Journal of Computer Engineering & Information Technology

Mini Review

A SCITECHNOL JOURNAL

Review on Predicting Student Academic Performance using Data Mining Classification Algorithm

Wasyihun Sema Admass*

Abstract

This paper has reviewed previous studies on predicting students' performance with various analytical methods. Most of the researchers have used cumulative grade point average (CGPA) and internal assessment as data sets. While for prediction techniques, the classification method is frequently used in educational data mining area. Under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students' performance. In conclusion, the meta-analysis on predicting students' performance has motivated us to carry out further research to be applied in our environment. It will help the educational system to monitor the students' performance in a systematic way.

Keywords

Student Performance, Prediction Technique, Data-Mining, Algorithms.

Introduction

The topic of explanation and prediction of academic performance is widely researched. The prediction of student performance should be topical debates in the education center. There are increasing research interests in education field using data mining. Application of Data mining techniques concerns to develop the methods that discover knowledge from data and used to uncover hidden or unknown information that is not apparent, but potentially useful [1].

In the area of educational center the data is increased rapidly so the researcher should have to transform in to useful information and knowledge, so data mining techniques play a special role in extracting useful and hidden patterns form tremendous amount of data. In the area of education, educational data mining EDM has become an emerging area for research interest amongst scientists and researchers across the globe. The EDM converts raw data from traditional and online education systems into important and useful information for educational institutes and research [2]. Different scholars perform research on the area of education to predict the academic performance of students. All of them agreed that predicting student's academic performance helps to identify the status of students as slow learner (poor), good learner(Good), Medium learner(Average), very good learner and excellent learner. Classifying the student academic performance by their status helps to improve the student from failure and the teachers to focus on the poor learner students. Students could improve their learning activities, allowing the administration to improve the systems performance. Thus, the application of data mining techniques can be focused on specific needs with different entities. In this systematic article review is reviewed to answer the following question as hypothesis. These are

- what importance attributes the researcher focused to predict student's academic performance
- What are the methods that different researchers used to predict student's academic performance
- What will be predicted as a future work from the given articles?
- Which of the data mining algorithms when used the most predictive set of students' academic performance.

Objectives

The objectives of this systematically article review on the prediction of student academic performance using data mining classification technique is the following.

- To identify the attributes used to predict the academic performance of students
- To identify the gaps in the existing prediction and indicating future work
- To identify the methods used in the existing prediction methods to predict students' performance.

Methodology

The reasoning for performing meta-analysis way of systematic article review is to find suitable methods for existing parameter, to fulfil the gaps in existing research and to place a new research activity in the suitable context.

- Searching: There is large amount of articles done on the area of education with different titles. To perform meta-analysis way of article review searching is important to get multiple articles from different journals by using keywords.

Usage of Data mining to predict students' academic performance

Many researcher do research on the area of education to predict the academic performance of students by using different data mining Techniques.

Surjeet Kumar Yadav and Saurabh Pal [3] conducted a research on 400 students to predict academic performance engineering students by using decision tree (ID3 and C4.5 and CART) algorithms. The researcher uses the past performance of the students to predict whether a new student will perform or not and it predict the result as pass and faille. The study experiments are conducted to find the best classifier for prediction of student's performance in First Year of engineering exam. From the classifiers accuracy it is clear that the true positive rate of the model for the FAIL class is 0.786 for ID3 and

***Corresponding author:** Wasyihun Sema Admass, Faculty of informatics and Department of information technology, University of Gondar, Gondar, Ethiopia, E-mail: wasit1014@gmail.com

Received: November 03 2021 **Accepted:** November 17, 2021 **Published:** November 24, 2021

C4.5 decision trees that means model is successfully identifying the students who are likely to fail. These students can be considered for proper counselling so as to improve their result. The study is also focused to identify those students which needed special attention.

Vrushali Mhetre and Prof. Mayura Nagar [4] paper focuses on predicting academic performance as slow learner, fast learner and average learner. For that they applied various data mining techniques and compare the Accuracy based on students attributes. This research work is done to identify the best feature selection and classification algorithms to examine slow, average and fast in education data set. to find the best attribute by comparing the performance of various feature selection techniques in the prediction of learners using different classification algorithms such as Naïve Bayes, J48, ZeroR and Random Tree using WEKA tool. The idea of this research work is to identify slow learners which help the faculties to give special attention to individual student's to improve their overall performance. Finally it has been investigated that Random Tree technique performs best with accuracy 95.4545% and identify students who are slow learners which further provide base for deciding Special aid to them.

Sagardeep Roy and Anchal Garg [5] conduct a research on predicting student academic performance using data mining techniques which has the goal to help student improve their skills, to find out what hinders student from achieving success and how to improve it. This paper is don on 32 attributes of a student by using Naïve Bayes classifier, J48 Decision Tree and MLP classification algorithms. The accuracy of these algorithm is Naïve Bayes classifier 68.6 %, J48 algorithm 73.92% and MLP has 51.13 % there for the result is J48 perform best accuracy than others. The result identify the abilities of students, their interests and weaknesses. Student performance can be influenced by different types of attributes. This can be social, demographic and related to school.

M. Mayilvaganan and D. Kalpanadevi [6] conduct a research to predict student academic performance using classification algorithm to classify the student as Excellent Leaner, Good Learner, and average learner, Slow Learner for diagnosis by using three main classification techniques such as decision tree, Naïve Bayesian methods, and knearest- neighbor. The research experiment result of these three algorithms the decision tree, shows that 30% slow learner, 20% of Average learner, 40% good learner and 10% of excellent learner and Naïve Bayes algorithm, it is also decision tree classification based on probability inference the result was found from Figure 1, shows the performance analysis by using this algorithm the result shows 30% of slow learner, 20% of Average learner, 40% good learner and 10% of excellent learner. The experiment of knearest- neighbor algorithm result shows 45% of slow learner, 10% of Average learner, 5% good learner and 40% of excellent learner. From these experiment and analysis of classification accuracy, K-Nearest Neighbor taken a less time for classifying the student performance as excellent learner, Good Learner, Average Learner and Slow Learner. Knearest Neighbor has best accuracy of time taken in classification when compared to other techniques by the significance of examination result and other activities are affected in the rule set. This study is very useful to identify the ratio of slow learner for rectify the failures early and take action to improve the weaker student in perfectly manner.

V. Ramesh, P. Parkavi and K. Ramar [7] conducted this research paper which focuses on identifying weak students and the identified student can be individually assisted by the educators so that their performance is better for the future. This study is also investigate the accuracy of some classification techniques for predicting the performance of students. The researcher uses four different classification algorithms: NaïveBayes, Multilayer Perception, J48, and REPTree. From the research experiment the result shows multilayer perception (MLP) classifier is most appropriate for predicting student performance which gives 72.38% of prediction and the paper concludes the important factors that affect the students performance related to the school.

Sajadin Sembiring [8] conduct a research on student performance prediction to predict the performance of students based on their grade (GPA). The researcher grouped all the grades in to five groups 'excellent, very good, good, average and poor' and the researcher categorized the value of each item in questionaries' with high, low and medium. The researcher uses two data mining techniques SSVM and kerner k-means clustering algorithm. This paper is done on 300 students of samples and every samples is expressed by ten characteristics parameters. We used five performance predictors that proposed in this study and five characteristics demographic data of student. From the research experiment result shows that the average testing accuracy for the lowest 61% for prediction "*good*" performance

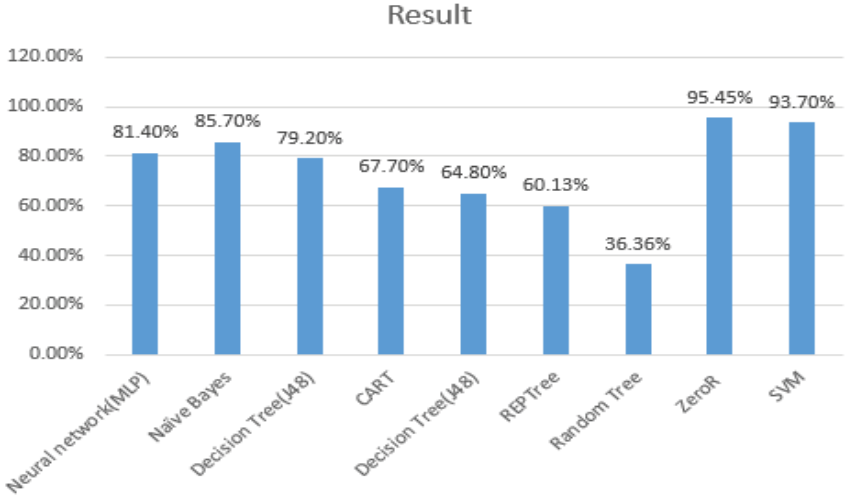

Figure 1: Accuracy result of different prediction methods.

and the highest 93.7% for the prediction "*poor*" performance. Based on the results obtained they are sufficient to prove that the rule model of prediction student performance by using predictor's of student performance proposed acceptable and good enough to serve as predictor of student performance.

Ahmed Mueen, [9] this research paper is conducted two achiever three basic objectives first objective was to predict student academic performance, second objective was to reduce number of attributes, and the last objective is to compare classification accuracy of different classifiers. The researcher uses three classifiers to achieve these objectives Naïve Bayes, Multilayer Perception (neural network) and C4.5 (J48). From these research experiment the accuracy of each classifier shows naïve Bayes 86%, Multilayer Perception 82.7% and decision Tree (J48) 79.2%. From this the researcher conclude that Naïve Bayes classifier has best accuracy to predict the student performance than others. Finally the researcher analyzed the dataset to identify factors which cause student to loss his academic status due to academic performance. We have found that poor performance of student was due to lack of participation in on-line discussion forum.

Important Attributes used to Predict Student Performance

The meta-analysis systematical article review helps to identify the important attributes used to predict academic student performance. The attribute's which are frequently used and which has great role is predicting student academic performance the attributes will be important attribute Table 1.

The attribute's that are frequently used is GPA and assessment. The researcher used the GPA frequently either directly or indirectly to predict the students' academic performance. GPA is a good predictor because it is tangible measurement for future education and career mobility CGPA is the most influence attributes in determining the survival of students in their study, whether they can complete their study or not. In this review, assessment was classified as assignment mark, quizzes, lab work, class test and attendance. All attributes will be grouped in one attribute called internal assessment. The attributes are mostly used among the researchers to predict students' performance. The next important attributes used to predict student academic performance is student demographic factors which includes gender, age, family background, and disability. The reason that the researcher used demographic factors of a student is to identify which sex has better attitude to learn and more strategic to study. The other attributes which are used by the researcher extracurricular activity high school background there are also several researchers in another study who have used psychometric factor to predict students' performance. A psychometric factor is identified as student interest, study behavior, engage time, and family support. They have used this attributes to make a system to look very clear, simple and user friendly. It helps the lecturer to evaluate students achievement based on their personal interest and behavior. However, these attributes are rarely to apply in predicting students' performance by several researchers because it focuses more on qualitative data and it is also hard to get a valid data from respondents.

Prediction Methods used in Predicting Student Performance

To predict student performance prediction modelling is used. To use predictive modelling in educational data mining different activity may be performed like classification, regression and categorization. The most popular task to predict students' performance is classification. From the classification techniques the researcher uses Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine.

The specific application of data mining techniques grouped by algorithms in predicting student performance will be described in the following:

Decision Tree

Decision tree is the most wieldy used classification algorithms in data mining techniques. The decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules. From the seven papers six of the paper have been used decision tree algorithms Table 2.

Naive Bayes

Naïve Bayes algorithm is the next option for the researcher to predict the students' academic performance. Among seven (7) papers five (5) of them have been used naïve Bayes algorithm as prediction method to estimate student performance. The following table shows the predicted result of naïve Bayes algorithm Table 3.

Neural Network

The next predictor method used by the researcher to estimate student performance is the neural network. The researcher uses Multi-layer perceptron algorithm to predict student performance. Among seven papers three of them have been used this neural network techniques. Look at the following table which shows the result estimated by neural network techniques Table 4.

K-Nearest Neighbor

The researcher have been also used K-nearest Neighbor data mining algorithm as predictor method to predict student performance. From the seven paper one paper is used K-nearest neighbor algorithm as predictor method. According to [6] the estimated result of the student by using k-nearest neighbor algorithm is 45% of slow learner, 10% of Average learner, 5% good learner and 40% of excellent learner from the required data set.

REPTree, Random Tree and ZeroR

Some of the researcher also used these algorithms to predict student academic performance. According to [7] REPTree is used to estimate the student performance based on demographic and psychometric attributes. The predicted result of REPTree is 60.13%. ZeroR also used by [4] to estimate the student performance, then the result shows 36.36%. Random Tree algorithm is used by [4] to predict student performance and the result is 95.45 % Table 5.

Support vector Machine and K-means Clustering

The researcher Sajadin Sembiring [8] has been used these two methodes to estimate the student performance. According to [8] after data set is prepared the data set is inserted to k means clustering algorithm. The number of clusters was determined as an external parameter then grouped in to five clusters. The researcher is also used SVM as prediction method and the result of the prediction. Average testing accuracy for the lowest 61% for prediction "*good*" performance and the highest 93.7% for the prediction "*poor*" performance. The following graph indicates the best prediction methodes used in these systematic review.

Table 1: Attribute Factor Used in Different Researcher.

Attribute	Authors
Students Branch,Students grade in High School,Students grade in Senior Secondary,Medium of teaching, Living location of teaching ,Student family size, Student family status, Family income,Family occupation,Result(Pass,Pro,Fail)	Surjeet Kumar Yadavand Saurabh Pal [1]
Student CBGS, curricular and extra-curicular activity, quize, assignment marks, projects, result,Learner{slow,average,fast}	Vrushali Mhetre and Prof. Mayura Nagar [4]
School, Type of Address, Parent's Cohabitation Status, family educational qualification, family employment type, Reason for opting a certain school, Time taken to travel to school, Weekly Study time, Educational support given by family, internet access, family relationship, free time out of school, workday alcohol Consumption, weakly alcohol consumption, current health status, Absences in school, first year grade, second year grade, grades	Sagardeep Roy, Anchal Garg [5]
Speciality, lower class grade, Higher Class Grade, Extra Knowledge or skill, Attendance, hours spend to study, resources, seminar performance, result, class test grade(internal), lab work, exercise, homework, quiz, over all semester mark	Mayilvaganan Kalpanadevi [6]
Grade obtained at secondary level father occupation, mother occupation, school area at secondary level, school area at higher secondary level, private tuition at secondary level group of study, student's community, school area at elementary level, parent's education	
Interest, Study Behaviour, Engage Time, Believe, and Family Support and GPA as dependent variable	
Grade Point Average (GPA), quiz1, quiz2, quiz average, Assignment submit, Assignment delay, labtest1, labtest2, lab test average, final exam grade, total time spent, hours spent studying daily, methods of study used, city of birth, transport method, distance to the college, subjects interest, motivation level, difficulty doing homework, facilities in college, having home tuition, level of father education, level of mother education, attendance	

Table 2: Decision Tree Accuracy Result.

Methods	Result	Author
Decision Tree	ID3=62.2, C4.5=67.7,CART=62.2	Surjeet Kumar Yadav and Saurabh Pal [1]
Decision Tree(J48)	=72.7	Vrushali Mhetre and Prof. Mayura Nagar
Decision Tree(J48)	73.92 %	Sagardeep Roy, Anchal Garg
Decision Tree(J48)	79.2%	Ahmed Mueen, Bassam Zafar, Umar Manzoor
Decision Tree(J48)	64.8%	V.Ramesh, P.Parkavi and K.Ramar [7]
Decision Tree(C4.5)	30% of slow learner, 20% of Average learner, 40% good learner and 10% of excellent learner	M. Mayilvaganan and D. Kalpanadevi [6]

Table 3: Naive Bayes Accuracy Result.

Methods	Result	Authors
Naïve Bayes	49.5 %	V.Ramesh, P.Parkavi and K.Ramar [7]
Naïve Bayes	30% of slow learner 20% of Average learner, 40% good learner and 10% of excellent learner	M. Mayilvaganan and D. Kalpanadevi [6]
Naïve Bayes	68.60%	Sagardeep Roy and Anchal Garg [5]
Naïve Bayes	68.1818%	Vrushali Mhetre and Prof. Mayura Nagar [4]
Naïve Bayes	85.7%	Ahmed Mueen, Bassam Zafar, Umar Manzoor [9]

Table 4: Neural Network Accuracy Result.

Methods	Result	Authors
Neural network(MLP)	72.38 %	V.Ramesh, P.Parkavi and K.Ramar [7]
Neural network(MLP)	51.1392%	Sagardeep Roy and Anchal Garg [5]
Neural network(MLP)	81.4%	Ahmed Mueen, Bassam Zafar, Umar Manzoor [9]

Table 5: 1 Accuracy Result for REPTree, Random Tree and ZeroR Algorithm.

Method	Result	Authors
REPTree	60.13%	V.Ramesh, P.Parkavi and K.Ramar [7]
Random Tree	36.36%	Vrushali Mhetre and Prof. Mayura Nagar [4]
ZeroR	95.45%	Vrushali Mhetre and Prof. Mayura Nagar [4]

Discussion

The discussion section presents the analyzed result all papers in predicting students' academic performance. The meta- analysis systematic review of this article mainly focuses on identifying the important attributes and methodes used in predicting student academic performance. On the other hand it also identify the research gaps and indicate future works for the researcher.

This review analyze the important attributes in two three basic category of attributes. These are tangible value attributes, demographic attributes and psychometric attributes. The tangible attributes are measurable values like GPA and assessment's the demographic attribute is gender, age, family background, disability and other related attributes and the psychometric attributes student interest, study behavior, engage time, and family support [3-7, 9] from these attributes the best predictor is GPA. All other attributes directly or indirectly predict the GPA of the student.

The next important factor is prediction methods. From the above graph ZeroR has the highest prediction accuracy by (95.45%) followed by SVM which is highest prediction for poor class by (93.70%). The reason that ZeroR algorithm has high prediction accuracy all the attributes are tangible to predict students' performance clearly Next, naïve Bayes and Neural network has high prediction with 85.7% and 81.4 % respectively [9], but naïve Bayes are the lowest prediction method with 45.6 % accuracy at [7] the lowest prediction methodes used in this systematic review is random tree with 36.36 % accuracy result.

Evaluation and critique

Strong side

All the papers are estimate the students' academic performance to identify the week or low performed students to announce the teachers to focus on weak or low performed students to prevent from failure and indicate the teacher should more interactive with student, provide proper guidance and motivate the student. These all researchers use tangible variable, demographic variable, psychometric variables to predict the performance of students (weak or Lowe performed students and strong well performed students).

Weak side

The researcher is focused only the Assessment attribute, demographic factor (attribute) and psychometric attributes. These attributes are not the only important factors that predict the student performance. Other attributes factors like psychological factor, alcohol, and consumption, romantic (love) relationship is also important factors used to estimate the performance of students

Conclusion

Estimating student performance is mostly important for the teachers and students to improve their teaching and learning process. This systematic review has reviewed the previous studies on predicting student academic performance with various prediction methodes. Most of the researcher used demographic attributes, psychometric attributes and tangible attributes but more frequently used attributes are student GPA and assessments (tangible) are used, while for prediction methodes classification method is frequently used. Form the classification method decision tree, naïve Bayes and neural network are widely used by researchers. In conclusion, the review on predicting students' performance has motivated us to carry out further research to be applied in our environment. It will help the educational system to monitor the students' performance in a systematic way and motivate to conduct a research on what is the reason that the student skill will be minimized (educational Froude for the student).

References

1. Sajadin Sembiring (2011) Prediction of student academic performance by an application of data mining techniques.
2. Pamela Chaudhury (2016) Enhancing the capabilities of Student Result Prediction System.
3. Surjeet Kumar Yadav (2012) Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Comp Sci and Info Tech J (WCSIT) 2(2): 51-56.
4. Vrushali Mhetre (2017) Classification based data mining algorithms topredict slow, average and fast learners in educational system using Weka. IEEE International Conference on Computing Methodologies and Communication.
5. Sagardeep Roy (2017) Predicting Academic Performance of Student Using Classification Techniques.
6. Mayilvaganan (2014) Comparison of Classification Techniques for predicting the performance of Students Academic Environment. International Conference on Communication and Network Technologies (ICCNT).
7. Ramesh (2013) Predicting Student Performance: A Statistical and Data Mining ApproachInt J Comput Appl 63: 975-8887.
8. Sajadin Sembiring (2011) Prediction of Student Academic Performance by an Application of Data Mining Techniques. International Conference on Management and Artificial Intelligenc 6.
9. Ahmed Mueen (2016) Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. IJ Modern Education and Computer Science.

Author Affiliation Top

Department of information technology, University of Gondar, Gondar, Ethiopia.