Enhancement of Students' Academic Performance and Association Rule Mining

Muhammad Saqib*, Muhammad Adil†, Hamail Raza Zaidi‡

Abstract

Data mining techniques have revolutionized the process of finding insights from data and, thus making it easier for managers to make better decisions. Relevant tools and appropriate methodologies have enabled several organizations to extract actionable knowledge available within a massive amount of data. This research study is an attempt to investigate the association among several factors and the academic performance of college/university students. We have discussed and analysed the potential benefits of using a few techniques of data mining in the education sector, and thus established an understanding of how those techniques can effectively help learners in overall academic performance improvement. The research started with a thorough literature review to support arguments. One of the techniques, i.e., Association Rule Mining (ARM), is explored further and compared with a few other techniques in this specific research. It has been found an effective approach in finding associations and discovering patterns in the data. The study also reveals that appropriate use of association rule mining and decision trees with influencing study elements may offer a high level of prediction accuracy of students' achievement, while concluding, it is recommended that other Data Mining algorithms may be taken into consideration for further such discoveries.

***Keywords*:** Machine Learning; Data Mining; E-Learning.

Introduction

Enhancing the effectiveness of the teaching and learning process is the key to greater students' academic success, and so finding ways to realize such is always at the front and center of educational providers' vision and missions (Abulhul, 2021). Given the abundance of students' performance data, approaches that can help mine and extract relationships and identify the influencing factors within that data that may lead to enhanced teaching and learning effectiveness are strongly being considered (Pratama et al., 2023). The Data Mining approach, also known as association mining and often referred to as association rule learning or market basket analysis, is widely used to identify patterns or connections between variables in large datasets, which can then be utilized to predict

*Corresponding Author: Department of Computing, Arden University, Berlin Campus, Berlin 10963, Germany, msaqib@arden.ac.uk

† Computing Department, IQRA National University, Peshawar, 25100, Pakistan, madil@inu.edu.pk

‡ Computing Department, IQRA National University, Peshawar, 25100, Pakistan, hamail.zaidi@inu.edu.pk

outcomes or gather knowledge about how the variables under study behave. This approach is recognized as an appropriate data mining method for application to this study topic. Besides the key influencing factors of students' performance, which include in-class activities, extracurricular involvement also increases the likelihood of students achieving higher grades (Suleiman et al., 2024). Regular class attendance is another influencing factor, but this tends to be more common among students who reside on campus. Students' employment during their studies, for instance, working for more than 20 hours a week, tends to increase the likelihood of lower academic performance, and the same for students spending more than two hours per day on social media (Summer et al., 2025).

Improving student academic performance can be achieved through a variety of approaches. One is to set clear expectations and goals, so students are aware of what is expected of them and what they need to achieve to succeed. Teachers should communicate course objectives and learning outcomes and provide effective feedback that helps students identify areas where they need to improve and provides guidance on how to do so. Teachers should give timely and specific feedback to help students learn and grow. Using differentiated instruction is important as every student is unique and has their learning style (Qorib, 2024).

Instructors should use differentiated instruction to accommodate different learning styles and abilities (Blaz, 2023). This can include using technology, hands-on activities, group work, and promoting active learning to encourage students to actively engage in the learning process (Malik & Zhu, 2023). Active learning can be achieved by using interactive teaching methods, such as discussions, debates, and projects. Fostering a positive learning environment where students can learn better as they feel safe and supported is also crucial. Teachers should create a positive learning environment by building relationships with students, encouraging collaboration, and providing opportunities for student-led learning. Using data such as test scores and assessments to inform teaching instructions can also be very helpful, as this allows them to identify areas where students may be struggling, thus adjusting their teaching accordingly. In addition, providing opportunities for extra help for students who are struggling is essential, and this can happen in the form of after-school tutoring or one-on-one meetings with the teacher. Thus, improving student academic performance requires a combination of effective teaching strategies, a positive learning environment, and personalized support (Pimdee et al., 2024).

The academic success of students is crucial to the functioning of universities as a high-performing academic track record is a key indicator of a top-tier educational institution. The key indicator can only be

meaningful once student performance is clearly defined, as currently, the definition varies from one author to another. The plethora of different ways that student performance may be defined makes it a bit challenging, but looking at a few of them can help to consolidate a reasonable definition. According to Usamah et al. (2013), evaluating students' progress via assessment and extracurricular activities is an effective way to improve education, while most research views the completion of high school to be an indicator of academic achievement (Bin Mat et al., 2013).

On the other hand, most of Malaysia's universities and colleges utilize semester and overall grades as a measure of student success. The final grade, which is used as the key measure of student success, combines the students' performance from the coursework, assessments, final exam, and participation in extracurricular activities. So far, it is widely accepted that both student performance and the efficiency of the educational process depend on regular assessment. A way forward is to have a well-thought-out strategy for how students should spend their time at school, which may be achieved by an appropriate analysis of their performance. There are several methods that are being discussed for gauging student progress now. One of the most common recent approaches to evaluating academic progress is using Data Mining. In recent years, the appreciation of the relevance of Data Mining used to extract insights from large datasets for better decision-making, regardless of industry type, has been on the rise.

The use of Data Mining has been and is still proving to be very beneficial. The increasing use of technology within the educational setting is without doubt creating large datasets about student progress and performance, and so the need for the use of Data Mining there is clearly justified. Integration of Data Mining use in classrooms has great momentum in the developed part of the world, but not so much in the developing part of the world. The usual barriers are the high infrastructure cost, the lack of education-related investments, and the slow pace of developing technical talents. However, the encouragement and the strive to leverage such educational benefits should continue. Data Mining in the classroom has great potential to assist educators in creating accommodating learning environments for learners regardless of their preferred learning styles and cognitive abilities. Data Mining provides the capabilities for extracting meaningful information and for discovering recurring or hidden patterns from the massive educational datasets being generated daily. As a result, the insights gained may be used for quick interventions supporting adaptive learning of individuals, for predicting student performance with recommendations, and for enhancing student learning experience. These are important steps for improving student performance and it may be quite challenging to embrace them without

Data Mining capabilities. In such a scenario, teachers can easily keep track of their students' progress and performance and can continuously adjust their teaching practices to better suit their students' needs. Students would also be motivated to take responsibility for their own learning and so enhance their own educational experiences (Bienkowski et al., 2012). Moreover, the teaching and learning process would improve and so would the student performance. Data mining methods may be tailored to meet the requirements of a variety of organizations and the above discussion has presented a glimpse of that. A systematic review is offered to help reach the following goals of this research:

- To investigate how different student-related factors are associated with academic performance.
- To better understand the present prediction approaches for predicting student performance.
- To examine how the data mining technique is associated with students' performance.
- To highlight some directions for the future related to data mining techniques in education.

In the following paragraphs, the research design of surveys and how they might be utilized to forecast student achievement are discussed. The discussion of research questions and the findings of current prediction algorithms is addressed in depth.

Literature Review

In this part, we showcase a range of previous research and practices in the field of pedagogical Data Mining. Machine learning and Data Mining are the mainstays of the current literature.

In 2002, Han and Kamber described software that facilitates multidimensional data analysis, classification, and summarization, which may be used for Data Mining (Han & Kamber, 2002). They also laid out how to use classification theory to foretell how different topics can be covered in a given syllabus. Such insights may be utilized to enhance the curriculum of any class offered at an institution. In 2011, Pandey et.al. presented the results of their analysis based on the work of sixty students, where they used Bayes classification for sorting by category, language, and other contextual factors like education level. Their findings indicated that they could predict whether incoming students would succeed. In 2012, Azhar Rauf et.al recommended that if the k-means clustering approach is used by calculating starting centroids rather than selecting them at random, it will reduce the number of iterations required and speed up the process of determining student clusters (Rauf et al., 2012).

In 2017, Khan surveyed 400 students at Aligarh Muslim University's senior secondary school in Aligarh, India. The aim was to determine which measures of cognition, personality, and demographic characteristics were most predictive of performance in the scientific stream at the upper secondary level. The method relied on cluster sampling, in which the whole population of interest was partitioned into similar subsets from which a representative sample was drawn. Their results showed that males from lower socioeconomic backgrounds performed better in the general stream of education, whereas girls from affluent backgrounds performed better in the scientific stream (Khan, 2017). In 2007, Galit et.al. provided an example of how student data might be used to forecast test outcomes and alert at-risk pupils in advance. Data mining in the classroom is an area that has seen little amount of research so far. This area of study that combines Data Mining with student learning is sometimes referred to as learning analytics and it is still being considered as an emerging field of study. Therefore, plenty more research must be conducted to facilitate educational institutions in finding ways to improve their educational systems using Data Mining. There is an abundance of student performance datasets within educational institutions but limited abilities to leverage them. Here, some previously published materials are evaluated.

The studies that have been conducted in this area of education are summarized by Romero & Ventura (2010), detailing the users, as well as various educational settings, and the data. They have also detailed the prevalent challenges in the classroom setting and how Data Mining approaches have helped to address them. Multidimensional cube analysis using OLAP technology was used by Hua-long Zhao (2008) to demonstrate how the curriculum selected by students may be affected by factors such as instructor, academic term, and individual learner. His use of the Star model from the data warehouse to the study of curricula may help decision-makers at all educational levels to create better plans for students' academic development (Zhao, 2008).

Student enrollment data has been mined using Data Mining methods by Siraj et al., (2009). They analyzed the strengths and weaknesses of three different predictive Data Mining methods, namely Neural Network, Logistic regression, and Decision tree, to draw conclusions on which one is the most effective. Planners may utilize the data to create a sound teaching and learning strategy for the institution based on that learning.

K-means clustering, a Data Mining approach, was discussed by Ayesha et al. (2010) as an appropriate approach for analyzing student performance in classrooms. In this case, the K-means clustering technique

proved to be an effective approach to uncovering information from a school setting. A real-world experiment was carried out in an ICT educational institute in Sri Lanka by Tissera et al. (2006). The results showed that to discover connections between topics covered in undergraduate curricula, a battery of Data Mining tasks should be conducted. This information is crucial for decision-makers who have a direct impact on the quality of educational programs since it sheds light on the syllabi of such programs.

Using Data Mining, Sun (2010) investigated the academic success of students. The purpose was to propose and implement a rule-discovery strategy that would be well-suited to the assessment of students' learning outcomes to enhance students' capacity for evaluating their own learning and, ultimately, to better serve their own learning and development. Data Mining techniques was used by Siraj et al. (2009) to extract information on student enrolment. To determine which of the three predictive Data Mining techniques (Neural Network, Logistic regression, and Decision tree) was most appropriate for such a scenario, they compared and contrasted their respective strengths and limitations. Based on the findings, the institution's planners may implement an effective approach. Ayesha et al. (2010) addressed the use of K-means clustering, a Data mining technique, to evaluate classroom performance. Here, K-means clustering was used to glean insights into the institutional context of a school (Musa et al., 2024). Tissera et al. (2006) discussed the findings of a field study conducted at an ICT training center in Sri Lanka.

Research Design and Methodology

Based on the above literature sources, there are gaps in the area of data mining models and students' academic performance; thus, our study is an attempt to further investigate and address the subject. We have examined so far that limited contextual applications of data mining, e.g., associated rule mining concerning education, have been studied. We have observed a lack of comparative analysis between different techniques in the education environment. Thus, this study is addressing all such issues, bridging the gap between data mining theory, techniques, and their real-world optimal use in the education sector.

In addition, our research study employs a specific data mining technique, i.e. association rule mining, to analyze data and predict academic performance. A few other models for comparison i.e. Naive Bayes, Support Vector Machine (SVM), Decision Trees, and k-Nearest Neighbors (kNN) are used.

The goals of a systematic relational review are to identify relevant approaches for a given parameter, to fill research gaps, and to situate a new

research endeavor in the appropriate framework. A systematic literature search is conducted to provide evidence for the hypotheses or questions that will guide the investigation. The next section will be finding the right research questions to direct the findings. To define the study's parameters, this is also helpful.

Research Questions

Understanding the present study of forecasting students' performance requires the right research questions. Table 1 displays the requirements for research questions, which are based on Kitchenham's framework for constructing them (Population, Intervention, Outcome, and Context, or PIOC).

Table 1: Details of Population, Intervention, Outcome, and Context (PIOC).

Criteria	`Details
Population	University (student performance)
Interventions	Accuracy in forecasting, methods for making reliable forecasts
Outcome	Predictive reliability, effective methods
Context	`Colleges and universities Research methods from all areas of empirical study, including but not limited to: pilot studies; questionnaires; experiments; and case studies

Given this context, the suggested research questions for this study are as follows:

- What are some factors related to students that significantly influence academic performance?
- What are some common ML Algorithms to predict students performance?
- How do techniques of data mining differ in their association with education?
- What are some gaps that can be addressed related to education and data mining techniques?

However, a preliminary study is recommended before diving into the details of this research. The study's primary aim is to determine whether the study's research questions are adequately aligned with those goals. The report then goes on to describe the research methods used to complete the pilot study.

For a systematic review to provide complete findings, the search strategy must be meticulously thought out to ensure that all relevant works are located within the search results. To that end, a thorough literature search was done to see if any answers could be found for the issues that had been stated. The search keywords for this systematic review were

created following the procedures outlined by Kitchenham et al. (2010). Here are some examples of search terms that came out of it: (student outcomes) AND (Data Mining in education) AND (systems, applications, methods, processes, techniques, methodologies, and procedures) AND (prediction OR estimation OR assessment). These choices make up the search strategy: IEEE Xplore, Spinger Link, Science Direct, and the ACM digital Library were among the resources combed for such search terms. Journal articles, workshop papers, and conference papers are also good reference points. Additionally, full-text search was conducted so as not to overlook relevant publications that may not have included the specified keywords in the title or abstract. Considering its publication history, the year 2002 marks a significant milestone. Since this material was first published in early 2015, our search was limited to that period and so after January 2015 papers were excluded. (Shahiri et al., 2015).

Analysis and Discussion

In this part, an analysis of the outcomes from the current studies aimed at forecasting student performance is conducted. This meta-analysis takes into account the most reliable predictors as well as the most crucial elements that may have an impact on students' final grades. A few screenshots as evidence of using a machine learning model and finding performance metrics are given below.

```
[12]  #Let us Use Train Test Split Method

      from sklearn.model_selection import train_test_split
```

```
[23]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
#To Call predict method for predictions on the test dataset

model.predict(X_test)
```

```
array([ 7778.2164802 ,  7039.64288515,  4432.83568254, 10018.04077753])
```

```
# Let us find the Accuracy of the Model using Score function

model.score (X_test, y_test)
```

```
0.8486015395887757
```

Figure 1 and Table 2 show some of the proposed machine learning models and performance metrics, i.e., accuracy, precision, recall, and F1

score. Neural Network has a 98% prediction accuracy, followed by the Decision Tree with 91% prediction accuracy. SVM and K-Nearest Neighbor methods follow with 83% prediction accuracy, and then Naive Bayes with 76% prediction accuracy, which is the least effective strategy for making predictions due to its simplistic nature. If we look at the performance metrics, the values will vary depending on the overall dataset quality, feature selection, and distribution, etc. While discussing the confusion matrix, the models correctly identified overall high-performing students in the results, while a few false positives were identified. The confusion matrix for each model can be discussed in a future study. The model correctly identified 86% of high-performing students, with a small number of false positives.

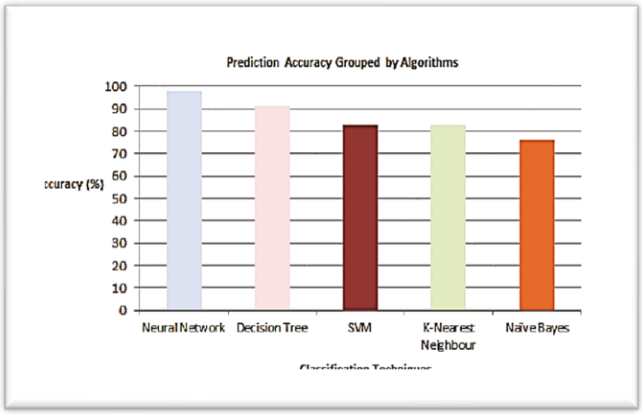

Figure 1: Prediction accuracy using a classification technique organized by algorithms.

Table 2: Details of Population, Intervention, Outcome, and Context (PIOC).

Decision Tree Classifier		Naïve Bayes	
Metric	Value	Metric	Value
Accuracy	90.6%	Accuracy	75%
Precision	87%	Precision	70%
Recall	86%	Recall	71%
F1 score	85.4%	F1 score	70.4%

The properties or features employed in the prediction process have a direct impact on the accuracy of the resulting predictions. Due to the weight given to primary characteristics, the Neural Network approach yielded the best predictive accuracy. Internal and external evaluations have been combined to form these characteristics. Accuracy drops by 1% when using just one variable type, that is, external evaluations. Internal

evaluations are the third most common kind of variable, and they predict success with an accuracy of 81% on average. This demonstrates the significance of external evaluations, which, in this case, use the grades students get in their final exams, in forecasting their future success. When it comes to predicting students' achievement, however, psychometric characteristics were the least reliable predictor, with only 69% prediction accuracy. When compared to the quantitative data often used by Neural Network algorithms, the qualitative data used by psychometric variables makes prediction more challenging. Yet, the maximum error prediction is still lower with Neural Networks. Predictions will be within a tenth of a percentage point of reality, at most. A further benefit of Neural Networks is their simplicity in capturing nonlinear interactions. Its capacity to quickly update past data, much like the human brain, has earned it the name "adaptive system." As a result, the model may continuously expand beyond the available data. A neural network's greatest strength is its capacity to master new information with just a small amount of input (Shahiri et al., 2015).

The Decision Tree technique comes in second with a prediction accuracy of 91%. Based on the Decision Tree technique, a student's cumulative grade point average (CGPA) is the most reliable indicator of their success. Two further studies corroborate this claim, showing that 90% accuracy in performance prediction is achieved when CGPA is used as the primary characteristic. To summarize, the Decision Tree works well with huge datasets, can process both numerical and categorical data, and the relationships between variables are straightforward to understand and comprehend. Additionally, psychometric criteria are the least accurate in predicting student achievement, with a result of just 65%. This shows that employing psychometric characteristics to predict student performance is not a good fit for Decision Tree.

Support Vector Machine, which has an approximate accuracy of 83%, can be a third option alongside the K-Nearest Neighbor. According to the findings, psychometric characteristics are the best features to use when using the SVM approach to predict student achievement. Accuracy in performance dropped to 73% when co-curriculars were included. K-Nearest Neighbor, on the other hand, performed very well with 83% prediction accuracy when three criteria were combined to predict students' performance, that is, internal evaluation, CGPA, and extracurricular activities. When compared to Decision Tree and Naive Bayes, K-Nearest Neighbor consistently produces higher accuracy outcomes. Some other studies have utilized extracurricular activities as a characteristic as well, but they have paired it with additional factors to improve prediction accuracy (Mayilvaganan & Kalpanadevi, 2015).

Finally, Naive Bayes has the lowest prediction accuracy, at 76%. Grade point average, student demography, high school history, scholarship, and social network engagement are among the characteristics considered. These properties were also used by the Neural Network and Decision Tree methods; however, the results showed that Naive Bayes provided the best accuracy. Naive Bayes works well as a predictive tool because the qualities employed are highly correlated (Xing et al., 2015).

As highlighted previously, association mining is useful in finding patterns or connections between variables in huge datasets, which can then be utilized to forecast outcomes or gather knowledge about how the variables under study behave. Finding rules that characterize the connections between various characteristics and academic performance is one possible application of association mining in this context. An association rule could read, for instance, "Students who regularly attend class and study for at least three hours per day are more likely to achieve a GPA of 3.0 or higher." This rule could be used to identify and construct interventions to aid students who may be at risk of falling behind academically.

Conclusion and future work

Predicting students' academic achievement is most helpful for enhancing the teaching and learning process. In this work, we looked back at how researchers have predicted students' grades using a variety of statistical techniques. In most studies, researchers have relied on a combination of CGPA and self-evaluation to conclude. In contrast, the classification approach is widely employed in the field of educational Data Mining to make predictions. For this reason, academics often use Neural Networks and Decision Trees, both of which fall under the umbrella of categorization approaches, when attempting to foretell students' academic outcomes. The results of the meta-analysis on predicting student performance have inspired us to do more studies for practical use in our setting. The ability to keep tabs on student progress can only benefit the educational system as a whole.

For future work: Improving student learning outcomes by extending the experiment to consider additional distinguishing characteristics. Also, further tests might need to be conducted using other Data Mining techniques to provide a wider perspective, to uncover other possible benefits, or to strengthen the accuracy of the results. Numerous programs are available to be used, and more variables to be considered.

References

Abulhul, Z. (2021). Teaching strategies for enhancing student's learning. *Journal of Practical Studies in Education, 2*(3), 1-4.

Ayesha, S., Mustafa, T., Sattar, A. R., & Khan, M. I. (2010). Data mining model for higher education system. *European Journal of Scientific Research*, *43*(1), 24-29.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: .* An Issue Brief. Office of Educational Technology, US Department of Education.

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. . *In 2013 IEEE 5th conference on engineering education* (pp. 126-130). IEEE.

Blaz, D. (2023). *Differentiated instruction: A guide for world language teachers.* Routledge.

Han, J., & Kamber, M. (2002). Data mining: concepts and techniques. *ACM Sigmod Record, 31*(2), 66-68.

Khan, A. M. (2017). Online databases usage by research scholars of the Aligarh Muslim University. . *DESIDOC Journal of Library & Information Technology*.

Malik, K. M., & Zhu, M. (2023). Do project-based learning, hands-on activities, and flipped teaching enhance student's learning of introductory theoretical computing classes? *Education and information technologies, 28*(3), 3581-3604.

Musa, M., Ahmadu, A., & Williams, C. (2024). Comparative Analysis of K-Means and Naïve Bayes Algorithms for Predicting Students' Academic Performance. . *International Journal of Development Mathematics (IJDM), 1*(3), 196-208.

Pimdee, P., Sukkamart, A., Nantha, C., Kantathanawat, T., & Leekitchwatana, P. (2024). Enhancing Thai student-teacher problem-solving skills and academic achievement through a blended problem-based learning approach in online flipped classrooms. *Heliyon Journal, 10*(7).

Pratama, M. P., Sampelolo, R., & Lura, H. (2023). Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. Klasikal. *Journal of education, language teaching and science, 5*(2), 350-357.

Qorib, M. (2024). Analysis Of Differentiated Instruction As A Learning Solution In Student Diversity In Inclusive And Moderate

Education. *International Journal Reglement & Society (IJRS), 5*(1), 43-55.

Rauf, A., Sheeba, S. M., Khusro, S., & Javed, H. (2012). Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity. *Middle-East Journal of Scientific Research, 12*(7), 959-963.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40*(6), 601-618.

Siraj, F., & Abdoulha, M. A. (2009, May). Uncovering hidden information within university's student enrollment data using data mining. In *2009 Third Asia International Conference on Modelling & Simulation* (pp. 413-418). IEEE.

Suleiman, I. B., Okunade, O. A., Dada, E. G., & Ezeanya, U. C. (2024). Key factors influencing students' academic performance. *Journal of Electrical Systems and Information Technology, 11*(1), 41.

Summer, R., McCoy, M., Trujillo, I., & Rodriguez, E. (2025). Support for working students: Understanding the impacts of employment on students' lives. . *Journal of College Student Retention: Research, Theory & Practice, 26*(4), 1123-1146.

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

Tissera, W. M. R., Athauda, R. I., & Fernando, H. C. (2006, December). Discovery of strongly related subjects in the undergraduate syllabi using data mining. In *2006 International Conference on Information and Automation* (pp. 57-62). IEEE.

Zhao, H. L. (2008). Application of OLAP to the analysis of the curriculum chosen by students. *In 2008 2nd International Conference on Anti-counterfeiting, Security and Identification* (pp. 97-100). IEEE.