Research Article

Survey on Prediction System for Student Academic Performance using Educational Data Mining

Nilesh V. Ingale[a], Dr. M. Sivakkumar[b], Dr. Varsha Namdeo[c]

[a] Asst. Professor in Computer Engineering at College of Engineering and Technology, North Maharashtra Knowledge City Jalgaon (MS) India.
[b] Scientis, Venkatashawara research Thanjuvur, India
[c] Professor in the Department of Computer Science and Engineering at SRK University, Bhopal, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: Grading students' academic performance is a more difficult and challenging work which will help educators to keep track of progress of performance of students. For this the Educational Data mining (EDM) is a most active and demanding research field. It target is to finding of useful information from the educational dataset by using data mining techniques. Most important tasks of EDM are the prediction of the students' performance. Various researchers all around the globe have published research work on prediction of students' performance [1]. EDM plays an important role in the world of business to help the educational institution for the prediction as well as for make necessary decisions depends on the students' academic performance. Now a day's enhancement of students' performance will affects the students' career and also the reputation of the institute [2]. The main aim of this review paper is to explore the methodology developed and used by these researchers and the findings of their research work in uncomplicated and simplest manner. We presented a comparative study on the effectiveness from prediction of student's performance by applying the EDM techniques. The study results make one important conclusion, which showing that the EDM methods are adequately effective for prediction of students' academic performance, and these predictions are useful to make the necessary decisions and actions by management and teachers.[3]

Keywords: Educational Data Mining (EDM), Machine Learning, Prediction Students' Academic Performance

1. Introduction

Analyzing the large volume of data to make useful summarized information may be a complex task for human being. Grading students academic performance is a more difficult and challenging work which will help educators to keep track of progress of performance of students. Enhancement of quality of education for improvement of student performance is more important for all educational institutes. Prior prediction of outcome of students is beneficial, knowing this in advance will affect the student performance [4]. Data mining is the domain which analyzes large repositories of information to extract useful as well as necessary information. Computers can process any type of data such as texts, numbers, images, facts and figures.

Data mining method plays a highly important role in the classification as well as differentiation of education data. This data is used in a properly organized way by data mining approach. This work can be done with analysis depends on the association, patterns, relations between these data so for getting useful information. The prediction in students' performance with high accuracy is more useful because it helps to finding the students with poor academic performance. In universities, enrollment system and academic performance define the student retention. Data Mining can be used within the educational area. This will be useful to increase our knowledge of learning process by recognizing the variables and evaluating them.

Mining of educational data from the educational environment called as Educational data mining [2]. In last decades, Educational data mining is becoming more popular and demanding field, which cause the exceptionally increase in count of researchers [5]. Recently lots of innovative methods are being developed by many researchers. The earlier proposed machine learning and data mining approaches are using by many researchers for deep understanding of data of educational institutes. This educational data is being used for the understanding the performance of a student.

Information exposed by using EDM approach can be used by a various kind of users like management, administrator, teachers, students and all who are connected with the educational field [1]. Basically Educational Data Mining methods includes,

- Classification and Profiling students
- Identify students learning methods
- Finding all students that are taken course together
- Predicting student's academic performance

The main objective is to get knowledge of the helpful patterns and identifying the necessary and functional information from the systems of educational data. These systems are nothing but syllabus management, course management and registration as well as admission system. These systems are more helpful to all students at various steps of educational institutes' likes universities, colleges and schools. Nowadays, the main aim of all educational institutions is the delivery of better and higher education. Recently, many institutes are using newly developed and innovative methods of education for developments in various education fields for making delivery of higher and better education [6].

2. Literature Review

Various studies are developed to improve prediction of student performance. Prediction of students academic performance is attracted various policy makers, researchers, as well as educators for long time.

K. Kasthuriarachchi and S. Liyanage [3] Designed a system which correlate with the students attendance, develop new methods for practical to encourage students to improve attendance and give the progress of students in classroom. This will help to take decision by the academicians, educators as well as parents. Also management of institute gives the financial support by means of scholarships to students from families having lower income. Moreover, Students performance will degraded due to the parent's negligence, therefore parents must monitoring continuous the their children's progress. These parameters are useful for the process of critical decision making in the educational institutes.

Ching-Chieh Kiu [4] designed a model for predicting student academic performance. By conducting analysis author find out the impact or effect of student background, coursework achievement and social activities on the prediction of performance. Author illustrates the significance of student background as well as social activities to be useful for prediction of student performance and also helpful for identifying the weak student. Early prediction by the model is helpful for the students to perform better to enhance the academic performance and teachers to make the teaching learning process more effective by invention in teaching process, which helps to the student's performance.

The research study presented by H. Muhonen et. al. [5] stated that educational discussions is naturally related with students' academic performance in the domain of language physics, chemistry and arts. By performing qualitative analysis explore the communication patterns which identified both teacher-initiated as well as student-initiated discussion learning, together with peer-centred discussions. In the domain of arts language lessons, discussion or teaching were represented by their neutral quality, whereas in domain of the chemistry and physics lesson, is typical high-quality learning. The outcome of study suggests that to improve the student learning, there is need of improvement in both the quality as well as discussions in the classroom.

R.-C. Zhang et. al. [6] done the study to on present teaching learning background through technology mediated learning (TML), The teaching learning method is more affected by the DA(dynamic assessment) approach. This study identifies the two research gaps on students' academic performance which is affected by DA. First, the previous research has more concentrates on pre- and post-test assessment based DA. The usage of Information technology and knowing the effect of DA based on computer for predication of students' performance over time is essential. Secondly, systems are designing based on TML assessment, the reason is that a many students get the support in remote TML. This system having many limitations such as class size is limited, system will not specify the others factors from outside the system which affects the result.

Padmaja Appalla et. al. [7] presents an efficient educational data mining (EDM) technique to enhance the e-learning. The proposed method contains main two modules, namely the server and the client module. Server module is used to read the documents present in the database and made the related knowledge representation from documents. Similarly client module, is retrieved the information based on the requirements of user. The proposed method is assessed from different parameters which are recall, precision, F-measure and recall. The overall results are getting by changing the number of documents, K-size and the keywords, and. The proposed method has generates the excellent result by generating high evaluation metrics, as illustrate by the average values of recall of 1, F-measure of 0.86 and precision of 0.81 for K = 2.

P. Ducange et. al. [8] summarized the actual benefits in the Virtual Learning Environments provided by the techniques for Big Data mining, for the big data authors analyzed all source of information. Research illustrates the privacy and security features enclosed by layered model. Authors highlight the various levels where we can apply the big data mining. He also emphasized the importance of transcending data which is purely numerical to epistemologically interpret new knowledge created by using educational data and awareness to students to provide the actual data which will help to train the systems to produce proper results.

In the research study of Sadiq Hussain et. al.[9] demonstrates the various data mining techniques by using the various tools of visualization. Apriori algorithm is used for association rule mining, to mine the interesting association rules. Three different classifiers selected for the classification and based on their result which consider parameters like classification errors and accuracy; concluded that the neural network gives better performance than other classifier. Neural network perform better on huge dataset and gives the best than other classifier with accuracy of 90.84%.Also, the authors trying to give some meaningful structures of clustering.

In the research Y. Mitrofanova et. al. [10] discussed different methods and tools of conceptual analysis used in EDM, All these methods classified based on the application stages and goals, which provides way to develop the smart education system. Data mining techniques are used to collect educational data, analyze them, visualize it and make it in presentable format is need for smart education systems. Further Outlier detection, data filtration for judgment, text and relationship mining are some area needs to done for development of EDM tools.

Chew Li Sa et. al. [11] developed the model for student performance analysis. This model focuses to develop a system which analyzes the performance of students. Author consider the course "TMC1013 System Analysis and Design" for the study. To make sure the predictions of the performance in the course various data mining technique like classification algorithms are used. The main contribution of the system is that helps to the teachers, academicians for analysis of student performance and finding the weak students' which may be fail in the course.

Ali Daudet. al. [12] developed a model to predict a academic performance of students, which based on learning analytics. . For research work data was collected from undergraduate and graduate courses of various universities of Pakistan. Research work included a attributes which consider a family details like family income, expenditure and their assets. To evaluate attributes like Precision Rate, Recall and F1-score matrices were used. Work used fivefold cross validation. Research work used classifiers like SVM, Decision Trees, C4.5 and Naïve Bayes. Research concluded that SVM gives better result than other classifiers. Proposed model provides some interesting facts like family's income and expenditure affects performance of students, married students performs well than bachelor students.

H. Amado-Salvatierra et. al. [13] presents an innovative framework that provides massive online environment for the learner to engage them for study. The framework is named as Full Engagement Educational Framework (FEEF). This main aim of the designing the framework which uses an Educational Data Mining (EDM) and Machine Learning (ML) techniques to provide the virtual assistant to give the suggestions to increase the engagements and identified the students which requires attentions to increase the performance. The customized notifications which can be from anywhere that is from within virtual environment or outside provided by these innovative framework. By these notifications the learner is more attentive learn the each content.

Jie Xu et. al. [14] proposed a machine learning method for predicting future performance of students' from degree programs. Course clustering method using a latent factor model-based approach was developed to finding related courses for implementation of basic predictors. A progressive prediction architecture based on an ensemble was developed to connect students' performance into the process of prediction. These data-driven methods are used by combining with other pedagogical methods to evaluate students' future performance and this will gives important information for teachers, academicians and advisors that will help to recommend future courses to students.

A. Jain and S. Solanki [15] elaborate the research work on techniques for prediction of student performance. Proposed machine learning technique is designed by parameter refining and selections of proper attributes. Proposed method gives better result in terms of accuracy and other parameters. Random forest classifier used for the classification. For better result multiclass analysis is used which will help students, teachers and administrator to get classification of students based on the students' performance, using this all the necessary action will be taken to improve the students' performance. This will alert the students about their performance in advance so they will improve it.

A. Tripathi et al. [16] the proposed work which uses naïve bayes approach analysis of predictions for the student performance. The comparisons of results are done in terms of accuracy and total execution time of proposed model with the existing systems. This work concluded that prediction analysis for the complex datasets if more challenging task to predict the students' performance. It is evaluated that accuracy of proposed model is high accuracy and required less execution time than the existing model.

Le Hoang Son and Hamido Fujita [17] proposed a new approach for solved the Multi-Input Multi-Output Student Academic Performance Prediction (MIMO SAPP) problem.. The MIMO SAPP defines to predict the student's future academic performance after they enrolled into a university. The new approach is uses a special learning strategy and multiple parameters sets to overcome the limitations of existing systems. Proposed method is defined as MANFIS-S (Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets). To confirm the

performance of the system, the new approach of multiple parameter sets is used which regulate the model with more relevant parameters. Whereas one additional new approach of local and global training for special learning. All the records from the database is used train the random parameter set in global training, which will revise and gain relevant subset of parameters. Finally, Fuzzy KNN is used to identify the group of new record which is introduced in the testing set. From experimental evaluation prove the Efficiency of MANFIS-S against the relevant algorithm by means of accuracy.

P. Kamal and S. Ahuja [18] find out the most challenging issue and try to give the solution. Author highlights the issue is First year results have the major list of drop out students and percentage is very low of first-year students. It noted the various parameters that affect the student's academic. Research's main aim to find the factors has impact on student's academic performance. He used the sample dataset of 480 students of BCA. The results of the research find that attentiveness in class, daily attendance and previous year percentage is major factors along with this some other the influential factors are there such as qualification of parents, their income. It is proved that 20% of identified factors have the significant relation with student's academic performance. By practical evaluation concluded that performance of ID3 better than other Regression analysis.

In the research done by O. Aissaoui et. al. [19], to predict the learning styles of learner, authors demonstrates two different educational data mining (EDM) techniques. Clustering technique is used to make the learner group using the K-modes algorithm. The classification techniques use the result generated by the clustering algorithm applied to a classified as training set to predict new learners' pattern. In this work, comparative analysis is performed for four classifiers using Weka tool. The results generated concluded that the ID3 algorithm gives the better accuracy and the lowest mean error and Kappa statistic, that's why ID3 algorithm is the efficient in classifying correct pattern.

R. Patil et. al. [20] introduced the new system to predict the results of fourth year based on the current and past academic performance of third year students. These data sets of academic performance of third year student have been used to calculate final solution as prediction. Classification data mining technique is used for the clustering the data into predefined classes or groups. As it is a supervised learning method, so to test data should be classify into the predefined classes or groups for that rules should be defined and there is need of labeled training data. In this research comparison between ID3, C4.5 and improved ID3 is done. From the experimental evaluation conclude that Improved ID3 algorithm provides better performance than heuristics algorithm like ID3 & C4.5 algorithm.

I. Shetty et. al. [21] used combination of various techniques to predict the student's performance from the previous data. These techniques consist rule-based learning, classification, neural network based algorithms and ensemble methods. Confusion matrix is used to calculate the accuracy because for the unbalanced data, this matrix is proved as a good metric. When the structure data is considered Ensemble methods work best as compared to neural networks. a large volume of data is required in the neural networks for training. Since authors used data in study is structured and less complex, hence ensemble methods performed better as compared with Multilayer perceptron and neural networks for prediction. This study may be useful to students for improvising the performance and results in future and teachers for finding the weak students for give the more attention to improving the performance and future results.

P. Sokkhey and T. Okazaki [22] presented the comparative study on predication of student performance in mathematics; this study focuses various approaches like machines learning (ML) algorithms, statistical analysis techniques and deep learning architecture. Grading students' academic performance is a more difficult and challenging work which will help educators to keep track of progress of performance of students. For improving the prediction find out enhanced prediction model by reviewing and compared the many existing techniques. These techniques are five different types of machine learning algorithms, statistical technique Structural Equation Modeling (SEM), and deep learning framework; Study executed and compared Deep Belief Network (DBN). Out of all the Random Forest (RF) was performing excellent to predict the student performance.

F. Kaunang and R. Rotikan [23] proposed a new prediction model for students' academic performance, which conducts a comparative analysis performed to compare evaluation of Decision Tree and Random Forest classification techniques using WEKA. Then it is concluded that performance of the Decision Tree is better than Random Forest. The results indicate factors which was affecting the Student's learning behaviour. The knowledge which is gathered from these experiments can be considered while evaluating the learning behaviour of students. Generated results can be useful for future study such as for make better performance add more attributes to dataset; can use other classification techniques for accurate result and comparative analysis.

E.B. Costa et. al. [24] demonstrates the EDM techniques to find out its effectiveness for predict the performance of students. They contribute the research which is differing from other related work. By using this

effective EDM techniques, identify the students who are fails at early stages of courses and then it will help to take the decision to reduce the rate of failure. To make these EDM techniques more effective, research analyse the effect of data pre-processing and tune up the algorithms. Specifically, author conducted this study by doing a comparative study on the four different techniques of EDM namely Support Vector Machine (SVM), Decision Tree, Naive Bayes and Neural Network. Experimental evaluation is performing two data sources taken form Brazilian university. These two data sources are independent data sources of distance education and on campus of introductory programming courses. This analysed EDM approach efficiently predict students' academic failures well in advance, this will help teachers or academicians to take the decision to improve the performance.

K. Rawat and I. Malhan [25] proposed the hybrid approach for classification based on machine learning for prediction of the performance evaluation. In this study four different machine learning algorithms such as IBK, NB, ANN and J48 were used. The evaluation of these methods with hybrid approach showed that classification by hybrid method gives better accuracy than other individual algorithms. Hence, the proposed hybrid approach is useful for the prediction of student's result will he1lp to take actions to improve performance.

3. Summary

Literature survey is done by studying various papers, using the different approaches performed on educational data to predict student's academic performance. Form [1] to [7] shows the generalized study on EDM process. Whereas, Big data Mining and Generalized data mining techniques are studied, [8] to [11]. For machine learning approach to predict the students performance referred [12] to [16]. For prediction of student's performance proposed models using hybrid approaches referred the existing models of [7] to [25].

4. Research gap

We study the various research will states the effects of data preprocessing and fine tuning algorithms on the effectiveness of these approaches. Many researchers performed experiment to provide proof for efficiency of existing EDM methods to in identifying weak students who may fail. It is important to be aware of some threats. All the studied systems work on specific educational data that means system is not producing general result. Only fine-tuning of Naive Bayes techniques and SVM was done automatically. Fine tuning task of algorithm is done manually may affect the efficiency. This task is main threat for systems. [3].

To predict performance of student researchers applied various classification techniques. Some are uses the multi-class classification gives best accuracy and requires a less execution time. For predicting student performance only a few of them consider choosing associated attributes as an important step and parameters for fine tuning of algorithm [5].

5. Conclusion

For improving the prediction find out enhanced prediction model by reviewing and compared the many existing techniques (final). This paper has presented a exhaustive survey of research works on Educational data mining (EDM).This paper reviews several existing researches and identifies other future pathways based on EDM insights. Clustering techniques help identify key variables such as student behavior in class, group learning, time need to spent learning a particular module, the classroom environment, and student motivation, etc. Clustering on EDM provides various useful factors and it can be multilevel nonhierarchical and hence the researchers must carefully choose the algorithm and the variables that result in better and accurate clusters and hence provide useful information.

References

1. V. Sathya Durga, J. Thangakumar (2019), "Students Performance Prediction through Educational Data Mining - An Uncomplicated Review", International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 01, pp. 1404-1406.
2. T. Devasia, Vinushree T P and V. Hegde (2016), "Prediction of students performance using Educational Data Mining," International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, pp. 91-95.
3. Kasthuriarachchi, K. T. S., & Liyanage, S. R. (2019), "Predicting Students' Academic Performance Using Utility Based Educational Data Mining", Frontier Computing, pp. 29–39.
4. C. Kiu (2018), "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, pp. 1-5.
5. Zhang, R.-C., Lai, H.-M., Cheng, P.-W., & Chen, C.-P. (2017), "Longitudinal effect of a computer-based graduated prompting assessment on students' academic performance", Computers & Education, 110, pp. 181–194.

6. F. J. Kaunang and R. Rotikan (2018), “Students' Academic Performance Prediction using Data Mining”, Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, pp. 1-5.
7. Son, L. H., & Fujita, H. (2018), “Neural-fuzzy with representative sets for prediction of student performance”, Applied Intelligence.
8. Ducange, P., Pecori, R., Sarti, L., & Vecchio, M. (2016), “Educational Big Data Mining: How to Enhance Virtual Learning Environments”, Advances in Intelligent Systems and Computing, pp. 681–690.
9. Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018), “Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study”, Cybernetics and Algorithms in Intelligent Systems, pp. 196–211.
10. Mitrofanova, Y. S., Sherstobitova, A. A., & Filippova, O. A. (2019), “Modeling Smart Learning Processes Based on Educational Data Mining Tools”, Smart Innovation, Systems and Technologies, pp. 561–571.
11. C. L. Sa, D. H. b. Abang Ibrahim, E. Dahliana Hossain and M. bin Hossin (2014), "Student performance analysis system (SPAS)," The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), Kuching, pp. 1-6.
12. Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi and Miltiadis D. Lytras (2017), "A Neural Network Approach for Students’ Performance Prediction", Proceedings of International World Wide Web Conference Commitee, Perth, Australia, pp. 415-421.
13. Amado-Salvatierra, H. R., & Rizzardini, R. H. (2018), “An Experience Using Educational Data Mining and Machine Learning Towards a Full Engagement Educational Framework”, Learning Technology for Education Challenges, pp. 239–248.
14. J. Xu, K. H. Moon and M. van der Schaar (2017), “A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs”, in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753.
15. Tripathi, S. Yadav and R. Rajan (2019), "Naive Bayes Classification Model for the Student Performance Prediction," 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, Kerala, India, pp. 1548-1553.
16. Appalla, P., Kuthadi, V. M., & Marwala, T. (2016), “An efficient educational data mining approach to support e-learning”, Wireless Networks, 23(4), pp. 1011–1024.
17. Muhonen, H., Pakarinen, E., Poikkeus, A.-M., Lerkkanen, M.-K., & Rasku-Puttonen, H. (2018), “Quality of educational dialogue and association with students’ academic performance”, Learning and Instruction, 55, pp. 67–79.
18. Kamal, P., & Ahuja, S. (2018), “Academic Performance Prediction Using Data Mining Techniques: Identification of Influential Factors Effecting the Academic Performance in Undergrad Professional Course”, Advances in Intelligent Systems and Computing, pp. 835–843.
19. El Aissaoui O., El Alami El Madani Y., Oughdir L., Dakkak A., El Allioui Y. (2020), “Mining Learners’ Behaviors: An Approach Based on Educational Data Mining Techniques”, Embedded Systems and Artificial Intelligence. Advances in Intelligent Systems and Computing, vol 1076. Springer, Singapore, pp. 655-670.
20. R. Patil, S. Salunke, M. Kalbhor and R. Lomte (2018), "Prediction System for Student Performance Using Data Mining Classification," Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp. 1-4.
21. Isha D Shetty, Dipshi Shetty, Sneha Roundhal (2019), “Student Performance Prediction”, International Journal of Computer Applications Technology and Research (IJCAT) , Volume 8, Issue 05, pp. 157-160.
22. P. Sokkhey and T. Okazaki (2019), "Comparative Study of Prediction Models on High School Student Performance in Mathematics," 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), JeJu, Korea (South), 2019, pp. 1-4.
23. Jain and S. Solanki (2019), "An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning," International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 1457-1462.
24. Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017), “Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses”, Computers in Human Behavior, 73, pp. 247–256.
25. Rawat, K. S., & Malhan, I. V. (2018), “A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining”, Lecture Notes in Networks and Systems, pp. 677–684.
26. G. Barata, S. Gama, J. Jorge and D. Gonçalves (2016), "Early Prediction of Student Profiles Based on Performance and Gaming Preferences," in IEEE Transactions on Learning Technologies, vol. 9, no. 3, pp. 272-284.

27. Romero and S. Ventura (2010), "Educational Data Mining: A Review of the State of the Art," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618.