INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 13, Issue 12, (2023) E-ISSN: 2222-6990

# A Systematic Literature Review on Predicting Students Academic Performance By Using Data Mining Techniques

Nurul Hafida Suhaimi[1], Habibah Ab Jalil[2], Iskandar Ishak[3]

[1,2]Faculty of Educational Studies, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia, [3]Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Corresponding Author Email: nurulhafida@gmail.com

**To Link this Article:** http://dx.doi.org/10.6007/IJARBSS/v13-i12/20329 DOI:10.6007/IJARBSS/v13-i12/20329

***Published Date:*** 22 December 2023

**Abstract**

There has been a lot of interest in education on the prediction of students' academic performance. The enormous increase in educational data offers the chance to gather data that can be used to assess the effectiveness of teachers, anticipate student dropout rates, predict overall academic achievement, revise the material to better suit the requirements of students, and much more. However, the lack of a mechanism in place to predict students' academic performance is still a concern in Malaysia. The research on existing prediction techniques is still inadequate and very few studies that have been done on the Malaysian context, especially that contribute to students' academic performance. Given the scarcity of research on existing prediction techniques in Malaysia context, a detailed literature review on employing data mining techniques to predict student performance is suggested. The primary goal of this article is to provide a thorough overview of data mining approaches to predict students' academic performance, as well as how various prediction techniques aid in determining the most significant students' attributes which contribute to students' performance. The findings of this paper offer an insight of the implementation of data mining in a specific dataset, and it summarizes the prediction algorithm with highest accuracy and attributes with significant contributions to students' academic performance.

**Keywords:** Attributes, Academic Performance, Educational Data Mining, School Students

## Introduction

Academic performance explanation and forecasting is a heavily explored topic. Predicting how students will perform should be a hot topic of discussion in the classroom. The use of data mining in the field of education is becoming increasingly popular among researchers. Data mining is the process of using computer algorithms and other computational approaches to discover previously unknown relationships or patterns in large datasets for practical use

(Sembiring, 2011). Academic success of students is crucial in universities. This is because a high-performing academic track record is a key indicator of a top-tier educational institution (Sembiring, 2011). The literature provides numerous definitions of student performance. According to Ramesh (2013), assessing students' performance through learning evaluation and extracurricular activities is a viable option. Graduation rates have been used as a proxy for academic success in most research.

For most Malaysian universities, final grades were the primary indicator of academic accomplishment. Students' final grades are affected by coursework, assessments, final exams, and participation in extracurricular activities (Chaudhury et al., 2016). Evaluation is critical for keeping students' performance at a high level and the learning process running smoothly. As educators analyze students' academic achievement, they are better able to plan a holistic strategy for their time in school (Yadav & Pal, 2012). Several approaches for assessing student progress have lately been created. Data mining is one of the most frequent methods for assessing academic progress.

Data mining techniques are critical in the education industry, where they are utilised to find hidden patterns in massive datasets and transform them into useful insights for researchers. Scientists and researchers worldwide are becoming interested in Educational Data Mining (EDM) as a new arena for academic investigation. EDM uses raw data from both offline and online education systems to better assist academic institutions and researchers (Chaudhury et al., 2016). Academic performance projection is a typical goal of educational scholars. Researchers acknowledged that determining a student's predicted academic success aids in categorizing them as either a bad learner (slow), a good learner (good), an average learner (average), a very good learner (outstanding), or a very good learner (very good). Pupils can avoid failing and teachers can concentrate on struggling students by categorizing their academic performance. The efficiency of the system may be enhanced if students took steps to enhance their educational experiences. This means that data mining methods can be applied to a wide variety of entities with a narrow emphasis. Hence, the following research questions will be tested in this systematic literature review:

1) What are the most attributes used to predict students' academic performance?
2) What are the most accurate prediction models to predict students' academic performance?

**Methodology of The Research**

Performing a meta-analysis is one method of systematic article review that aims to uncover appropriate approaches for a given parameter, fill in research gaps, and set a new study in its proper theoretical and methodological context. Numerous articles under a wide variety of headings have been written about the field of education. Meta-analysis article reviews need extensive keyword searches for several publications across multiple journals.

**Search Strategy**

It might be difficult to find the best balance of sensitivity and specificity when doing database searches or designing search algorithms for systematic reviews. Despite the fact that various methodologies exist that offer principles for systematic search procedures, no one has yet documented a complete search strategy in sufficient detail for others to replicate it. The authors created a protocol that outlines every step required to establish a systematic search strategy for use in the systematic review. The search phrases (keywords and free-text synonyms) were copied and pasted from a thesaurus into a text document, followed by search

syntax (containing field codes, parentheses, and Boolean operators) on a new line. Potentially relevant candidate search terms were found and the assurance of phrase completeness increased by comparing thesaurus results with free-text search results. Word macros were created to translate between database syntax and interface syntax automatically. Only publications addressing machine learning, educational data mining, predictive models, students' academic performance, elementary, secondary, and university levels of education were examined for this study.

**Selection Criteria**

Studies in peer-reviewed journals or conference proceedings produced in English, studies that studied the prediction of students' performance at higher educational/high school/primary school levels.

**Inclusion Criteria**

The inclusion criteria were per below

- Research papers on Students Performance Prediction
- Papers from 2010 to 2022 era.
- Papers written in English.
- Not a review paper

**Exclusion Criteria**

The exclusion criteria were per below:

- Studies not using ML or DM techniques
- Duplicates paper

**Duration of Completion**

Only journals from the past ten years were taken into consideration for this comprehensive literature assessment. 2011 through 2021 are the relevant years. The majority of academic magazines have traditionally been geared toward elementary and secondary school students. Research on the subject of academic achievement among college students around the world accounted for only 25% of the papers. There have been discovered numerous magazines that were published between 2011 and 2021. The subject of education is covered by more than fifty journals. 25 publications were initially taken into consideration, but only 15 of them satisfied the inclusion and exclusion requirements for this systematic literature review.

For the majority of the journals, data from OpenHub, Scopus, ScienceDirect, and Taylor and Francis were mined. A third of the papers were peer-reviewed, while the remaining were theses and dissertations that students who attended foreign universities in 2020 and 2021 wrote. From the sources indicated above, fifteen periodicals were selected; four were from OpenHub, two from Scopus, one from ScienceDirect, and one from Taylor and Francis.

**Results and Interpretation**

**Usage of data mining to predict student's academic performance.**

In an effort to forecast students' future success in the classroom, some researchers have looked into the topic of education using data mining techniques. 400 engineering students' academic progress was predicted using decision tree (ID3, C4.5, and CART) algorithms by Yadav and Pal (2012). By comparing a student's past performance to that of other students who were similar to them, a researcher can predict a new student's success or failure.

Experiments are carried out to predict how well students would perform on the First-Year Engineering Exam. With a true positive rate of 0.786 for the ID3 and C4.5 decision trees, the model is correctly identifying students who are likely to fail using the FAIL class. Giving these pupils aid may improve their performance. The study also emphasises locating students who are in danger.

Researchers, Mhetre and Nagar (2017) have worked hard to identify the characteristics that predict whether a student would study at a fast, medium, or slow speed. Numerous data-mining techniques were used, and demographic comparisons of the students were done. Researchers look for the most efficient feature selection and classification methods to analyse the slow, average, and fast students in a specific education data collection. WEKA was used to compare the accuracy of the predictions made by Naive Bayes, J48, ZeroR, and Random Tree, four distinct classification algorithms. This study aims to identify slow learners so that teachers can focus on them and assist them in enhancing their academic performance. Last but not least, research shows that the Random Tree technique has the highest accuracy (95.4545%) in identifying slow students, offering a rock-solid basis for determining whether or not to offer these kids specialised aid. In order to help students improve themselves, Roy and Garg's (2017) study into projecting student academic performance using data mining techniques identifies the challenges that stand in the way of their academic success and offers remedies. In this study, we classify a dataset of 32 student characteristics using the Naive Bayes classifier, J48 Decision Tree, and MLP approaches. The accuracy of the Naive Bayes classifier is 68.6%, that of the J48 classifier is 73.92%, and that of the MLP classifier is 51.13%. The results show each student's areas of strength and need for development. The general success of a student can be impacted by numerous variables. These elements could be societal, demographic, or academic.

Researchers Mayilvaganan and Kalpanadevi (2014) use three main classification techniques—a decision tree, Naive Bayesian methods, and k-nearest-neighbor—to classify students as "Excellent Learners," "Good Learners," "average Learners," or "Slow Learners"—in order to predict their academic performance. For this experiment, the participants were separated into five groups based on the data gathered using the knearest-neighbor algorithm: the slow learners (45%), the average learners (10%), the good learners (5%) and the outstanding learners (40%) According to the findings of the trials and the accuracy analysis, K-Nearest Neighbor took less time to classify student performance into the categories of excellent, good, average, and slow. The test Knearest Neighbor has the best accuracy of time taken in categorising when analysing the significance of the test result and how other tasks are affected by the rule set. In order to address issues early and improve the performance of the lower-income pupils, it is crucial for this research to identify the percentage of sluggish kids.

This study, conducted by Ramesh et al. (2013), aims to identify students who are having academic difficulties so that their teachers can provide them with individualised support and help them become better students in the future. This study also looks at the effectiveness of several classification approaches in predicting student achievement. The classification techniques used in the study are NaiveBayes, Multilayer Perception, J48, and REPTree. The experiment results demonstrate that the multilayer perception (MLP) classifier, with an accuracy of 72.38 percent, is the most accurate at predicting student achievement. The most significant elements that affect students' performance in regard to the learning environment are also identified in the article.

According to Ahmed Mueen (2016), the three main objectives of this study have all been accomplished. The main objective was to predict student achievement in the classroom, then

to streamline the number of parameters and compare the effectiveness of different classifiers. The study uses Naive Bayes, a neural network called Multilayer Perception, and C4.5 classifiers to achieve these objectives (J48). The studies in this paper show that naïve Bayes is 86% accurate, Multilayer Perception is 82.7% accurate, and Decision Tree (J48) is 79.2% accurate. The researchers came to the conclusion that the Naive Bayes classifier offers the most accurate forecasts of student performance as a result of the findings. Finally, the dataset was assessed to identify the factors that influence a student's academic probation. Academic failure has been connected to students' low engagement in the online discussion forum.

According to their grades, Sembiring (2011) researches ways to predict students' academic success (GPA). The relevance of each component was graded by the study's author on a scale from high (5) to low (1), and the results were then gathered into five distinct categories (excellent, very good, good, average, and poor). The support vector machine and the kerner k-means clustering method are two data mining approaches that are used in the study. For the purpose of gathering information for this study, 150 students were chosen at random, and ten categories of their distinctive characteristics were noted. We examined data on student demographics as well as the five essential predictors of academic success. The trials' findings indicate that "excellent" prediction performance falls between the ranges of 61% to 93.7% for the average testing accuracy. The data gathered is adequate to demonstrate the validity and efficacy of the suggested rule model for predicting student achievement utilising predictors of student success.

"Educational data mining" was created, as El-Halees (2009) describes, to glean knowledge more effectively from data in the educational setting. In his study, student achievement was examined using educational data mining. Additionally, he gathered data from the database course's pupils. Prior to employing data mining algorithms to do tasks like classifying the data, assembling groups of related records, and spotting anomalies, he first prepared the data. The four techniques used represented students' capacities based on the learned material.

Tair and El-Halees (2012) employed data mining approaches to find important information in educational systems. The College of Science and Technology employs educational data mining to improve graduate students' performance over a 15-year period, deal with their poor grades, and gain insightful knowledge from their records (from 1993 to 2007). After obtaining the data, they employed data mining techniques to look for patterns and irregularities. In each instance, they gave examples of the information they had retrieved and talked about its importance in class.

Garć ıa and Mora (2011) developed a model to forecast students' academic success based on sociodemographic and academic variables. This model uses the Naive Bayes classifier and the Rapid miner tool to achieve a classification accuracy of 60%. Bharadwaj and Pal (2012) chose 300 students from five universities for their study on student performance. According to the Bayesian classification approach, which examined 17 factors, students' living arrangements and the type of instruction they received had a big impact on their academic performance.

The Al-Radaideh et al. (2006) methodology, which suggests employing data mining classification algorithms to assign a value to pertinent data, can be used to predict students' academic results. The three techniques that were used were ID3, Nave Bayes, and C4.5. Their findings demonstrated that the decision tree outperformed the other two models in terms of making correct predictions.

Lakshmi et al. (2014) presented details of the students' behaviour. Higher scores were given access to a mastery domain that was more difficult, according to the ID3 approach, which was

used to categorise the students' performance. The best classification method for creating decision trees is the ID3 algorithm, which combines top-down, bottom-up, and greedy search tactics. The preferred metric for determining the most efficient categorization parameters is information gain.

Data mining was used in Ali's (2013) study of educational systems. He obtained the data as soon as the children were admitted. Using data mining techniques, information on students is categorised and clustered while accounting for their demographic, behavioural, and psychological traits. It was helpful in describing student performance as a function of their secondary school grade point average or percentage, whether it was good or negative.

The two most common indicators are grade point average and evaluation. The researcher mostly used the students' GPA to forecast their academic achievement. The cumulative grade point average is a trustworthy measure of accomplishment in school and in the workplace. The most crucial element in determining whether a student will be able to graduate from college is their cumulative grade point average (CGPA). We have categorised the several types of evaluation used in this analysis, including assignments, quizzes, labs, examinations, and classroom participation. The collection of all traits will be referred to as "internal assessment." The traits are typically used by academics to forecast the future success of their students. In order to predict academic success, student demographic factors like gender, age, family history, and disability are used. The study used participant demographic data to identify which gender demonstrates a more favourable disposition toward learning and more efficient study practises. In addition to academic achievement and extracurricular activities, the researcher additionally takes more evidence in favour of the use of psychometric characteristics comes from numerous other studies that used them to forecast students' progress.

**Prediction of students' academic performance**

Models for predicting how students would perform are utilised. Predictive modelling can be used in pedagogical data mining in a variety of ways, including classification, regression, and categorization. Classification is the most commonly used task for this purpose. The researcher used a variety of different categorization methods, including decision trees, neural networks, and naive bayes. Predicting student success through the use of data mining approaches organised into algorithms will be detailed below.

**Summarized Attributes of Different Data Mining Techniques Used to Predict Students' Academic Performance**

The summary of these findings are demonstrated in Table 1.

Table 1
*Summarized Attributes of Different Data Mining Techniques Used to Predict Students' Academic Performance*

| **Attributes** | **Author/s** | **Data mining techniques** | | | | |
|---|---|---|---|---|---|---|
| | | **Decision Tree** | **Nave Bayes** | **Neural Network** | **Kneares t neighbo ur** | **Kerner k-** |
| Efficicent Feature Selection | Mhetre & Nagar (2017) | / | x | x | x | x |
| Past Performance | Yadav & Pal (2012) | / | x | x | x | x |
| Societal, demographic, and academic | Sagardeep & Garg (2017) | x | / | x | x | x |
| Academic Difficulties | Mayilvaganan & Kalpanadevi (2014) | / | / | x | / | x |
| Academic Difficulties | Ramesh et al. (2013) | x | / | x | x | x |
| Low engagement in the online discussion forum | Ahmad Mueen (2016) | x | / | x | x | x |
| Distinctive characteristics | Sembiring (2011) | x | x | x | x | / |
| Irregular patterns of courses outline | Tair & El-Hales (2012) | / | x | x | x | x |
| Sociodemographic | Garcı́a & Mora (2011) | / | x | x | x | x |
| Sociodemographic | Al-Radaideh et al. (2006) | x | / | x | x | x |
| Students' behaviour | Lakshmi et al. (2014) | / | x | x | x | x |
| Demographic, behavioral and psychological traits | Ali (2013) | / | x | x | x | x |

**Discussion**

The discussion section contains the results of all analyses of the studies' ability to forecast students' academic progress. The objective of this article's meta-analysis systematic review was to identify the variables and methodologies that are most effective at forecasting

students' performance in the classroom. On the other hand, it identifies areas where additional investigation is necessary and offers recommendations for future research.

Fifty relevant publications that fit the datasets have been identified based on the key terms utilised to further this systematic literature research. 25 of the 50 journals that were considered have been found to be pertinent. Out of 25 journals, 10 journals were chosen based on inclusion and exclusion criteria. The journals have been found using a variety of resource hubs and according to the inclusion criteria. The exclusion criterion resulted in the exclusion of 15 items.

The most important traits will be categorized into two big classes in this review. Along with the more obvious monetary value, these factors also include demographics and psychometrics. The more tangible qualities include metrics like grades and test results. In terms of both student psychometrics (student interest, study behaviour, engage time, and family support) and student demographics (gender, age, family background, handicap, etc.), GPA is the greatest predictive predictor (Baker, 2010; Mayilvaganan & Kalpanadevi, 2014; Mhetre & Nagar, 2017; Ramesh et al., 2013; Roy & Garg, 2017; Yadav & Pal, 2012). All other characteristics can accurately predict a student's GPA.

The accuracy of future forecasts is critical after historical data. Based on the data shown above, ZeroR has the best prediction accuracy (95.45%), whereas SVM has the worst prediction accuracy (93.70%). Due to the presence of all features required to precisely forecast student performance, ZeroR has a high prediction accuracy. The accuracy of a Neural Network is quite high, at 85.7% (Mueen, 2016) and 81.4% (Baker, 2010), while the accuracy of naive Bayes is the lowest of the prediction methods, coming in at 45.6% at (Ramesh et al., 2013). In this thorough investigation, it was discovered that random trees had the lowest prediction accuracy (36.36%).

**Conclusion and Recommendations**

Using predicted student performance to improve the classroom environment can be very beneficial for both teachers and students. The literature on forecasting student academic achievement was analysed in this systematic review utilising a number of techniques. Although most research focused on demographic indicators, psychometric characteristics, and tangible characteristics like students' GPAs and tests, classification methods are frequently used for making predictions. Neural networks, decision trees, and naive Bayes are all examples of "classification methods," which are widely used in science. We have conducted additional study for local implementation as a result of the overall analysis of anticipating student achievement. Schools will be encouraged to investigate the causes of any drops in proficiency as a result of being better equipped to monitor their students' progress (educational Froude for the student)

**References**

Ahmed Mueen (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *IJ Modern Education and Computer Science*.

Ali, M. M. (2013). Role of Data Mining in Education Sector. *International Journal of Computer Science and Mobile Computing, 2*(4), 374-383.

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining Student Data Using Decision Trees. In *International Arab Conference on Information Technology, Yarmouk University, Jordan*.

Baker, R. S. J. D. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112-118.

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement Using Classification. *Arxiv Preprint Arxiv*, 1201, 3418.

Chaudhury, P., Mishra, S., Tripathy, H. K., & Kishore, B. (2016). Enhancing the capabilities of Student Result Prediction System. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16)* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/2905055.2905150

El-Halees, A. (2009). *Mining Student's Data to Analyze E- Learning Behavior: A Case Study*.

Garć ıa, E. P. I., & Mora, P. M. (2011). Model Prediction of Academic Performance for First Year Students. In *Artificial Intelligence (Micai), 2011 10th Mexican International Conference* (pp. 169-174). IEEE.

Lakshmi, D., Arundathi, S., & Jagadeesh, D. (2014). *Data Mining: A Prediction for Student's Performance Using Decision Tree Id3 Method*.

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of Classification Techniques for predicting the performance of Students Academic Environment. *International Conference on Communication and Network Technologies (ICCNT).*

Mhetre, V., & Nagar, M. (2017). Classification based data mining algorithms to predict slow, average, and fast learners in educational system using Weka. *IEEE International Conference on Computing Methodologies and Communication*. https://doi.org/10.1109/ICCMC.2017.8282735

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *Int J Comput Appl, 63*, 975-8887.

Roy, S., & Garg, A. (2017). Predicting Academic Performance of Student Using Classification Techniques. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).* IEEE. https://doi.org/10.1109/UPCON.2017.8251112

Sembiring, S. (2011). Prediction of Student Academic Performance by an Application of Data Mining Techniques. *International Conference on Management and Artificial Intelligence, 6*.

Tair, M. M. A., & El-Halees, A. M. (2012). Mining Educa- tional Data to Improve Students' Performance: A Case Study. In *International Journal of Information, 2*(2), 140-146.

Yadav, S. K., & Pal, S. (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. https://doi.org/10.48550/arXiv.1203.3832