International Journal of Engineering & Technology, 7 (4.5) (2018) 467-470

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Survey on data mining approach for analysis and prediction of student performance

Chala Simon [1] *, Ybralem Bugusa [1]

[1] *Department of Computer Science and Engineering, Symbiosis International university, Pune, India*
**Corresponding author E-mail: chala.fikadu@sitpune.edu.in*

Abstract

The quality of education is measured by the academic performance of students and the results they produce. Since the student academic performance is the made up of the environmental, psychological, socio-economic and other factors, it is challenging to measure the academic performance of students. Such difficulties can be reduced by investigation of various factors that influence the student perfor- mance. Many researchers have been used different approaches to identifying the variables that help to predict students' performance. This survey paper examines various data mining methodologies that have been used to analyze and predict students' performance.

Keywords: *Student Performance; Data Mining; Prediction; Classification Algorithm.*

1. Introduction

Student Performance is defined as how well the student has performed in class, and how well the student has grasped the offered material. The performance achieved by the students measures the quality of education delivered by educational institutes. A quality education gives students the skills they have to partake as productive, innovative and responsible individuals from society. Besides various factors may influence the quality of education as well as the student motivation and potential. Deeply analyzing and identifying this factor enables to build a new strategy and better deci- sion for the future development of quality education specifically and for nation's future social and economic prosperity in general. To achieve this objective a careful information assessment or data about the institute, teachers, the profile of students and data mining technique are crucial for better decision.

Data mining is the process of identifying relevant patterns from huge data's along with the various machine learning and statistical methods to support future decision-making. Today, due to availability of vast amount of data and the need for changing such data into helpful information and knowledge, data mining has pulled a lot of consideration in research industry and society as a whole. Knowledge Discovery in Databases (KDD) is an-other name of datamining, which is the area of discovering new and potentially valuable information from enormous databases Arockiam, L., et al. [2], When it comes to the education sector, it can provide the tasks that are used to study the student performance, i.e., predic- tion and analysis with the data available about all factors that in- fluence the quality of education. It means the academic perfor- mance of the student isn't an effect of just a single main factor besides it heavily depends on various factors like personal, socio- economic, psychological and other environ-mental variables. Examining and choosing the most relevant and influential variables from this factor is better to know what it will come next.

Despite this, prediction and analysis are an imperative point of reference in an educational environment for improvement of student performance. Student's academic performance is an essential factor in building their future Baker, et al. [3], Tang et al [1].

In this work, authors have summarized various classification approach used to evaluate students' performance using important variables or predictors. The ultimate aim of classification method is to make a possible prediction of main class with a higher accu- racy in given dataset.

2. Literature review

There are a number of datamining techniques used in analysis and prediction based on the dataset. Association rule mining is used to analyse the relationship between dependent variables and independent variables. Classification and Regression models are the most commonly used in predicting the target class from the given data. This literature review is used to explore the important factors and different data mining algorithms that have been used in predicting students' performance.

In Angeline, et al [4], student's assignment marks, class tests, attendance, lab-work, previous semester grade and their participation in extra-curricular tasks are the basis for internal assessment. In addition, an external assessment of a student is on the bases of marks scored on the final exam. The proposed model used in this Paper helps to predict the students about poor, normal and great in light of class performance additionally class attendance from the generated rules. Result: From the extracted pattern Apriori algorithm is found to be effective in predicting the student under three categories: good, average and poor.

Quadri, et al. [5] presents decision tree method to measure the performance of academic based on their cumulative grade point average (CGPA). In this paper the factors that can influence students' dropouts selected by the decision tree and the effect of each risk factor is quantified by logistic regression.

Thiele, Tamara, et al. [6] utilizes the relationship between type of school, school performance, socio-economic deprivation, school grades, neighbourhood involvement, gender and academic success. Multivariable logistic regression is presented to identify the factors,

which were independently associated with academic performance. According to this paper, more affluent students per- formed better than students from lowest income.

In Ahmed, et al. [7] presents prediction the final grade of stu- dents, which based on the decision tree (ID3) classification. Some variables were collected from the database of students to predict the student's final grade. This study will benefit the students to improve the student's performance, to identify those students who needed additional attention to reducing failing ration and taking appropriate action at the right time.

Bunkar, Kamal et al. [8] generates the classification rule, which is based on decision tree to classify the student performance and identify the probability of failing students. It uses generated rules to which allows predicting the final result in un-der studying course. The classification algorithm successfully identifies the students who are expected to fail.

In Bhardwaj, et al. [9] authors are used in student database to predict the students' division based on last year database. This study will benefit the students and the teachers to improve the division of the student. The study investigates that not only the effort of the student's influential factor but other factors have significant influence over students' performance. This proposal will improve the insights of previous methods.

Yadav, et al [10] paper used C4.5, ID3, and CART decision tree algorithms predict performance of engineering students' in the final exam. The result of the decision tree method shows the num- ber of students who are expected to fail or pass to next year. This study helps to improve the performance of the students who were predicted to fail or pass. The results of the comparative analysis show that low achiever students can be benefited from the predic- tion to improve their potential.

In Satyanarayana, et al. [11] Decision Trees - J48, Naïve Bayes and Random Forest are used to increase the quality of student data by eliminating the noisy classes, and hence improving pre-diction accuracy. Personal, socio-economic, psychological and other environmental attributes are used to measure student academic per- formance. Apriori, Filtered Associator and Tertius are used to identify factors that can affect student result. The paper empiri- cally compares the selected technique with single model-based techniques and show that using hybrid models, it gives better pre-

Diction accuracies and also provides better rules for identifying the factors that influence student results.

Paper Mythili, et al. [12] presents data mining methodologies to study and analyse the school students' performance based on classification techniques, which are useful to gauge students', performance. This paper considers the classification accuracy, confusion matrices and the execution time of various data mining classification algorithms. Multi-layer perceptron, Random Forest algorithm, Decision tree C4.5 (J48), and Lazy based classifier (IB1), Rule-based classifier is used to classify students' performance. The paper investigated that Random Forest performance performs better than that of other algorithms used in this study.

In Osmanbegović, et al [13] Paper authors conduct different algorithms that result in research results in a reduction of data dimensionality and student's performances prediction by their personal demographic and societal features. Random Forest and J48 generate classification model achieve accuracy higher than 71%.

In Al-Radaideh et al. [14] work decision tree method is used in prediction of the student's final grade regarding C++ course, ID3, C4.5, and the Naïve Bayes classifications were used. The results of this study show that Decision Tree model had a better predic- tion than other models.

Baradwaj, et al. [16] mainly studies the prediction of students' performance by using datamining techniques. The decision tree algorithm is used to evaluate the student performance at the end of the semester on the basis of gathered information from manage- ment system of students such as, class attendance, test, Seminar and Assignment marks. The paper also investigates the classifier accuracy in predicting the students' performance.

Kushwah, et al. [15] authors conduct a study on the comparison of various data mining algorithms in the research community. K- nearest neighbour (kNN) classification algorithm is more powerful, easily understandable, easily implemented and works in a different situation. It finds that a group of k objects in the training set that is related to the test object. And also, it is based on the label assignment on the majority of a particular class in this neighbourhood

Table 1: Summary of Previous Works

Paper	Methodology	Result
Quadri, M. M., & Kalyankar,N.V. [5]	decision tree techniques logistic regression	A decision tree algorithm picks the factors that can influence dro- pouts. The dropouts and the consequence of each risk factor measured by logistic regression.
Thiele, T., et al. [6]	Multivariable logistic regres- sion	School grades are representative of 'true academic' potential by comparing group differences in attainment at school compared to university.
Bunkar, Kamal, et al. [8]	decision tree	It uses generated rules to predict the final grade in a understudy- ing course. Identifying the students who are probable to fail.
Bhardwaj, Brijesh Kumar, and Saurabh [9]	Bayesian classification	The result achieves that the factors like students' grade in the se- nior secondary exam, living location, the medium of teaching, family annual income, qualification of mother's, other habit of students, and family status of the students were highly correlated with the academic performance of students.
Ahmed, Abeer Badr El Din [7]	decision tree(ID3)	It predicts student final grade data from student database. This study will help the students to improve the student's perfor- mance. Identify students who needed additional help not to fail in exam.
Yadav, Surjeet Ku- mar, and Saurabh Pal [10]	C4.5, ID3, CART decision tree	The result shows that, the decision tree can identify the number of students who are expected to fail or pass to next year.
Ashwin Satyana- rayana, Mariusz Nuck- owski [11]	Apriori, Filtered Associator and Tertius Decision Trees-J48 Naïve Bayes, and Random Forest	Various approaches are used for removing noisy data and result prediction. Apriori, Filter Associator, and Tertius are used to see- lect the best predictor variables. This provides better rules for understanding the factors those influ- ence better student outcomes.
Mythili, M. S., and AR Mohamed Shana- vas. [12]	C4.5 J48 Random Forest, Multilayer Perception IB1 Decision Table	Various methods are compared and investigated that Random For- est performance is best than that of other algorithms employed in the study based on accuracy achieved, confusion matrix result and time is taken to execute.
Osmanbegović, E., Suljić, M., & Agić, H. (2015). [13]	Random Forest J48	The methodology is used in the reduction of data dimensionality and prediction of student's performances. Their personal demographic and societal features use data. Achieves results higher than 71%.
Al-Radaideh, et al. [14]	ID3, C4.5, NaïveBayes	It compared decision tree and naive Bayes data mining techniques and summarized that decision tree achieves best.
Shiv Pratap Singh Kushwah, et al. [15]	K-nearest neighbor (KNN)	Compare several techniques and states that K-nearest neighbor (kNN) classification algorithm is easy to understand and to im- plement and also more powerful than others

3. Data mining methods

3.1. Association rule mining

Association rule mining is a method in which frequent patterns, correlations, associations, or causal structures are extracted from datasets.

a) Apriori Algorithm- is the most popular tool used by dif- ferent researchers in frequent pattern and association rule mining. It is used to discover a correlation between variables in a large dataset.
b) FP-Growth Algorithm - is fast algorithm for determin- ing frequent item sets in the record. It is also memory ef- ficient but expensive to build FP-Tree.

3.2. Classification algorithms

There are a number of classifier algorithms used in predicting student academic performance.

a) Decision tree - A decision tree is the most effective and well-known method for classification and prediction in educational data mining. A Decision tree normally begins with a single node, which branches into the conceivable outcome.
b) Naive Bayes - The Naive Bayes Classifier is based on Bayesian theorem and is specifically for high dimensional inputs. In addition to its simplicity, Naive Bayes can often perform better classification methods. Naive Bayes classi- fier is widely used in prediction students' academic per- formance.
c) Support Vector Machines - Support Vector Machines (SVM) are based on the notion of decision planes that states decision boundaries. A decision plane is used to separate the objects from contained under different classes. The standard.

SVM takes a sorted dataset and predicts, for each given in- formation, which of two conceivable classes includes the in- formation, making the SVM a non-probabilistic binary lin- ear classifier.

d) Random Forest - Random forest algorithm is the most popular algorithm, which can use for both classification and the regression kind of problems. Random forest algorithm is one of a supervised classification algorithm, which creates a forest using a number of trees. The forest gives the high ac- curacy results since it is constructed from the higher the number of trees.

4. Data preparation

In this study, the dataset obtained from the Kaggle educational dataset, which is collected from learning management system, will be used. The dataset contains 480 student data rows and 16 va- riables. The dataset consists of three main groupings of features:

a) Demographic: - gender, nationality
b) Academic background: - educational stage, grade Level, and section
c) Behavioral features: - raised hand in class, opening resources, answering the survey by parents, and school satisfaction.

Since real-world data is often inconsistent, incomplete, and lacking in certain behaviors and is probably to contain several errors, Data preprocessing is a known method to solve such problems. Data preprocessing prepares unprocessed data for further processing. In this work, the dataset obtained will be preprocessed, such as data cleaning, transformation, reduction and other tasks will be applied. Many Data mining techniques used by many researchers. However, the proposed method selects the best clas- sifier based on comparative analysis.

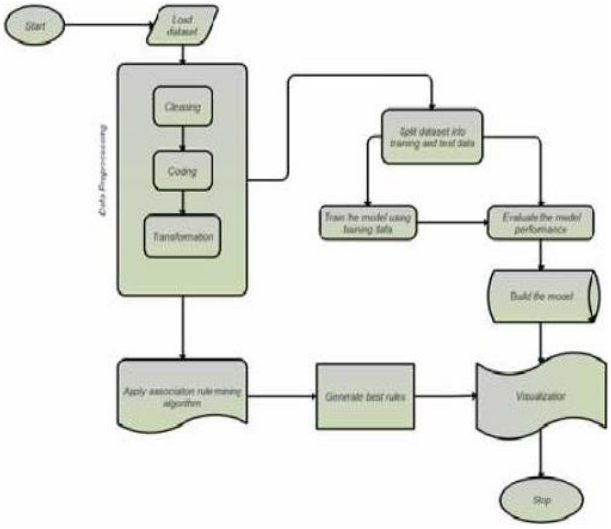

Fig. 1: Flow Diagram of Proposed System.

5. Conclusion and future work

This paper aims to identify related factors and several data mining algorithms used in predicting the performance of students. Predicting the student performance is the most important tool to help the students and schools in improving their learning and teaching process. Various studies have been reviewed that works on pre- dicting students' performance with various methods. A number of factors that can affect the student performance also have been identified. From this review, most of the studies have been achieved a better result by using Apriori algorithm in identifying the relationship between different factors that can affect the stu- dent performance. And also, a decision tree is the most used algo- rithm in predicting the student performance. The future work of this paper will be comparing and using advanced machine learning algorithm to get better prediction result.

References

[1] Tang, Tiffany Ya, and Gordon McCalla. "Smart recommendation for an evolving e-learning ystem: Architecture and experiment." In- ternational Journal on elearning 4.1 (2005): 105.
[2] Arockiam, L., et al. "Deriving Association between Urban and Rural Students Programming Skills." International Journal on Computer Science and Engineering 2.3 (2010).
[3] Baker, Ryan Shaun, Albert T. Corbett, and Kenneth R. Koedinger. "Detecting student misuse of intelligent tutoring systems." International conference on intelligent tutoring systems. Springer, Berlin, Heidelberg, 2004.
[4] [Angeline, D. Magdalene Delighta. "Association rule generation for student performance analysis using apriori algorithm." The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1.1 (2013): 12-16.
[5] Quadri, Mr MN, and N. V. Kalyankar. "Drop out feature of student data for academic performance using decision tree techniques." Global Journal of Computer Science and Technology (2010).
[6] Thiele, Tamara, et al. "Predicting students' academic performance based on school and socio-demographic characteristics." Studies in Higher Education 41.8 (2016): 1424-1446.
[7] Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Mining: A prediction for Student's Performance Using Classifica- tion Method." World Journal of Computer Application and Tech- nology 2.2 (2014): 43-47.
[8] Bunkar, Kamal, et al. "Data mining: Prediction for performance improvement of graduate students using classification." Wireless and Optical Communications Networks (WOCN), 2012 Ninth Inter- national Conference on. IEEE, 2012.
[9] Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." arXiv preprint arXiv: 1201.3418 (2012).

[10] Yadav, Surjeet Kumar, and Saurabh Pal. "Data mining: A prediction for performance improvement of engineering students using classification." arXiv preprint arXiv: 1203.3832 (2012).
[11] Satyanarayana, Ashwin, and Mariusz Nuckowski. "Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance." (2016).
[12] Mythili, M. S., and AR Mohamed Shanavas. "An analysis of students' performance using classification algorithms." IOSR Journal of Computer Engineering 16.1 (2014): 63-9.
[13] Osmanbegović, Edin, Mirza Suljić, and Hariz Agić. "Determining dominant factor for student's performance prediction by using data mining classification algorithms." Tranzicija 16.34 (2015): 147-158.
[14] Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. "Mining student data using decision trees." International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan. 2006.
[15] Kushwah, Shiv Pratap Singh, Keshav Rawat, and Pradeep Gupta. "Analysis and comparison of efficient techniques of clustering algorithms in data mining." International Journal of Innovative Tech- nology and Exploring Engineering (IJITEE) 1.1 (2012): 2278-3075.
[16] Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint ar- Xiv: 1201.3417 (2012).