

Berhanu & Abera (2015) conducted a study to predict students' academic performance on basis of their previous academic records. They used a decision tree algorithm. Data was collected from the college of Agriculture, Department of Horticulture – Dilla University. The data include five years period and processing was conducted using Rapid Miner. They focused to help students to improve their academic records and give focus on particular courses and enable them to graduate with good grade.

Abu Saa, Al-Emran & Shaalan (2019) conducted a research for investigating students' future academic performance based on dataset extracted from a student information system of a private university in the United Arab of Emirates (UAE). They execute different data mining algorithm on their selected data set in which Random Forest algorithm had the highest Accuracy score over the other algorithm. This model act as an early warning system for predicting failures and low academic performances of students.

Mim, Islam & Paul (2018) conducted a study to investigate the impact of social media on students' academic performance. The data was collected through structured Questionnaire from 345 randomly selected students of Mawlana Bhashani Science and Technology University (MBSTU), Tangail, Bangladesh. The demographic data and educational information analyze by descriptive statistics while a multiple regression model was applied to show the influence of social media on academic performance of students. The findings of their study indicate that large number of students had a negative impact on their academics of using social networking sites for long hours but a portion of students provided a positive result those were using social media for educational purpose. The study suggested that social media should be used in a limited and positive way.

Mingle, Adams & Adjei (2016) conducted a study to examine social media usage and academic performance in public and private senior high schools. The study was focused on finding the activities performed by students on social media and how its effect on spelling during examination, and how participation affected students' grades before and after using social media. The study used the cross-sectional survey method and employed questionnaire to collect responses from two public senior schools and two private senior high schools. The study revealed that student of private schools spent more hours on social media and high proportion of students of private schools experienced low grade as compare to public schools so study suggested that limited and proper use of social media for all students in order to maintain or improve their academic performance.

Baradwaj et al. (2011) conducted research in which present data mining model for higher education system. In this research, they used decision tree method for extract knowledge that predict students' performance in end semester examination. It helps earlier in identifying the dropouts and students who need special attention.

Yadav et al. (2012) conducted a study in which decision tree algorithms are applied on engineering student's data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year. The results help to improve the performance of the students in future. The comparative analysis of the results helped the weaker students to improve in future.

Tariq, Waqas, et al. (2012) conducted a research in which presents impact of social networks on education, students and impact on life of youth. They mainly focused how social networking websites are harmful for youth and how to reduce dangerous impact of social media from youth and teenagers.

Angra and Ahuja (2017) conducted study in which uses data collected form a survey to predict the students' performance, they used rapid miner software to implement various classification methods to analyze and predict students' performance by taking into consideration the behavior of students and also consider students online activities.

Bunkar, Singh et al. (2012) presented paper in which attempt to apply the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. They used data obtained from first year undergraduate university students. The classification rule generation process was implemented which was based on the decision tree as a classification method where the generated rules are studied and evaluated. A system that facilitates the use of the generated rules is built which allows students to predict the final grade in a course under study.

A systematical system on anticipating student performance is proposed by Kurdi, Al-Khafagi et al. (2018) they utilize data mining procedures to enhance students' accomplishments. Association rules and decision trees were utilized to investigate the acquired data and the results could actually help in improving students' success rate and bring the benefits and impacts to students, educators and academic institutions.

A study by Parmar, Vaghela et al. (2015) in which they used a method in which they distribute training and testing task of classification on each Node and central Node respectively, improves classification and prediction task on large and distributed data. To help in decision making, they suggested a method to generalize the information contained in those models, apply specific classification methods were used to generalize these rules to create global model which helps in predicting students' performance.

Singh and Singh (2016) presented a multidimensional Data Model using different types of data sources by using ETL (E-Extract, T-Transform, and L- Load) processes for the creation of multiple data marts which were then analyzed by using data mining techniques to generated Student's semester wise performance as well as a faculty semester wise report.

Another study conducted by Srivastava, Karigar et al. (2018) in which different data mining techniques were utilized to identify reasons for under par student's educational performance, they also adopt data mining techniques as the mathematical foundation for the heuristic process.

Mhetre et al. (2017) conducted a study in which a data was collected from Sardar Patel Institute of Technology College MCA Department and on that data, classification-based algorithms were applied in order to identify best performing student and display it by predictive data mining model. In this study Naïve Bayes, J48, ZeroR and Random Tree using WEKA were used and comparison was made among these four classifiers to predict their accuracy on data. In all data mining classifiers Random Tree performs best. This research will help the institutions to identify students who are slow learners in future which may provide base for deciding special attention to them.

M. Sivasakthi (2017) conducted research into five supervised data mining algorithms such as Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA were applied on the data set to predict programming performance of the students, were evaluated based on their predictive accuracy. Furthermore, a comparison of all five classifiers is also done in order to predict the accuracy and to find the best performing classification algorithm among all. The results indicate that the MLP performs best. This research may help the institutions to identify the students who need special aid in programming it also helps to improve methodology to help students and teachers to improve student's introductory programming performance.

Patil, Rahul, et al. (2018) conducted a study in which develop a system for predicting student performance on basis of their current and previous academic performance. Data mining techniques under Classification was applied on collected data set. The ID3, C4.5, Improved weighted modified ID3 classification algorithms applied on particular data set. In all algorithm Improved ID3 algorithm gives better performance as compared to traditional ID3 & C4.5. This model will provide support to improve academic performance of students in future.

Berens, J et al. (2019) conducted a study in which they develop an early detection system using administrative student data from a state and private university to predict student dropout as a basis for a targeted intervention.

Saarela, M. and Kärkkäinen, T. (2015) present a study in which detect the main courses in order to find out general success in oriented studies.

SUMMARY

In last decade a lot of research was conducted on Educational Data Mining. In these research, researchers mostly developed models using different algorithms like Naive Bayesian, Decision Tree, Regression, Neural networks etc., and applied those models on educational data. In order to find out how Student Academic Performance can be improved in future.

On the other hand, several other studies were done in which researchers examine the impact of Social Media on Students' Academic Performance.

The primary aim of our study is to interconnect these two domains. In this regard a model was developed using Naive Bayesian algorithm in order to predict that how Social Media impact Students' Academic performance positively or negatively through predicting their future academic results.

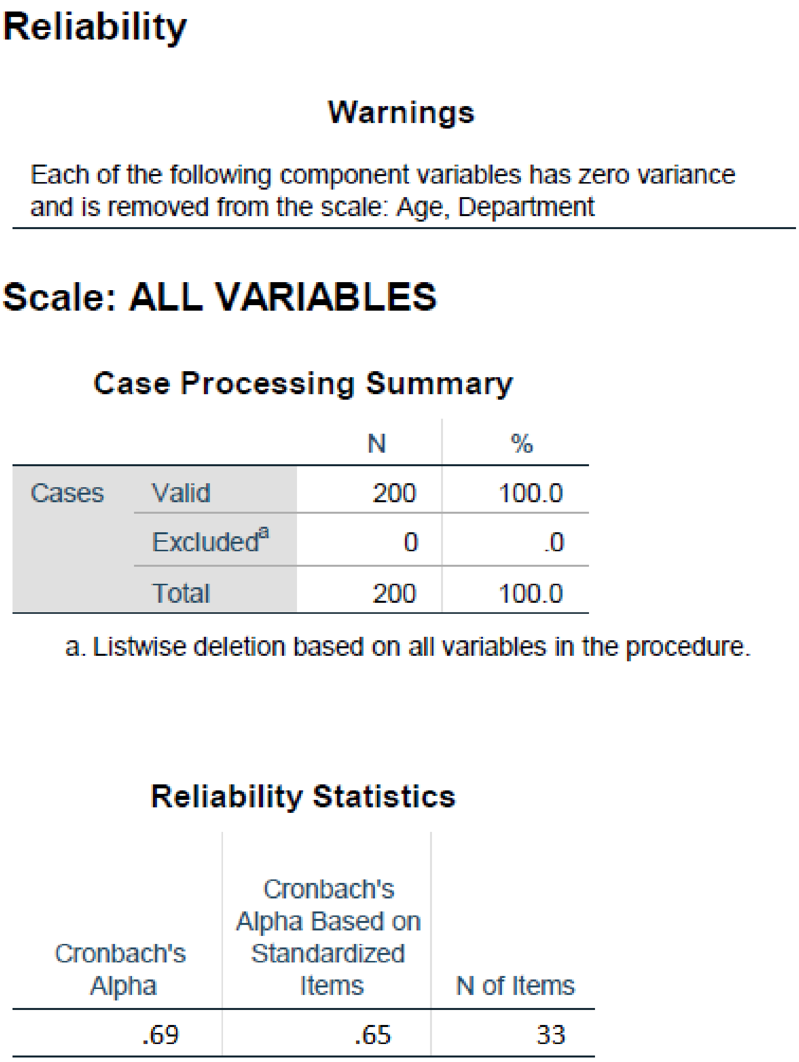
3. METHODOLOGY

In this study our main objective is to use data mining process on educational data and find useful patterns that help students to improve their academic performance in future. The main focus of this study is to determine association between students' education and social media, and how the use of social media impact student academic performance using data mining technique. So, their academic performance could be predicted in the coming semesters. In this regard, a survey was conducted through structured Questionnaire having multiple questions that were related to their use of social media for education purpose, use of social media for social purpose and few were related about their previous academic performances which will later be pre-processed and converted into nominal data so it will be used in the data mining process to discover the relations between the social media and the students' performance. The student performance is measured by the Grade Point Average (GPA), which is a real number out of [0,4]. The data were collected from the students enrolled in computer science department of IQRA University, Karachi, Pakistan.

4. Data Gathering and Conversion

The data was collected through structured Questionnaire, which was divided in to two parts, section 1 was designed for collecting personal information of the respondent and section 2 was designed for collecting data related to the social media usage of respondent. Research question instrument is adapted from the thesis titled "Social Media and Academic Performance of Students in University of Lagos" (Osharive. Peter, 2015). A few questions were modified in this study as the primary objective is to investigate social media use of students for educational and social purpose.

The reliability of Questionnaire is validated by applying Cronbach's Alpha. The result of Cronbach's Alpha is shown in figure 1. According to Keith S Taber the result of Cronbach's Alpha is acceptable (Taber, Keith S, 2018).



Reliability

Warnings

Each of the following component variables has zero variance and is removed from the scale: Age, Department

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	200	100.0
	Excluded[a]	0	.0
	Total	200	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.69	.65	33

Figure 1: Result of Cronbach Alpha

The sample size of students (N=200) of undergraduate's program of computer science department were selected as respondent which are from IQRA University, Karachi Pakistan. After data collection only, those fields were selected which were deemed necessary for our research study and for the data mining process. Table 1 shows a list of those attributes and their possible values.

Table 1: Attributes and Possible Values

S.No	Description	Possible Values
Section A		
	Name	
	Institute	
1	Gender	Male, Female

S.No	Description	Possible Values
2	Age	18 - 25, 26 - 35, 36 - 45, Above 45
3	Qualification	Intermediate, Undergraduate, Graduate, Postgraduate
4	Mother Qualification	Intermediate, Graduate, Postgraduate, M. Phil \Ph.D.
5	Father Qualification	Intermediate, Graduate, Postgraduate, M. Phil \Ph.D.
6	Family Income	20k - 50k, 50k - 80k, 80k - 120K, Above 120k
7	Daily Social Media Usage	1 - 2 hours, 3 - 4 hours, 4 - 5 hours, Above 5 hours
8	Current GPA	
Section B		
Social Purpose		
9	SP 1 - I spend most of my time in using social media	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
10	SP 2 - I mostly use online networking for my social purpose.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
11	SP 3 - I like doing video streaming for my social use.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
12	SP4 - I usually do online gaming as my social activities.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
13	SP 5 - I am more satisfied with online shopping as compare to personal shopping.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
14	SP 6 - I often use social networking sites for photo sharing	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
Education Purpose		
15	EP 1 - I mostly use E-book for my education purpose.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
16	EP 2 - I like doing video streaming for my educational use.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
17	EP 3 - I often use online networking for my educational purpose.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
18	EP 4 - I usually access university portal for my different educational reasons	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
19	EP 5 - I mostly take help of social networking sites for my assignment preparation	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
Types of Social Media Used and its Impact		
20	IPSM 2 - I mostly use Facebook for my social purpose.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
21	IPSM 3 - I prefer to use Facebook for my educational purpose	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
22	IPSM 4 - I prefer to use WhatsApp for my social purpose.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
23	IPSM 5 - I often use WhatsApp for my educational purpose	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
Influence of Social Media		
24	INSM 2 - Excess use of social media is a problematic issue that affect my academic career.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
25	INSM 9 - Use of electronic messaging application improve my participation in academic group activities	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
26	INSM 10 - Use of electronic messaging application gives me awareness about new development not only socially but also in the field of education.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
27	INSM 11 - In my opinion use of social media by the people around me help me in my educational performance.	1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree and 5 = Strongly disagree
Educational Performance		
28	EDP1 - I Categorize myself as achiever in the range of	10 - 20, 21 - 40, 41 - 60, 61 - 80, 81 – 100
29	EDP2 - In my CP (class performance), I mostly achieve in the range of	10 - 20, 21 - 40, 41 - 60, 61 - 80, 81 – 100

30	EDP3 - In my class assignment, I mostly achieve in the range of	10 - 20, 21 - 40, 41 - 60, 61 - 80, 81 – 100
31	EDP4 - In my class quiz, I mostly achieve in the range of	10 - 20, 21 - 40, 41 - 60, 61 - 80, 81 – 100
32	EDP5 - In my class presentation, I mostly achieve in the range of	10 - 20, 21 - 40, 41 - 60, 61 - 80, 81 – 100

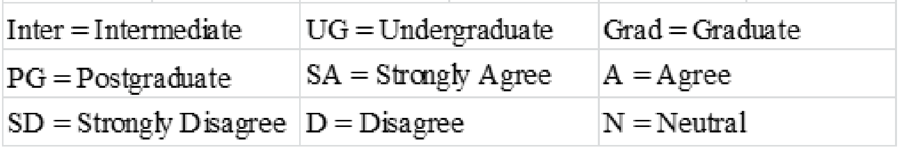
5. Data Analysis

The dataset is initially analysis in statistical manner as well as by graphical presentation so it may help to understand the data before applying complex datamining algorithms on it. Table 2 shows the data ranges in the dataset according to their attributes.

Table 2: Data Ranges in Dataset

Attributes	**Respondent results**				
Age	18 – 25 = 200	26 – 35 = 0	36 – 45 = 0	Above 45 = 0	
Gender	Male = 179	Female = 21			
Qualification	Inter = 3	UG = 197	Grad = 0	PG = 0	
Mother Qualification	Inter = 53	Grad = 117	PG = 20	M. Phil / Ph.D. = 10	
Father Qualification	Inter = 33	Grad = 131	PG = 27	M. Phil / Ph.D. = 9	
Family Income	20K – 50K= 25	50K – 80K = 66	80K – 120K = 64	Above 120K = 45	
Daily Social Media Use	1 – 2 hrs = 69	3 – 4 hrs = 90	4 – 5 hrs = 25	Above 5 hrs = 16	
SP1	SA = 19	A = 55	N = 65	D = 49	SD = 12
SP2	SA = 25	A = 77	N = 50	D = 33	SD = 15
SP3	SA = 13	A = 54	N = 53	D = 52	SD = 28
SP4	SA = 20	A = 59	N = 53	D = 45	SD = 23
SP5	SA = 11	A = 25	N = 62	D = 60	SD = 42
SP6	SA = 42	A = 82	N = 58	D = 16	SD = 2

Attributes	**Respondent results**				
EP1	SA = 47	A = 100	N = 38	D = 14	SD = 1
EP2	SA = 70	A = 89	N = 33	D = 8	SD = 0
EP3	SA = 61	A = 105	N = 28	D = 5	SD = 1
EP4	SA = 67	A = 85	N = 36	D = 11	SD = 1
EP5	SA = 59	A = 81	N = 40	D = 13	SD = 7
IPSM2	SA = 20	A = 89	N = 56	D = 26	SD = 9
IPSM3	SA = 14	A = 40	N = 70	D = 57	SD = 19
IPSM4	SA = 41	A = 89	N = 52	D = 7	SD = 11
IPSM5	SA = 58	A = 84	N = 35	D = 20	SD = 3
INSM2	SA = 40	A = 81	N = 44	D = 29	SD = 6
INSM9	SA = 48	A = 87	N = 43	D = 22	SD = 0
INSM10	SA = 53	A = 77	N = 56	D = 13	SD = 1
INSM11	SA = 33	A = 84	N = 52	D = 25	SD = 6
EDP1	SA = 57	A = 46	N = 44	D = 20	SD = 33
EDP2	SA = 36	A = 47	N = 36	D = 48	SD = 33
EDP3	SA = 53	A = 39	N = 17	D = 36	SD = 55
EDP4	SA = 67	A = 34	N = 22	D = 41	SD = 36
EDP5	SA = 54	A = 28	N = 29	D = 30	SD = 59



Inter = Intermediate	UG = Undergraduate	Grad = Graduate
PG = Postgraduate	SA = Strongly Agree	A = Agree
SD = Strongly Disagree	D = Disagree	N = Neutral

All data mining work which is presented in this paper was performed using the software named Rapid Miner and Naive Bayesian algorithm was used to perform analysis. Analysis is categorized into two categories which are discussed below

5.1.PRIMARY ANALYSIS:

In this study the finding that discover from Primary Variables are fall in category of primary analysis. Primary Variables are basically those variables that are directly related to research topic. Few important findings are discussed below:

Q1. Daily Social Media Usage:

In this variable four options were provided to students which are 1 - 2 hours, 3 - 4 hours, 4 - 5 hours, Above 5 hours.

Findings and Graphical Representation

General observation that observed from this variable is that students' now a days are using social media for at least 2 hours a day. Students' whose social media usage is more than 4 hours are not performing well in their studies and are getting average grades and Students' who are using social media for less than 3 hours are more in number and are performing well as depicted in figure 2.

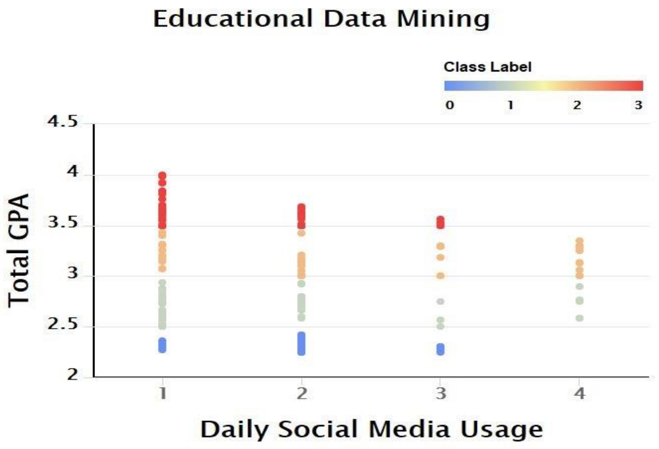


Figure 2: Graphical Representation of Q1 'Daily Social Media Usage attribute'

Q2. Social Purpose

The main intent of this section was to ask students about their social media usage for social purpose. This section includes six questions which were asked from students from SP1 to SP6.

Findings and Graphical Representation

The overall analysis that discover from this section is that Students' who are mostly using Social media only for Social Purpose has no major impact on their performance in anyway as shown in figure 3.

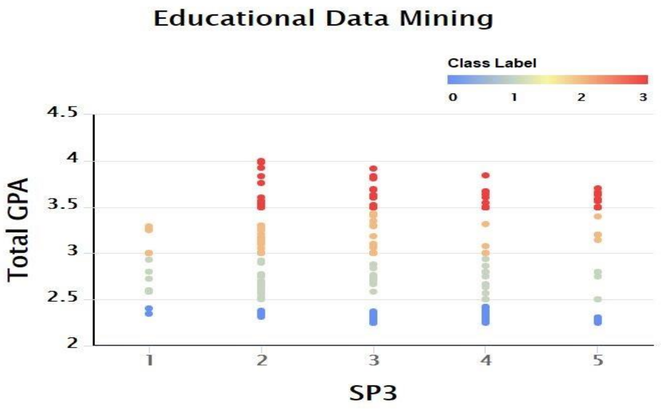


Figure 3: Graphical Representation of SP3

Q3. Education Purpose

The main intent of this section was to ask students about their social media usage for education purpose. In this section five questions were asked from students from EP1 to EP5.

Findings and Graphical Representation

The overall analysis that discover from this section is that Students' who are mostly using social media for educational purposes seems to be getting advantage and have major impact on their educational performance as shown in figure 4.

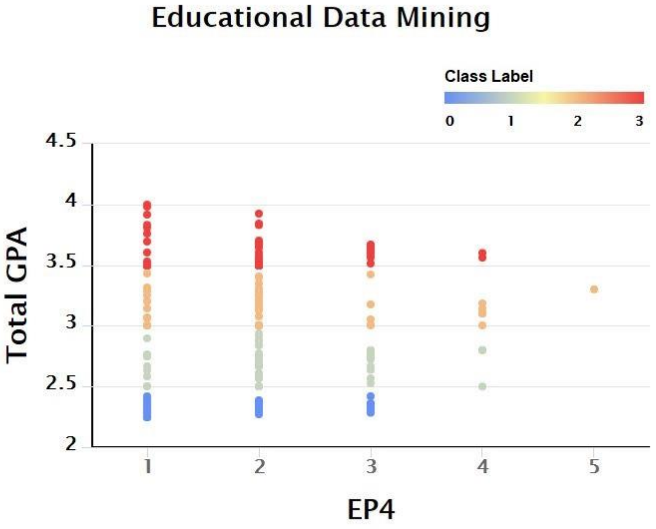


Figure 4: Graphical Representation of EP4

Q4. Types of Social Media Used and its Impact

The main intent of this section was to ask students about their type of social media usage for education and Social purpose. In this section four questions were asked from students from IPSM 2 to IPSM 5.

Findings and Graphical Representation

The overall analysis that discover from this section is that Students' are preferring WhatsApp on Facebook as their Social media tool for educational purpose as shown in figure 5.

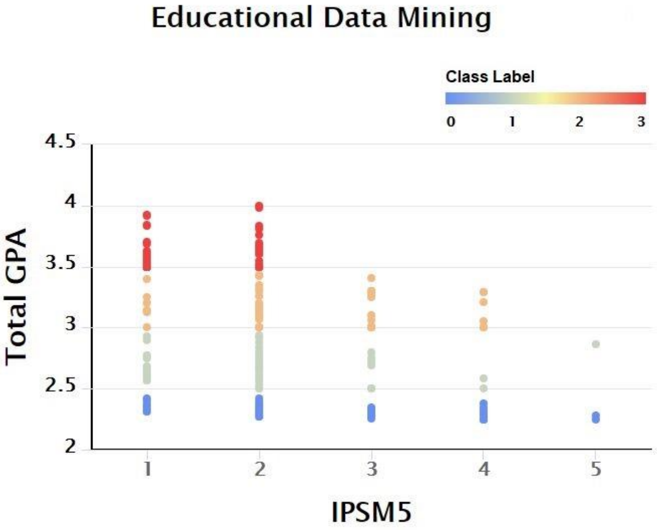


Figure 5: Graphical Representation of IPSM5

Q5. Influence of Social Media

The main intent of this section was to ask students that how social media usage influence their overall performance both for educational and social purpose. In this section four question were asked from students that were INSM 2, INSM 9, INSM 10 and INSM 11.

Findings and Graphical Representation

The overall analysis that discover from this section is that students are claiming that their use of social media help them to improve their awareness in the field of education as shown in figure 6 and 7.

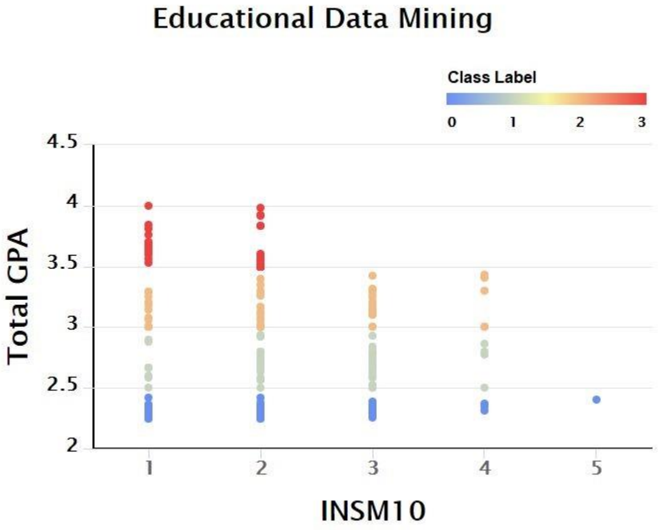


Figure 6: Graphical Representation of INSM10

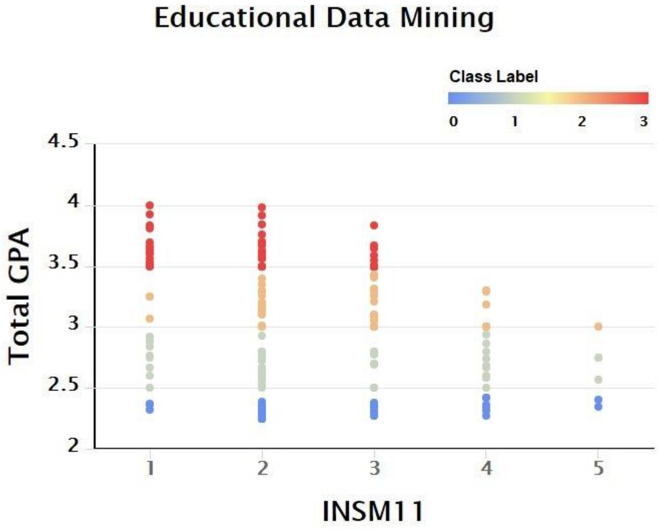


Figure 7: Graphical Representation of INSM11

5.2. SECONDARY ANALYSIS:

In this study the finding that discover from Secondary Variables are fall in category of Secondary analysis. Secondary Variables are basically those variables that are related to student personal information. Few important findings are discussed below:

Q6. Gender:
In this variable two options were provided
Gender: Male=1, Female=2

Finding & Graphical Representation

The overall analysis that discover from this variable is that in computer sciences Male Students' were more in number and there were few female students enrolled in this department. As depicted in figure 8, the level of interest of both genders towards Computer Sciences.

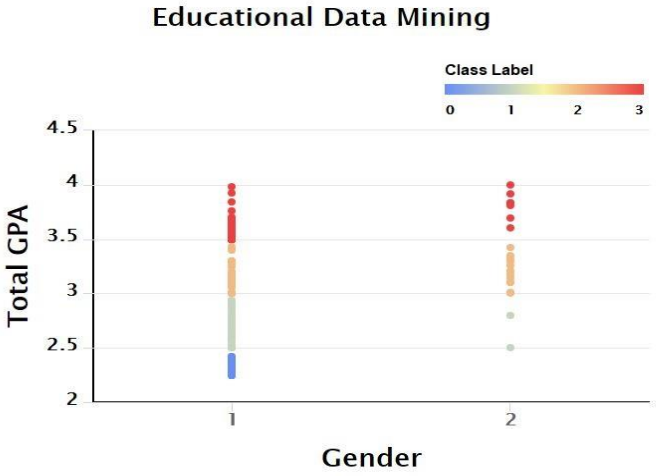


Figure 8: Graphical Representation of Gender attribute

Q7. Mother & Father Qualification:
In both variables four options were provided to students
Intermediate = 1, Graduate = 2, Postgraduate = 3,
M. Phil / Ph.D. = 4

Finding & Graphical Representation

The overall analysis of these two variables are that mother's and father's qualification have an impact on students' performance. Students' perform better if their parents are more qualified shown in figure 9 and figure 10.

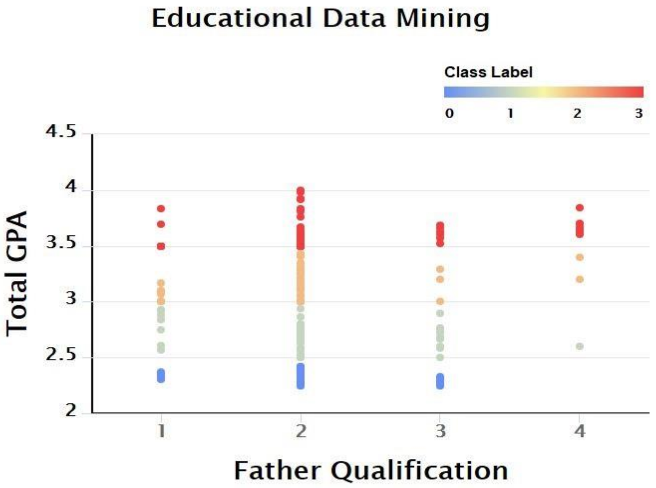


Figure 9: Graphical Representation of Father's Qualification

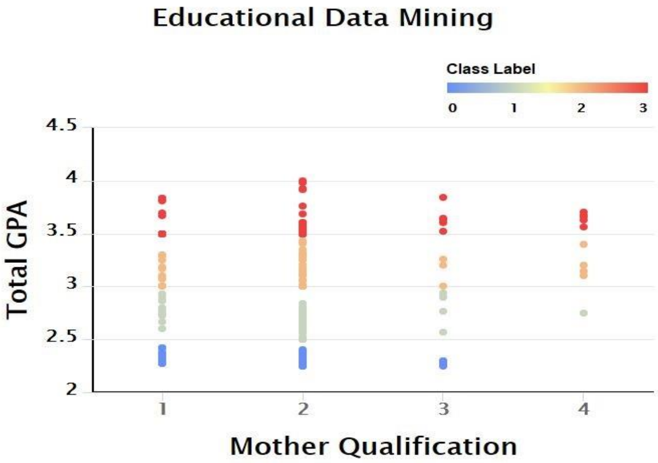


Figure 10: Graphical Representation of Mother's Qualification

Q8. Family Income in Pakistan Rupees:
In this variable four options were provided to students
20k - 50k = 1, 50k- 80k = 2, 80k-120K = 3, Above 120k=4

Finding & Graphical Representation

The overall analysis that discover from this variable is that students with low family income seems to work hard and are getting better grades /GPA as compare to students belong to higher income families depicted in figure 11.

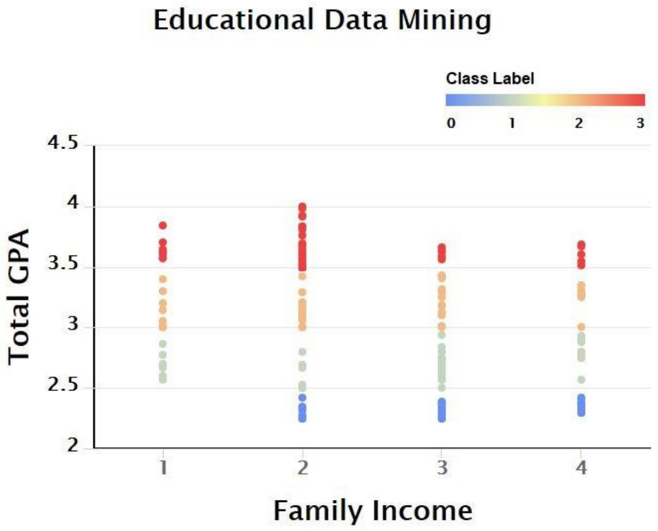


Figure 11: Graphical Representation of Family Income

6. Data Mining Process and Results

There are several well-known data mining techniques such as Classification, Clustering, Artificial Intelligence, etc. Classification is one of the most used and appropriate technique in data mining for predicting future data on basis of previously learned training dataset. There are multiple classification techniques available in data mining, such as, Naïve Bayes Decision Tree, Neural Networks etc. Result provides more accuracy with Naive Bayesian algorithm over other methods like Regression, Decision Tree, Neural networks etc., for prediction and comparison.

7. Model Design and Prediction Result

A 03-fold cross validation was used due to small size of the data to verify and validate the outcomes of the used algorithms and provide accuracy and precision measures. All data mining implementation and processing in this study was done using Rapid Miner. The model design of process that was used in this study is shown below

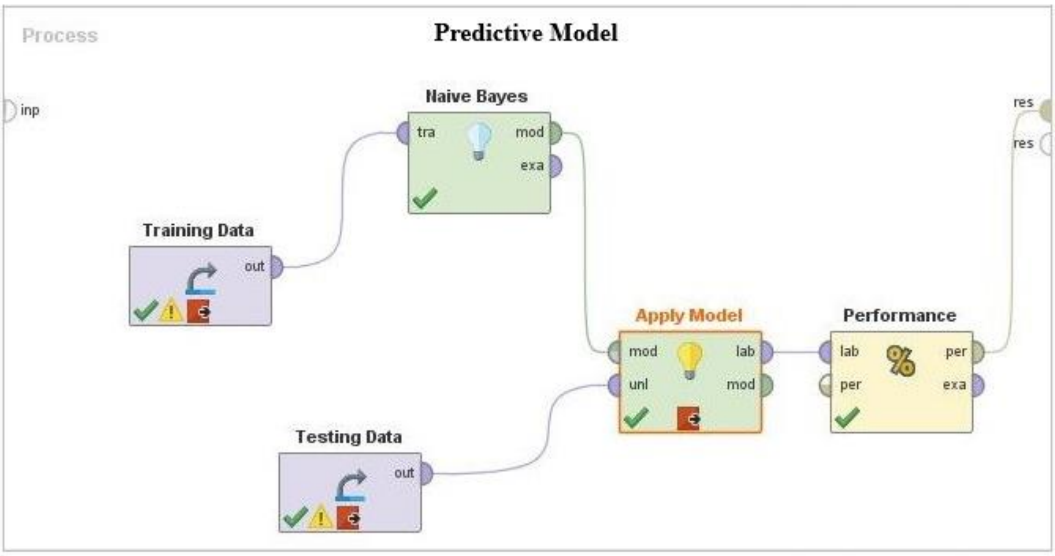


Figure 9: Predictive Model developed in Rapid Miner

8. Result

After running the Naive Bayes algorithm with 03-fold cross validation on data the following confusion matrix were generated. Results of First Set: Accuracy 85.0%

Accuracy: 85.0%					
	True 0	**True 3**	**True 2**	**True 1**	**Class precisions**
Pred 0	11	0	0	0	100.0%
Pred 3	0	10	1	0	90.9%
Pred 2	0	0	7	5	58.3%
Pred 1	0	0	0	6	100.0%
Class recall	100.0%	100.0%	87.6%	54.6%	
Cohen's Kappa = 0.801					

Results of Second Set: Accuracy 90.0%

Accuracy: 90.0%					
	True 2	**True 1**	**True 3**	**True 0**	**Class precisions**
Pred 2	9	3	0	0	75.0%
Pred 1	0	9	0	0	100.0%
Pred 3	1	0	9	0	90.0%
Pred 0	0	0	0	9	100.0%
Class recall	90.0%	75.0%	100.0%	100.0%	
Cohen's Kappa = 0.867					

Results of Third Set: Accuracy 92.7%

Accuracy: 92.7%					
	True 3	**True 1**	**True 0**	**True 2**	**Class precisions**
Pred 3	15	0	0	1	93.8%
Pred 1	0	8	0	0	100.0%
Pred 0	0	0	11	0	100.0%
Pred 2	0	2	0	4	66.7%
Class recall	100.0%	80.0%	100.0%	80.0%	
Cohen's Kappa = 0.899					

The execution in first fold of data set 85.0% accuracy was achieved which shows that the design a Naive Bayesian predictive model for this study predict 85.0% of student results correctly.

The execution in Second fold of data set 90.0% accuracy was achieved which show the percentage of correct predicted result by the design model for the following study.

In third and last validation 92.7% accuracy was achieved on the design model of this study.

This study also discovered that student who are using social media for educational purpose seems to be getting advantage and have a good impact on their academic results. Students who are using social media only for social purpose their performance has not been impacted in any way (shown in figure 4)
Additionally, this study also explored that students are mostly using WhatsApp as their social media tool for their education tasks but for social purpose they are using both Facebook and WhatsApp as social media tool. (shows in figure 5)
During the study it was discovered that students claiming that use of social media are beneficial for their academic as its help them to improve awareness in field of education (shown in figures 6 and 7)
Another finding is that students' academic performance is not completely related to their academic hard work but there are several other factors like family income, mother's and father's qualification that are highly associated with the academic performance of students' and have a great impact on their performance (shown in figures 9, 10 and 11)

9. Conclusion

In this paper, the classification data mining method was used to create predictive model which can effectively use to predict students' academic performance from a collected data. In this study a survey was conducted from University students of undergraduate program. The collected dataset was mainly about students' previous academic performance and their social media usage both for educational and social purpose. Then the collected dataset was preprocessed and data mining task was executed on dataset in order to design appropriate Naive Bayesian predictive model for it. Finally, interesting patterns were found with the design model. This research study discovers that social media use for educational purpose proves to be beneficial for students' academic performance and it was also explored that students' academic performance is not totally related to their academic hard work but there are several other factors like family income, mother's and father's qualification that are highly associated with the academic performance of students' and have a great impact on their performance. This study can help educational institute and students in developing and maintaining a properly planned course of action to achieve better future academic results and it can also help in finding interesting patterns and behavior of students' by apply this model on student's data regularly which may be beneficial for them in many ways. In Future we plan to further this research by acquiring a dataset of students form multiple domains, which we hope will give us more insights about the behavior of students' belonging to different field of studies by applying different classification and data mining algorithms on it to predict their future academic performances.

10. ACKNOWLEDGEMENT

We would like to show our gratitude to IQRA University Karachi, Pakistan for allowing us to collect data for this study, which was collected according to their research ethics protocols. We would also like to acknowledge the support and encouragement Dr. Imtaiz Arif (Academic Director, IQRA University) provided during this study. In addition to this, we also wish to thank all staff members who support us in collecting data. Finally, we would like to thank all participants who took part in the questionnaire.

References

1. Al-Saleem, Mona, et al. 2015. Mining educational data to predict students' academic performance. International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Cham
2. Angra, S. and S. Ahuja 2017. Implementation of data mining algorithms on student's data using rapid miner. International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), IEEE.
3. Baker, R.S., 2019. Challenges for the future of educational data mining: The Baker learning analytics prizes. JEDM| Journal of Educational Data Mining, 1-17.
4. Baradwaj, Brijesh Kumar, and Saurabh Pal. 2011.Mining Educational Data to Analyze Students‟ Performance. IJACSA International Journal of Advanced Computer Science and Applications.
5. Berens, J., Schneider, K., Gortz, S., Oster, S. and Burghoff, J., 2019. Early detection of students at risk-predicting student dropouts using administrative student data from German universities and machine learning methods. JEDM| Journal of Educational Data Mining, 1-41.
6. Berhanu, Fiseha, and Addisalem Abera. 2015.Students' Performance Prediction based on their Academic Record.International Journal of Computer Applications, 0975-8887.
7. Bound, J., Lovenheim, M.F. and Turner, S., 2010. Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. American Economic Journal: Applied Economics, 129-57.
8. Bunkar, K., et al. 2012. Data mining: Prediction for performance improvement of graduate students using classification. Ninth International Conference on Wireless and Optical Communications Networks (WOCN), IEEE.
9. Chandra, E. and Nandhini, K., 2010. Knowledge mining from student data. European journal of scientific research, 47(1), 156-163.
10. Cui, Y., Chu, M.W. and Chen, F., 2019. Analyzing Student Process Data in Game-Based Assessments with Bayesian Knowledge Tracing and Dynamic Bayesian Networks. JEDM| Journal of Educational Data Mining, 80-100.
11. Daniel, B., 2015. B ig D ata and analytics in higher education: Opportunities and challenges. British journal of educational technology, 904-920.
12. Dekker, G.W., Pechenizkiy, M. and Vleeshouwers, J.M., 2009. Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining.
13. Devasia, Tismy, T. P. Vinushree, and Vinayak Hegde. 2016. Prediction of students performance using Educational Data Mining. International Conference on Data Mining and Advanced Computing (SAPIENCE). IEEE.
14. Hofmann, M. and Klinkenberg, R. eds., 2016. RapidMiner: Data mining use cases and business analytics applications. CRC Press.
15. Ismail, Mohd Erfy, et al. 2019 Factors that influence students' learning: an observation on vocational college students Journal of Technical Education and Training.
16. Kurdi, M. M., et al. 2018. Mining Educational Data to Analyze Students' Behavior and Performance. JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO: TICET-ICCA-GECO), IEEE.
17. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. and Loumos, V., 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & Education, 950-965.
18. Maheshwari, E., Roy, C., Pandey, M. and Rautray, S.S., 2020. Prediction of Factors Associated with the Dropout Rates of Primary to High School Students in India Using Data Mining Tools. In Frontiers in Intelligent Computing: Theory and Applications, 242-251.
19. Mhetre, Vrushali, and Mayura Nagar. 2017 Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA. International Conference on Computing Methodologies and Communication (ICCMC). IEEE.

20. Mim, Faijun Nahar, Mohammad Ashraful Islam, and Gowranga Kumar. 2018.Impact of the use of social media on students' academic performance and behaviour change. International Journal of Statistics and Applied Mathematics, 299-302.
21. Mingle, Jeffrey, Musah Adams, and E. A. Adjei. 2016. A comparative analysis of social media usage and academic performance in public and private senior high schools.
22. Moseley, L.G. and Mead, D.M., 2008. Predicting who will drop out of nursing courses: a machine learning exercise. Nurse education today, 469-475.
23. Osharive. Peter. 2015.Social Media and Academic Performance of Students in University of Lagos [Online]. Available: https://www.academia.edu/11356882/social_media_and_academic_performance
24. Parmar, K., et al. 2015. Performance prediction of students using distributed Data mining.International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE.
25. Patil, Rahul, et al. 2018. Prediction System for Student Performance Using Data Mining Classification. Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE.
26. Pekrun, R., Goetz, T., Frenzel, A.C., Barchfeld, P. and Perry, R.P., 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemporary educational psychology, 36-48.
27. Pelaez, K., Levine, R., Fan, J., Guarcello, M. and Laumakis, M., 2019. Using a Latent Class Forest to Identify At-Risk Students in Higher Education. JEDM| Journal of Educational Data Mining, 18-46.
28. Saa, Amjad Abu. 2016. Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications, 212-220.
29. Saa, Amjad Abu, Mostafa Al-Emran, and Khaled Shaalan.2019 "Mining student information system records to predict students' academic performance." International conference on advanced machine learning technologies and applications. Springer, Cham.
30. Saarela, M. and Kärkkäinen, T., 2015. Analysing student performance using sparse data of core bachelor courses. Journal of educational data mining.
31. Samad, et al. 2019.The impact of social networking sites on students' social wellbeing and academic performance Education and Information Technologies.
32. Singh, R. P. and K. Singh. 2016. Design and research of data analysis system for student education improvement (Case study: student progression system in university) International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), IEEE.
33. Sivasakthi, M. 2017. Classification and prediction-based data mining algorithms to predict students' introductory programming performance. International Conference on Inventive Computing and Informatics (ICICI). IEEE.
34. Srivastava, S., et al. 2018. Educational Data Mining: Classifier Comparison for the Course Selection Process International Conference on Smart Computing and Electronic Enterprise (ICSCEE), IEEE.
35. Sweeney, M., Rangwala, H., Lester, J. and Johri, A., 2016. Next-term student performance prediction: A recommender systems approach. arXiv preprint arXiv:1604.01840.
36. Taber, Keith S. 2018. The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 1273-1296.
37. Tariq, Waqas, et al. 2012.The impact of social media and social networks on education and students of Pakistan. International Journal of Computer Science Issues.
38. Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S. and Ragos, O., 2020. Implementing AutoML in Educational Data Mining for Prediction Tasks. Applied Sciences.
39. Yadav, Surjeet Kumar, and Saurabh Pal. 2012.Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal.
40. Yang, D., Kraut, R. and Rose, C.P., 2016. Exploring the effect of student confusion in massive open online courses. Journal of Educational Data Mining, 52-83.