Journal of Advanced Zoology

ISSN: 0253-7214

Volume ***45*** *Issue* ***S-4*** ***Year 2024*** *Page* ***167-173***

Data Mining Techniques To Predict Student Academic Performance In Higher Education: Literature Review

Mrs. Suwarna Mulay[1*], Dr. Shubhangi Potdar[2]

[1]Assistant Professor, BYK College of Commerce, Nashik, Maharashtra, India.*
Email: suwarna.mulay@gmail.com
[2]Associate Professor, DVVPF's Institute of Business Management and Rural Development, Ahmednagar, Maharashtra, India. Email: shubhangipotdar@rediffmail.com

****Corresponding Author:** Mrs. Suwarna Mulay*
Email: suwarna.mulay@gmail.com

Abstract

Educational system is significantly changing in today's world. Recently, the New Education Policy (NEP)-2020 isstarted implementing in India. Students can have various options for getting education according to their choices and requirements. NEP-2020 is more student- centric rather than making them compulsory to get the degree with prescribed syllabus. AI has a major role in NEP-2020. Data mining technology plays a vital role in this new higher education system. As the Higher Education Institutions are growing rapidly, it is necessary for them to impart quality education for enrollment of students. Institutions can maintain educational quality by improving their results. This can be achieved by predicting student academic performance with the help of data mining algorithms.Classification, clustering, regression and association rule mining are the data mining techniques which can be implemented on student dataset to predict the final grade. This study focuses on prediction of student performance using classification and regression data mining techniques. The aim of this literature review is to study various data mining tools, algorithms and the important attributes that affect the student academic performance.

Keywords: Association rule mining, Classification, Clustering, Data Mining, NEP-2020, Regression.

INTRODUCTION

Data mining is the dominant technology which is introduced in the recent years. It is an essential technology used to predict future trends and analyze the big data collected with various online sources. In industries data mining tools are used for decision making and planning their marketing strategies. Data mining technology is an emerging trend in the global competitive business. Recently, this technique is also used in education sector. It plays significant role in the higher education system. Use of data mining technology in education sector is also known as Educational Data Mining (EDM). Educational Data mining is the recent technology which helps the Higher Educational Institutions in their decision making process. It is useful for Knowledge Discovery in the Databases (KDD).As the use of Internet is tremendously increased among the students, huge

amount of educational data is generated over the Internet. Data mining technology extracts hidden pattern from this dataset to predict the future trends.Data mining has various applications in higher education sector such as predicting final grade of the student, student profiling, to improve teachers' teaching performance, admission process, subject selection, strategic decision making for institution etc.Imparting quality education will become more important for organizations to survive in the competitive environment. EDM helps higher educational institutions in strategic decision making by mining the educational data and to improve the educational quality.If the student performance is predicted at early stage of completing the graduation it will helpful for student, teacher and institution to take necessary action for improving the result and achieving the success. EDM is concerned with overall development of student, teacher and educational institution in order to improve their performance and efficiency.

RELATED WORK

Aderibighe & Odunayo (2019), worked on predictive analysis to predict final CGPA of Engineering students based on their performance of first three years of Engineering. KNIME and regression algorithms were applied on the students' data set to predict the output. Dataset of the students of Covenant University, Nigeria was considered for the study. Data was collected from engineering students from seven different streams studying in the years from 2002 to 2014. The maximum accuracy was achieved using Logistic Regression. This research was useful for Nigerian University to improve their results as well as to improve the quality of higher education.

Annaisa & Harwati (2017),included in the article that the educational institutions can maintain the quality of the education by analyzing the students' performance. Proper action can be taken by the institutions for improving the results with the help of effective data mining techniques and predict student academic success at early stage of his graduation. They also proved that the feature selection can increase the accuracy rate of the prediction. Bayesian Network is outperforming Decision Tree since it has higher accuracy rate.

Yulison Chrisnanto et al (2021), In this research paper authors discussed that, the use of data mining techniques can maximize the ability of analyzing student performance, improves teaching and learning process and solve certain learning problems. They had also discussed various data mining algorithms which can be easily implemented in higher education sector. Data mining tools such as Python and RapidMiner were used for the processing.

Sadiq Hussain et al (2018), discussed various classification algorithms to predict student academic performance in higher education. Total 24 attributes were taken into consideration, out of which 2 attributes were deleted in data cleaning process and 12 highly influenced attributes were considered for predicting student academic performance. WEKA data mining tool was used to carry out the research. Various data mining algorithms such as J48, PART, RandomForest, BayesNet etc. were applied on the dataset to predict the academic performance.

Hanan (2020), explained that the early prediction of student academic performance at school level is helpful for universities for future admissions. Identifying academically sound students can become easy with the help of data mining techniques. Decision making for admissions to various courses at university can be made easy by predicting student academic performance using data mining techniques. Algorithms such as Decision tree, SVM, Naïve Bayes, Artificial Neural Network were applied on different data sets of students to predict their academic performance. WEKA data mining tool was used to carry out the research. Artificial Neural Network has highest accuracy in predicting student performance.

Abbas Mhawes et al (2021), analysed the data of engineering students in the mathematics subject. The sample of 84 students of second year of Computer Engineering from Mustansiriya University was considered for the study. Dataset was divided into two groups-training set and test set. Various classification algorithms were applied on both the data sets to evaluate student performance in mathematics subject. The algorithms such as Naïve Bayes, K Star, J48, Multilayer perception, SMO were applied on the dataset. Researchers has used WEKA data mining tool for conducting the research work. K-Star algorithm has maximum accuracy in predicting the performance in mathematics subject.

Sarah Alturki et al (2021), focused on predicting final grade of student at early stage of graduation. In this study researcher has considered pre-enrollment and post-enrollment factors for predicting the final grade. Data mining algorithms such as C 4.5, Simple Cart, LADTrees, Naïve Bayes, BayesNet, ADTree, and RandomForest etc. were applied on the dataset to find the better outcome. 70% of dataset was considered as training dataset and 30% dataset was considered as testing dataset. The results shown that the Naïve Bayes algorithm has higher accuracy with 69.67% in predicting the final graduation grade. The study explained that GPA of secondary school is also one of the most important factor to be considered while predicting the final

grade. Researcher concluded that the orientation year i.e. first year of graduation does not have any impact on student success.

A.Harika et al (2022), explained the procedure for implementation of data mining techniques on student dataset. Author explained that the classification technique was used to build the training dataset and testing was made with the help of testing dataset. Data mining algorithms like C 4.5, Simple Cart, LADTrees, Naïve Bayes, BayesNet, ADTree, and RandomForest were applied on the training dataset. It was analysed that RandomForest algorithm has higher accuracy than the other implemented algorithms. The accuracy is 71%. Students can get chance to improve their performance in future semesters if the student graduation grade is predicted earlier.

M T Sembiring, R. H. Tambunan(2020),discussed that, completion of graduation on time with good performance is important in higher education. Classification techniques of data mining were used in the research. Considering various factors affecting the academic performance, evaluation has been carried out. 483 records were taken as a sample data. The classification technique of Naïve Bayes algorithm was implemented on the data to calculate the student performance. Author proved that Naïve Bayes is one of the effective data mining algorithm to predict student academic performance. The accuracy of this algorithm in the current study is 70.83%.

Leena H. Alamriet al(2020), explained that predicting student academic grade helps institutions to improve the quality of their result. Various attributes were identified which affect the final grade of the student. SVM and Random Forest classification techniques were used to carry out the experiment. Data was collected from the students of Portuguese language and Mathematics subjects. Accuracy was calculated using classification and regression techniques.

Ali Salah Hashim et al.(2020),the study focused on various supervised machine learning algorithms such as DT, NB, Logistic Regression, SVM, KNN, SMO, Neural Network etc. The attributes that were considered for the study are student Number, gender, birthdate, study year, registration, employment, activity point, examination point, final point and grade. Logistic regression shows highest accuracy of 88.8%.

Shiwani Rana, Roopali Garg (2016), two types of machine learning algorithms i.e. supervised and unsupervised were explained in the paper. Student performance was predicted using supervised machine learning algorithms such as Naïve Bayes and Logistic Regression and unsupervised algorithms such as K-mean and Hierarchical algorithm. A brief comparison was made between supervised and unsupervised algorithms to find out the best results.

Ihsan A. Abu Amra, Ashraf Y. A. Maghari (2017), focused on KNN and Naïve Bayes algorithms to predict the performance of secondary school students. The attributes like gender, DOB, specialization, city, secondary school name, status, father's job, student status etc. were considered for performance prediction. Researchers shown that Naïve Bayes algorithm has accuracy 93.17% and KNN has accuracy 63.45%. It was concluded that Naïve Bayes algorithm has more accuracy to predict the next year's performance of the students.

Mrinal Pandey, Vivek Kumar Sharma (2013), four types of decision tree algorithms were discussed in the paper. These algorithms are J48, Simple Cart, RepTree and NBTree. Efficiency of these algorithms was checked with the help of percentage split and cross validation methods. WEKA data mining tool was used for data analysis and performance prediction.Important attributes for predicting the grade was considered. It was found that J48 was the best algorithm to predict the final grade of the student.

K. Govindasamya and T. Velmurugan (2017), have performed analysis of student performances at UG and PG level. Classification algorithms such as C4.5, KNN, Naïve Bayes and clustering algorithms such as EM and K-mean were used to predict student performance. Analysis was made on the basis of accuracy of these algorithms. The researchers concluded that the accuracy of algorithms can be increased by increasing the size of the dataset.

Oladele Tinuke Omolewa et. al. (2019), discussed the use of clustering algorithms in educational data mining. Researchers used k-mean clustering and multiple linear regression algorithms for predicting the student performance. The specific attributes were selected which affect the performance. Student performance prediction model was developed using multiple linear regression.

RESULTS AND DISCUSSION

One of the important applications of Educational Data Mining is student performance prediction. Many studies reported that student performance prediction is useful to improve the result of higher education institutions. It also help the students to get enough time to improve themselves and complete the degree on time.

The analysis of this literature review shows that the performance prediction was primarily carried out at undergraduate level. Very few studies were based on the performance prediction at post graduate level. This study also shows that Engineering College student dataset was generally considered. Limited literatures had used dataset of Arts and Science college students.Student academic performance can be affected by various attributes such as secondary school attended, family background, geographical locations, Family education, Family size etc. Many of these attributes were common in most of the literatures considered here. Table-1 shows the summary of all these literatures studied in this literature review paper. Table-2 shows number of attributes, dataset and sample size used in related literatures and Table-3 shows algorithms, tools and best accuracy of the related literatures.

Table-1: Objectives and Educational levelmentioned in Related Literatures

Reference	Objective	UG/PG
Aderibighe & Odunayo (2019)	Predict final CGPA of Engineering students based on their performance of first three years	UG
Annaisa & Harwati (2017)	To select attributes having higher influence on performance prediction and to apply Bayesian Network and decision tree algorithms on it.	UG
Yulison Chrisnanto et al (2021),	To maximize the ability of performance prediction by applying various data mining algorithms	UG
Sadiq Hussain et al (2018)	To understand how past academic, socio-economic factors and demographic factors affect the performance of students.	UG/PG
Hanan (2020)	Use of student performance prediction in University admission system.	UG
Abbas Mhawes et al (2021)	To predict performance of Engineering students in Mathematics subject.	UG
Sarah Alturki et al (2021)	To predict final grade of the student and to identify honorary students at the early stage.	UG
A.Harika et al (2022)	To train the dataset to prepare a feasible model for predicting student performance with respectable accuracy rate.	UG
M T Sembiring, R. H. Tambunan (2020)	To predict academic performance of the students of Industrial Engineering, USU	UG
Leena H. Alamri et. al.(2020)	Predicting student performance in order to improve the overall result of an organization.	UG
Ali Salah Hashim et. al. (2020)	To compare various machine learning algorithms for predicting student performance and to find out the best accuracy of an algorithm.	UG
Shiwani Rana, Roopali Garg (2016)	To evaluate student performance with classification and clustering algorithms.	UG
Ihsan A. Abu Amra, Ashraf Y. A. Maghari (2017)	To propose student performance prediction model using KNN and Naïve Bayes algorithms.	UG
Mrinal Pandey, Vivek Kumar Sharma (2013)	Predicting student performance to identify weak students and to take necessary action to improve their result.	UG
K. Govindasamya and T. Velmurugan (2017)	To analyse student results at UG and PG degree using classification and clustering algorithms.	UG and PG
Oladele Tinuke Omolewa et. al (2019)	To evaluate student performance using K-Mean clustering and Multiple Linear Regression	UG

Table-2: No. ofAttributes, Dataset and Sample size used in Related Literatures

Reference	No. of Attributes	Dataset	Sample Size
Aderibighe & Odunayo (2019)	6	Engineering, Nigerian University admitted and graduated from 2002-2014	1841
Annaisa & Harwati (2017)	9	University of Islam Indonesia's Information System	178
Yulison Chrisnanto et al (2021),	3	Students of UNJANI 2015-16 and 2016-17	91
Sadiq Hussain et al (2018)	16	Doomdooma College, Duliajan College, and Digboi College of Assam, India	300
Hanan (2020)	3	Computer Science and Information College, Saudi University	2039
Abbas Mhawes et al (2021)	23	Computer Engineering, Mustansiriya University	84
Sarah Alturki et al (2021)	12	College of Computer and Information Sciences, Saudi University	300
A.Harika et al (2022)	8	K.S. School of Computer science and Engineering from 2010 to 2017 batches	662
M T Sembiring, R. H.	8	Student of Industrial Engineering Department, Sumatra	483

Tambunan (2020)		University	
Leena H. Alamri et. al.(2020)	10	Mathematics Dataset, Portuguese Dataset	396 649
Ali Salah Hashim et. al. (2020)	11	College of Computer Science and Information Technology, University of Basre, Academic Year 2017-18, 2018-19	499
Shiwani Rana, Roopali Garg (2016)	7	Third Semester students of BE Digital Electronics, UIET, Chandigarh	First 5 students
Ihsan A. Abu Amra, Ashraf Y. A. Maghari (2017)	14	Student record collected from Ministry of Education Gaza	500
Mrinal Pandey, Vivek Kumar Sharma (2013)	15	Manav Rachna College of Engineering, Faridabad, Haryana, India	524
K. Govindasamya and T. Velmurugan (2017)	20	Private Arts and Science colleges, Chennai, Tamilnadu, India	108
Oladele Tinuke Omolewa et. al (2019)	10	University of California	395

Table-3: Algorithms, Data Mining Tools and Best Accuracy in the related literatures.

Reference	Algorithms	Data Mining Tool	Best Accuracy
Aderibighe & Odunayo (2019)	KNIME, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression	WEKA	Logistic Regression-89.15%
Annaisa & Harwati (2017)	Bayesian Network, Decision Tree	WEKA	Bayesian Network-98.08%
Yulison Chrisnanto et al (2021),	ID3, SVM	Rapid Miner	ID3
Sadiq Hussain et al (2018)	Random Forest, PART, J48, BayesNet	WEKA	Random Forest-99%
Hanan (2020)	ANN, Decision Tree, SVM, Naïve Bayes	WEKA	ANN-79.22%
Abbas Mhawes et al (2021)	J48, Naïve Bayes, Multilayer Perceptron, KStar, SMO	WEKA	KStar – 93%
Sarah Alturki et al (2021)	J48, Simple Cart, LAD Tree, Naïve Bayes, Bayes Net, Random Forest	WEKA	Naïve Bayes 63.33%
A.Harika et al (2022)	Random Forest, KNN, Support Vector, Logistic Regression, Naïve Bayes	SPSS	Random Forest – 71%
M T Sembiring, R. H. Tambunan (2020)	Naïve Bayes	Rapid Miner	Naïve Bayes 70.83%
Leena H. Alamri et. al.(2020)	SVM, Random Forest	Python PyCharm	Mathematics Dataset-SVM- 92.43%, Portuguese Dataset-Random Forest – 94.35%
Ali Salah Hashim et. al. (2020)	Decision Tree, NB, Logistic Regression, KNN, MLP, SMO, Neural Network, PART, JRip, OneR	MS Excel with DM add-in feature, MS SQL Server DM capabilities	Logistic Regression-68.7%
Shiwani Rana, Roopali Garg (2016)	Naïve Bayes, Logistic Regression	Python	Logistic Regression
Ihsan A. Abu Amra, Ashraf Y. A. Maghari (2017)	KNN, Naïve Bayes	Rapid Miner	Naïve Bayes – 93.17%
Mrinal Pandey, Vivek Kumar Sharma (2013)	J48, Simple Cart, RepTree, NBTree	WEKA	J48 – 80.15%
K. Govindasamya and T. Velmurugan (2017)	C4.5, Naïve Bayes, KNN	WEKA, MATLAB	C4.5 – 62.7%
Oladele Tinuke Omolewa et. al (2019)	Multiple Linear Regression	Python	Multiple Linear Regression

Data mining techniques like classification &regression were used by the existing researches to predict the student performance. Supervised machine learning algorithms such as classification and regression were used by most of the researchers to carry out their research work. Table-1 shows that 16 papers in the study had used classification and regression techniques for student performance prediction.

Fig-1 shows graphical representation of frequency of various classification and regression algorithms used in the related literatures of Table-3. This figure shows that Naïve Bayes algorithm was most frequently used for performance prediction in supervised machine learning.

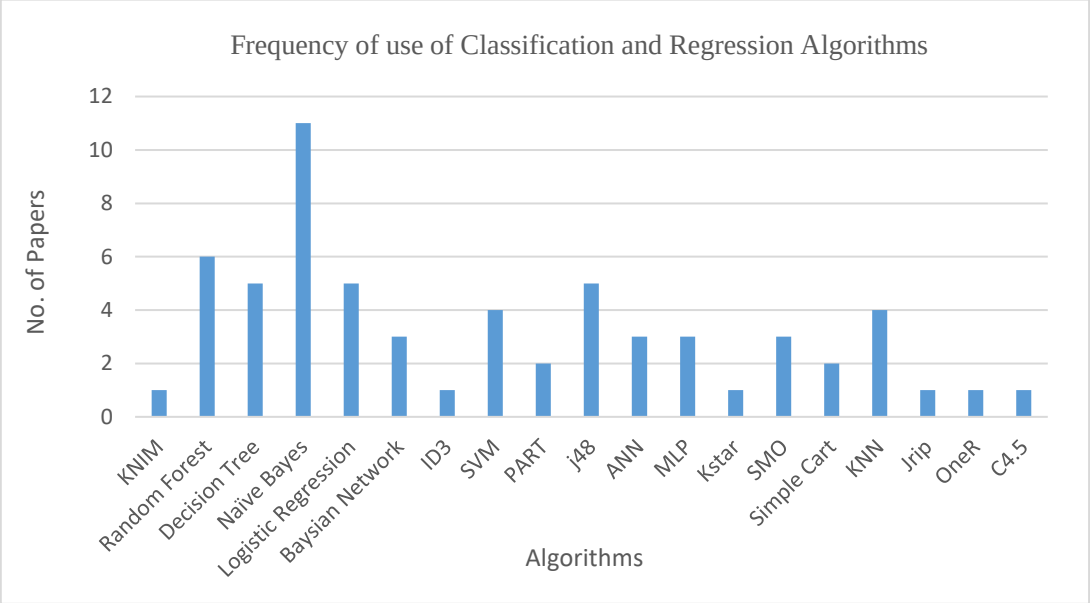

Fig.-1: Visual Representation of Table-1

CONCLUSION

Higher education institutions are the service industries providing educational service to the students. Improving student performance is one of the biggest task for all higher education institutions to maintain the quality of their results which plays important role in NEP-2020. Educational quality can be maintained by predicting student grade at early stage of graduation. Data mining technology helps to achieve this task. This literature review study discusses the use of data mining technology to predict student academic performance. Data mining methods such as classification & regression were mentioned in many important literatures related to educational data mining. The algorithms such as Naïve Bayes, Random Forest, SVM, Logistic Regression, ANN, FCM, EM, K-Mean were used to predict student academic success. Student performance can be affected by various important attributes which are also explained in this paper.Overall, these literature sources provide insights into the use of data mining technology in improving the student academic performance in higher education.

REFERENCES

1. Abbas, Ahmed, Ali and Salman (2021), "Evaluating thePerformance of Engineering's Students In Mathematic Subject based on Academic Decision-Making Techniques", Webology, Volume 18, Number 2, December, 2021
2. Aderibighe Israel Adekitan & Odunayo Salau (2019), "The impact of engineering students' performance in the first three years on their graduation result using educational data mining", ScienceDirect Journal, Vol 5, Issue 2.
3. A Harika, Akshatha K A, Anirudh A, Eesha B S, Prof. Sandhya A Kulkarni (2022), "Student Marks Prediction Using Machine Learning Techniques", International Journal of creative Research thoughts, Volume 10, Issue 7 July 2022 | ISSN: 2320-2882
4. Ali Salah Hashim et al (2020), "Student Performance Prediction Model based on Supervised Machine Learning Algorithms", IOP Conf. Series: Materials Science and Engineering 928 (2020) 032019
5. Annaisa and Harwati (2017), "A Comparative study using student's performance using Educational Data Mining Techniques.", IOP Conference Series 215
6. Hanan Abdullah Mengash (2020), "Using Data Mining Techniques to Predict Student performance to Support Decision Making in University Admission Systems", IEEE Access Volume-8, February 14, 2020
7. Ihsan A. Abu Amra, Ashraf Y. A. Maghari (2017), "Students Performance Prediction Using KNN and Naïve Bayesian", 2017 8th International Conference on Information Technology (ICIT)
8. K. Govindasamya and T. Velmurugan (2017), "A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction", International Journal of Control Theory and Applications ISSN: 0974–5572 © International Science Press Volume 10 • Number 23 • 2017

9. Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi et al.(2020), "Predicting Student Academic Performance using Support Vector Machine and Random Forest", ICETM 2020, December 17–19, 2020, London, United Kingdom
10. Mrinal Pandey, Vivek Kumar Sharma (2013), "A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction", International Journal of Computer Applications (0975 – 8887) Volume 61– No.13, January 2013
11. M T Sembiring, R H Tambunan (2020), "Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)", IOP Conf. Series: Materials Science and Engineering 1122 (2021) 012069
12. Oladele Tinuke Omolewa, Aro Taye Oladele, Adegun Adekanmi Adeyinka and Ogundokun Roseline Oluwaseun (2019), "Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions", Journal of Engineering and Applied Sciences 14 (22): 8254-8260, 2019 ISSN: 1816-949X
13. Sadiq Hussain, Neama Abdulaziz, Ribata Najoua (2018), "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA", Indonesian Journal of Electrical Engineering and Computer Science Vol. 9, No. 2, February 2018, pp. 447-459, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v9.i2.
14. Sarah Alturki, Nazik Alturki (2021), "Using EducationalData Mining to Predict Students' Academic Performance forApplying Early Interventions", Journal of Information Technology Education: Innovations in Practice, Volume 20,2021
15. Shiwani Rana, Roopali Garg(2016),"Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute", 2016 Second International Conference on Computational Intelligence & Communication Technology.
16. Yulison Herry Chrisnanto, Gunawan Abdullah (2021), "The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI)", Matrix: Jurnal Manajemen Teknologi dan Informatika Volume 11 Issue 1 Year 2021