VIEW OF EDUCATIONAL DATA MINING WITH MACHINE LEARNING IN PREDICTION OF STUDENT PERFORMANCE

Shalini Zanzote Ninoria Associate Professor, College of Computing Science and IT, Teerthankar Mahaveer University

Abstract

The growth of any nation strongly depends on the education system of the nation. Drastic changes can be seen in the education system across the whole world now a day. During pandemic it can be observed well recently. There is a major target in educational institutions is to provide quality education for students so that academic performance can be enhanced. Data mining methods and techniques play an important role in the today's world, and it is used for decision making in education system to make decisions related to the students' academic status. Presently, dropping of students has increased with the higher education. This situation is directly pointing to the academic performance fame of the institute. There are various existing systems available with current academic systems but there is no scope of relating intelligence with it or training the models. Data mining has provided a solution to sort out this most challenging problem of the educational problem by analyzing the students' performance using the techniques like linear regression, logistic regression and artificial neural network. Educational Data Mining (EDM) is the field of study concerned with mining educational data to find out interesting patterns and knowledge in educational organizations hence gained impressive attention in recent years. Machine Learning is the most powerful real application of data mining techniques as can The current study is concerned with this recent challenge in the field of academics , specifically, the student performance. In this paper the main focus has centered on detailed study of capacity of EDM in academics and give usage of data mining techniques in the field of EDM with academic performance prediction useful in the student performance measure and faculty performance measure as well so that new horizons can be added in the field of EDM and can be is used to obtain the new style to discover various intelligent paths to predict different semantic knowledge. The study also gives insight to provide and incorporate intelligence in the model using the application of data mining techniques in machine learning..

Keywords: Data Mining; Education Data Mining ; Prediction; Patterns; Machine Learning

Introduction

We are in an age of information. In this information age, we have been collecting tremendous amounts of information. This data presents a great untapped opportunity for knowledge discovery. The most popular area of research i.e. Data Mining for finding hidden facts and to assist policy makers to hypothecate the upcoming scenarios in future days. Data mining emerged in 1990s and has a big impact in business, industry, and science. According to Agarwal et.al.,(1993) defined Data Mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel and understandable patterns in large databases. Fournier et.al., (2017) described its application in various areas such as finance, telecommunications, healthcare, sales marketing, banking, etc. is already well known. The goal of data mining is to predict the future or to understand the past as per Ninoria et.al.,(2017). Agrawal et al.,(1994) stated that Knowledge Discovery in Database (KDD) aims at finding meaningful and useful information in immense amounts of data. Ninoria et.al.,(2019) states that two fundamental issues in KDD, having numerous applications in various domains, are frequent itemset mining (FIM) and association rule mining (ARM).Ninoria et.al.(2020) gives a major application of association rule mining in retails with utility concept in pattern identification. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. Educational data mining (also referred to as "EDM") is defined as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in. Silva et. al.,(2017) presents in study that Educational Data Mining (EDM) is an interdisciplinary research area created as the application of data mining in the educational field. It uses different methods and techniques from machine learning, statistics, data mining and data analysis, to analyze data collected during teaching and learning. Further the present research comprise four sections begin with some literature in related work and later in the next section problem definition have discussed. The next section holds open research issues along with the applications of the study and lastly conclusion can be seen for further improvement.

Statement of the Problem

The major goal of EDM is to study the learning theories to provide the establishment for the selection of instructional strategies and allow for reliable prediction of their effectiveness to determine the current knowledge level of students and academic scenario and give reliable prediction of their effectiveness in teaching, learning and assessment methods for upcoming digital era of academics. Predicting student learning performance is a problem that maps student information to his/her grades. Usually, this problem could be formalized into major application domain of data mining in the context of machine learning problems, i.e., clustering, classification, and regression, ensemble techniques.

Clustering : Clustering can be defined as the identification and classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) so that the data in each subset (ideally) share some common trait of similar classes of objects. The work of Oyelade et al.(2010) formalized student performance prediction into a cluster problem, where the students are grouped into multi-clusters G = {g1...gk}, where k is the number of clusters and then the objective student’s performance is predicted in the specific cluster.

Classification: Classification models describe data relationships and predict values for future observations. Classification is the task of learning a target function that maps each attribute set X to one of the predefined class labels Y. There are different classification techniques, namely Decision Tree based Methods, Rule-based Methods, Memory based reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines. In classification Swami et.al.,(2012) has used test data to estimate the accuracy of the classification rules. If the accuracy is acceptable, the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. In academic performance prediction the output can be using various classification models will be discrete grades for a student or academician. Al-Barrak et al.,(2016) studied and evaluated the "if-then" rules to improve prediction accuracy in the higher education system. Based on different feature selection methods and pruning rules, the DT model has three main algorithms, i.e., ID3, CART, and C4.5. Baker et al.,(2014) compared the three DT algorithms. They carried out experiments to seek the best one. Among these DT algorithms and other machine learning algorithms, the DT showed a higher precision on their used data set. Silva et al.,(2017) investigated the decision tree and the Bayesian Network to predict the academic performance of undergraduates and postgraduates from two academic institutions. In their experiment, the accuracy of the DT is always 3–12% higher than the Bayesian Network.

Regression: Ranadive et.al.,(2014) has adapted Regression techniques for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining, independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The regression model is a function that represents the mapping between input variables and output variables. The regression problem is equivalent to function fitting: selecting a function curve to fit the known data well and predict the unknown data well. Polyzou et.al.,(2016) investigated that in SPP, regression techniques are often used to predict the continuous scores of students in specific courses.

Neural Networks : Neural Network used gradient descent method based on biological nervous system having multiple interrelated processing elements. These elements are known as neurons. Rules are extracted from the trained Neural Network to improve interoperability of the learned network. To solve a particular problem NN used neurons which are organized processing elements. Neural Network is used for classification and pattern recognition. An NN changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase. In NN multiclass, problem may be addressed by using multilayer feed forward technique, in which Neurons have been employed in the output layer rather using one neuron. Many researchers collected these features from student' self assesses by using ANN models to predict student's performance. Researchers asked for student' comments per lesson to reflect their learning attitude and understanding degree of course content and learning difficulty. With this data, Sorour et al.(2017) conducted experiments with the Latent semantic analysis (LSA) technique and ANN model, achieving an average prediction accuracy of about 82.6%. Luo et al.(2015) employed Word2Vec and ANN to predict student grades in each lesson based on their comments .

Support Vector Machines : Zhou et.al.,(2016) performed SVM which splits the data by seeking the maximized margin between two classes. Due to SVM's powerful capability of classification, it has been investigated many times for student performance prediction studies or used as a baseline method. According to psychology, the behaviors potentially affect the student evaluation. Xu et al. divided students into three categories based on the detailed records of learning activities on MOOCs platforms, i.e., certification earning, video watching, and course sampling according to Xu et.al.,(2016). Zhao et.al.,(2022) gives a predictor based on SVM to predict certification obtaining .However, SVM suffers from computation cost in big data due to its optimization limitation.

Objectives of the study

- To classifying and predicting students' performance, dropouts as well as teachers' performance which can help educators to track academic progress to improve the teaching process, it can help students in course selection and educational management to be more efficient and effective.
- To determine a student's current knowledge levels and predicting a student's future success and provide novel guidelines for educational assessment.
- To provide extensive study in the field of educational data mining to give roadways to various future researchers.

Review of Literature

Educational Data Mining has coined with the vision to handle the challenges occurring continuously in the current rapid growing digital era. Many researches as taken up the field and contributed well in the same. Alfiani et. Al.,(2015) states that in the field of academics Data Mining has also been used to analyze the academic performance including faculties teaching performance and specially the student performance. There have been several investigations made under this field .Many researches has utilized data mining techniques in machine learning domain for the prediction in the vision to of next decade. Some of the major noteworthy contribution can be seen for example, Gulati and Sharma (2012) has also claimed that "Data Mining" in education can improve the academic system in different parameters such as orientation, student performance and organizations management. Kumar et. al.,(2011) used classification technique which is most trusted technique of data mining and providing many algorithms which are result oriented in machine learning as used the Naive Bayes algorithm for predicting student performance based on selected variables including students notes,seminars,papers,attendance and other which comes out to be total 13 variables and the results were utilized to construct a model which can be used to predefine the students who are at high risk of failure or dropout and hence guidance and counseling program is activated. Ayesha Mustafa e. al.,(2010) stated a study on evaluation of learning in early time as at the beginning and end of the courses. Varghese et. al.,(2010) in their research used clustering techniques in that the "K means" algorithm for 8000 students with five variables used for cluster. The results concludes that there is a strong connect between attendance and student performance. Bresfelean (2007) conducted a study based on students performance based on student results. According to Sun (2010), to monitor and guide a quality education the relationship between assessment and learning is an important tool. The research study of Cortez and Silva,(2008) conducted in Portugal for education system and the results claims that a good and accurate prediction can be achieved using educational data mining also help to improve the management of education in schools and the effectiveness of learning. Noaman and

Al-Twijri e. al.,(2015) published a recent study on entry requirements of the University of Saudi Arabia. Luo et.al.,(2015) states that some studies have also focused on the impact of Moodle used for applying Data Mining in academics.Chen et.al.,(2011) studied that the student learning on digital platforms can be seen as a application of data ming in Sun .Aslam and Ashraf (2014) took clustering algorithm to give a model of student learning. Grafsgaard et. al., (2014) developed a system which was appreciated will utilization of machine learning techniques used in recognizing facial expressions based on frustration or understanding of students in the classroom. They also used algorithms for detecting unspoken behaviors and associated these finding to acquire the knowledge. Seong Jae Lee (2014) has also worked on the model which can be considered as a record for human behavior prediction models.

Research Methodology

This study focused on systematically reviewing different existing noteworthy contributions in the recent field of research in Educational Data Mining studies that were searched from various sources primarily available in peer databases in an attempt to explain the current state of understanding in this field also analyzed the existing research literature and figured out several research gaps which can be preferred open research issues in future. A comprehensive and systematic methodology used to present here to show what research trends can be revealed and what major research topics and open issues are existed in EDM research with special focus on the utilization of data mining techniques in the world of machine learning. Various machines learning techniques view has been shown for reformulation of the problem.

Results and Discussion

The central work of institutions of higher learning is teaching, the focus is to improve the quality of education, but students are precisely the basis for measuring the quality of teaching. EDM in context of machine learning can apply high quality improved concepts at university students' grades, previously unknown effects on student achievement factors of mining, to provide some valuable reference for teachers and administrators in terms of outcome with highly influential conditions to provide the necessary decision support for teaching, student management and policy makers, to better carry out teaching, in order to improve the quality of teaching in Colleges and universities. The outcome will definitely help in the improvement of quality education nationwide.

The significance can be seen in the current scenario can be seen in various existing applications which are practically implemented and many of the researches has attracted to improve and provide more optimized applications. Some of the practical applications are the Recommender System in which combined student performance prediction tasks with recommendation systems to enhance education outcomes by making a personalized educational plan can be seen. Another can be seen as Early Warning System, many researchers study the common problem of dropping out. Bayer et al. studied the structured data of students' social behaviors, e.g., e-mail and discussion board conversations. Many other applications are also an burning issues for the researchers.

Table No.1: Summary Analysis of Machine Learning Techniques in EDM

Sr.No	Techniques and Methods	Limitations
1	Classification & Decision trees	Do not work best for non correlated variables.
2	Linear regression	Cannot work with noisy data.
3	Classification & Clustering in Support Vector Machine	Mostly useful for non- linearly separable data
4	Ensemble Techniques with bagging	Introduces a loss of interpretability
5	Clustering K-nearest neighbour	Cannot perform well if large dataset and high dimensionality with noisy data is there

The above table gives a clear vision of some of the methods which are used in the previous existing systems with some limitations which can be an open research issues for the upcoming researchers.

Conclusion

This study found in the current work definitely give concrete factors which can be considered for the student performance prediction as can be used to predict the student's failure risk as well as success risk. The outcome will give assistance in perfect strategies development to the policy makers nationwide. Based on the existing information, it has been noticed that other than the academic data, personal data also affect the success rate. Educational Data Mining giving remarkable impact in academics especially in students performance prediction and various techniques can be used in machine learning for better learning in early age so that the failures or dropout conditions can be avoided

References

1. Agrawal, R., & Srikant, R. (1994), "Fast algorithms for mining association rules", In Proc. 20th int. conf. very large data bases, VLDB,Vol. 1215, pp. 487-499.
2. Agrawal, R., Imieliński, T., & Swami, A. (1993), "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD international conference on Management of data,pp. 207-216.
3. Al-Barrak, M. A., & Al-Razgan, M. (2016), "Predicting students final GPA using decision trees: a case study", International journal of information and education technology, 6(7), 528.
4. Alfiani, A. P., & Wulandari, F. A. (2015), "Mapping student's performance based on data mining approach (a case study)", Agriculture and Agricultural Science Procedia, 3, 173-177.
5. Al-Twijri, M. I., & Noaman, A. Y. (2015), "A new data mining model adopted for higher institutions", Procedia Computer Science, 65, 836-844.
6. Aslam, S., & Ashraf, I. (2014), "Data mining algorithms and their applications in education data mining", International Journal, 2(7).
7. Ayesha, S., Mustafa, T., Sattar, A. R., & Khan, M. I. (2010), "Data mining model for higher education system", European Journal of Scientific Research, 43(1), 24-29.
8. Baker, R. S., & Inventado, P. S. (2014), "Educational data mining and learning analytics",In Learning analytics,Springer, New York, NY,pp. 61-75.
9. Bresfelean, V. P. (2007, June), "Analysis and predictions on students' behavior using decision trees in Weka environment", In 2007 29th International Conference on Information Technology Interfaces ,IEEE, pp. 51-56.
10. Chen, S. S., Huang, T. C. K., & Lin, Z. M. (2011), "New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports", Journal of Systems and Software, 84(10), 1638-1651.
11. Cortez, P., & Silva, A. M. G. (2008), "Using data mining to predict secondary school student performance".
12. Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., & Le, H. B. (2017), " A survey of itemset mining", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1207,74,.
13. Grafsgaard, J., Wiggins, J., Boyer, K. E., Wiebe, E., & Lester, J. (2014), " Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue.", In Educational Data Mining 2014.
14. Gulati, P., & Sharma, A. (2012).,"Educational data mining for improving educational quality", Int. J. Comput. Sci. Inf. Technol. Secur, 2(3), 648-650.
15. Kumar, V., & Chadha, A. (2011), "An empirical study of the applications of data mining techniques in higher education", International Journal of Advanced Computer Science and Applications, 2(3).
16. Lee, S. J., Liu, Y. E., & Popovic, Z. (2014), "Learning individual behavior in an educational game: a data-driven approach", In Educational Data Mining 2014.

17. Luo, J., Sorour, S. E., Goda, K., & Mine, T. (2015), "Predicting Student Grade Based on Free-Style Comments Using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons", International Educational Data Mining Society.

18. Ninoria, S. Z., & Thakur, S. S. (2017), "A Survey on High Utility Itemsets Mining", International Journal of Computer Applications, 975(8887), 44-50.

19. Ninoria, S. Z., & Thakur, S. S. (2019), "Review On Rare Itemset Mining",International Journal of Computer Sciences and Engineering, NCRTI.

20. Ninoria, S. Z., & Thakur, S. S. (2020), "Review on High Utility Rare Itemset Mining", In Social Networking and Computational Intelligence, Springer, Singapore. Chicago , pp. 373-388.

21. Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010), "Application of k Means Clustering algorithm for prediction of Students Academic Performance",arXiv preprint arXiv:1002.2425.

22. Polyzou, A., & Karypis, G. (2016), "Grade prediction with models specific to students and courses", International Journal of Data Science and Analytics, 2(3), 159-171.

23. Ranadive, F., & Surti, A. Z. (2014), "Hybrid agent based educational data mining model for student performance improvement",International Journal of Modern Communication Technologies & Research (IJMCTR) ISSN: 2321, 850.

24. Romero, C., Ventura, S., & García, E. (2008), "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, 51(1), 368-384.

25. Silva, C., & Fonseca, J. (2017), "Educational data mining: a literature review", Europe and MENA Cooperation Advances in Information and Communication Technologies, 87-94.

26. Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2014, October), "Predicting students' grades based on free style comments data by artificial neural network", In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings (pp. 1-9). IEEE.

27. Sun, H. (2010), " Research on student learning result system based on data mining", IJCSNS, 10(4), 203.

28. Swamy, M. N., & Hanumanthappa, M. (2012), "Predicting academic success from student enrolment data using decision tree technique", Int. J. Appl. Inf. Syst, 4(3), 1-6.

29. Varghese, B. M., Unnikrishnan, A., Sciencist, G., Kochi, N. P. O. L., & Kochi, C. U. S. A. T. (2010), "Clustering student data to characterize performance patterns", Int. J. Adv. Comput. Sci. Appl, 2, 138-140.

30. Xu, B., & Yang, D. (2016), "Motivation classification and grade prediction for MOOCs learners", Computational intelligence and neuroscience, 2016.

31. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021), "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis", Frontiers in Psychology, 12, 698490-698490.

32. Zhao, X. (2022), "Leveraging Data Mining Technique to Enhancing Online Education and Its Efficiency Study", Mathematical Problems in Engineering.

33. Zhou, X., Zhang, X., & Wang, B. (2016), "Online support vector machine: A survey", In Harmony Search Algorithm, Springer, Berlin, Heidelber, pp. 269-278.