

# USING DATA MINING TECHNIQUES TO MEASURE STUDENTS' ACADEMIC PERFORMANCE AND THE EFFECTIVENESS OF LEARNING METHODS: A LITERATURE REVIEW

<sup>1</sup>MARYAM AHMAD, <sup>2</sup>NOREEN IZZA ARSHAD, <sup>3</sup>ALIZA BT SARLAN

<sup>1</sup>MSc Student, Universiti Teknologi PETRONAS, Department of Computer & Information Sciences, Malaysia

<sup>2</sup>Associate Professor, Universiti Teknologi PETRONAS, Department of Computer & Information Sciences, Malaysia

<sup>3</sup>Senior Lecturer, Department Chair, Universiti Teknologi PETRONAS, Department of Computer & Information Sciences, Malaysia

E-mail: <sup>1</sup>maryam\_20000052@utp.edu.my, <sup>2</sup>noreenizza@utp.edu.my, <sup>3</sup>aliza\_sarlan@utp.edu.my

## ABSTRACT

Analyzing students' academic performance in online learning to improve the overall quality and effectiveness of education has been one of the main focuses of Higher Educational Institutions (HEIs). A practical analysis utilizing the academic performance data to improve the quality of online learning has become a vital issue urgently required to guide HEIs for the improvement of the academic performance of students. The changes affected in the Covid-19 framework have affected the academic performance of students and educators. This study aims to summarize the various aspects of educational data mining and how it can be utilized to improve the teaching process. Using EDM, the study analyzed students' academic performance for the past five years. It focused on the various learning methods that the students used. The study provided a detailed analysis of the multiple attributes that influenced the students' academic performance. We presented 12 out of 56 papers/documents that fit the inclusion and exclusion criteria of students' academic performance based on the educational setting. This study revealed that the most commonly used methods for assessing students' academic performance are not done in the face-to-face learning method.

**Keywords:** *Data Mining Techniques, Academic Performance, Attributes, Variables, Learning Methods*

## 1. INTRODUCTION

The monitoring of students' academic performance is becoming a challenge due to the recent growth in the number of Higher Educational Institutions (HEIs). Many institutions are increasingly shifting to the online and hybrid delivery approach of teaching [1]. Some of the reasons include aiding their quest for greater operational efficiency during the pandemic crisis. Therefore, it is crucial and timely to evaluate students' academic performance obtained through various delivery learning approaches that include traditional (face-to-face), online learning, and hybrid [2]. Data mining is one of the most popular techniques that could be applied to analyze students' performance and has been widely applied in the educational sector [3]. Therefore, the development

of data mining in education is popularly known as "educational data mining (EDM)" [4, 5].

EDM is a process used to extract useful information and patterns from a vast educational database [6]. Educators and institutions can use helpful information to analyze students' performance when different delivery learning methods are employed [7]. As a result, it would assist the educators in providing a practical teaching approach based on the patterns and outcomes of EDM analyses [8]. EDM also allows educators to monitor their student's achievements from the output of EDM, referring to distinct delivery approaches. Thus, HEIs can utilize the EDM application for different needs (e.g., to analyze the practical study delivery method approach and depending on the available data).

The primary purpose of HEIs offering various teaching delivery methods is to provide quality education to students, enhancing their academic performance [9]. Student's academic performance is attributed to multiple factors (or attributes) that reflect their overall academic performance [10]. However, much research on EDM has been conducted in the areas that include the educational sector, economic activities, statisticians, forecasters, and engineering and medicine. Nevertheless, limited research employs EDM to compare the effectiveness of delivery approaches [6, 7]. That concludes that it is significant to ensure that an appropriate teaching method is used for the right course and a suitable group of students [11, 12]. This study acknowledges the extensive vast work on predictive analysis that has been conducted by many researchers [13-20]; however, the intention is not to show prediction as part of this work. The attributes that have been used extensively in many previous studies for EDM include; Cumulative Grade Points Average (CGPA) [21], demographic [22], students background [23], behavior [24], and gender [25].

The current paper analyzes students' academic performance using online and hybrid learning methods and various educational data mining techniques. The analysis proposed the effective delivery learning method and attributes that helped contribute to the student's academic performance.

## 2. LITERATURE REVIEW

This section will review three delivery learning methods used for educational purposes and related attributes of students' academic performance analysis.

### 2.1 The Delivery Learning Methods

The current pandemic situation has shifted study to online and blended delivery modes. However, these techniques created challenges in assessing students' educational performance due to the lack of interaction between students and educators. This challenge increased the requirement of educational data mining techniques to consider the dimensions such as learning objective, learning preferences, learning activities, performance, participation, and achievement. These are the outcome of the students' involvement in different learning platforms and activities, enabling the researchers to decide on the result data [9].

### 2.2 Online Delivery Learning Method

Many educators utilize learning management systems to support and enhance online learning. These systems help educators monitor and track the students' learning progress [26]. These actions help in analyzing the usage of the system and its users. They also contain various data points related to the students' behavior [27]. E-learning is a fast and efficient method of learning that enables people from different parts of the world to access and develop their knowledge in a short time [28]. It is an instructional model that allows students and educators to collaborate and deliver lectures and training sessions simultaneously. This teaching method has gained widespread acceptance in education institutions [28].

Many researchers have mentioned that using that data from LMS, including students' participation, conducted activities online, and others, should be leveraged for more valuable students' academic performance analysis [21]. However, LMS data has not been very effective in analyzing the students' academic performance. The use of the EDM technique to analyze LMS data has not been explored [29].

### 2.3 Blended Delivery Learning Method

Blended learning optimizes achievement by applying technology to match each student's learning preference. The application can transfer the skill at the right time and to the specific individual to achieve these skills at the right time [30]. Many HEIs worldwide have drastically implemented blended learning as a method of teaching and learning [1]. This learning method combines face-to-face and online delivery methods to complement each other [31]. This method has several benefits, including time efficiency, cost reduction, and a convenient location for the students and the learners [32], which is much different from the educator-centered approach that focuses on the learner [33]. Nevertheless, the effectiveness of blended learning is less reported using EDM. Therefore, the educators are looking for methods to utilize the EDM technique for effectively looking at the academic performance of students when blended learning is performed [34].

Based on the studies mentioned above in traditional learning. Many studies have suggested adopting EDM to analyze students' academic performance in online and blended learning modes. The need is crucial to ensure timely, accurate, and effective performance analyses to take necessary

measures to improve the students' academic performance [29], as discussed in the next section.

## 2.4 Data Mining Techniques

Data Mining is discovering the patterns in data. The patterns discovered must be meaningful and possess some advantages in finding, identifying, and validating the patterns and using these patterns in making predictions to improve students' academic performance [35]. With the increase in educational data and technological advancement, it becomes paramount to use data mining techniques in discovering hidden information in educational data. The data mining techniques are also used to find ideas from machine learning, patterns recognition, statistics, and databases systems for applications with an enormous amount of data, the high dimensionality of data, heterogeneous data, and unstructured data [36].

Overall, educational data mining analyzes data from educational settings to better understand students and their learning environment [37]. Figure 1 illustrates data mining in education systems, which involves the development of a hypothesis, testing, and refinement. Academics and teachers carry out this process in designing, planning, and building educational systems.

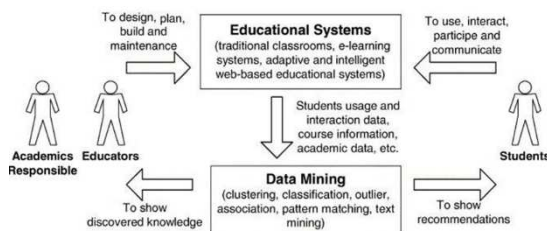


Figure 1: The cycle of applying data mining in education systems [8]

Educational data mining technique (EDM) uses different techniques in finding hidden information or relationships in educational databases, such as classification, clustering, and association. With the use of these methods, many sorts of data can be discovered that can aid educators in making a decision by HEIs based on the students' performance and can be used to analyze students' academic performance in various delivery methods [38, 39].

Nevertheless, it is reported that EDM techniques are complicated and often time-consuming [40]. Some researchers highlighted that the use of EDM could only be comprehended and conducted by practitioners in computer science or education [41]. This is due to the complexity of the

EDM techniques. Some of the data mining techniques used in the education setting based on the review done by [42] are further discussed in the following sections.

### 2.4.1 Classification

Classification has been reported as one of the most widely used data mining techniques for analyzing the educational database. Classification involves a two-step process that predicts data classification and constructs a model to classify it [43, 44]. To support that this data mining technique is one of the most extensively used, research mentioned that the classification technique had been applied from the year 2010 and to the first quarter of 2013 with 42.15%. A classifying algorithm is used to create a model that can classify different kinds of data. The model's accuracy is estimated and if it is acceptable, use it to classify new data. [45] portrayed how unique data mining procedures can be applied to improve students' performance in a learning management system.

Many researchers use many classification methods used in EDM to extract hidden patterns or relationships between attributes and classes in the dataset. [46] described four primary classification methods widely used in classification: Support Vector Machine (SVM), Artificial Neural Networks, and Decision tree. These methods are discussed in the following sections.

### 2.4.2 Artificial Neural Network (ANN)

ANN is an artificial intelligence system that mimics the functions of the mind. Each layer consists of a set of interconnected neurons. [36] applied ANN alongside other classification techniques to build up a classifier model ANN Q2H, with 91.3% performance accuracy. In comparison, Decision Tree C5.0 outperforms ANN with 92.3% performance accuracy. Furthermore, ANN is proven to have some advantages and disadvantages, as indicated by [46].

### 2.4.3 Decision Tree

Decision trees are commonly used in EDM research, and it is a tree-like structures with the nodes internally denoted by oval shapes representing the leaf nodes. This method used in educational research is the most popular [47]. The internal nodes contain the values of an expression, and each partition has a class tag linked to it. [48]. With decision trees in educational data mining, the model can then predict the students' performance based on past results [49].

#### 2.4.4 Association Rule Mining

Association rule mining is a technique utilized for analyzing market baskets. This method helps formulate policies related to the study of markets [50, 51]. The association rule is commonly used to evaluate the academic performance of students. However, it can be beneficial if the number of obtained rules is too high or not interesting. In addition, this technique can help minimize the time spent deciding the two input parameters and utilize it for data mining in educational settings [52].

#### 2.4.5 Clustering

The clustering technique is an unsupervised machine learning where similar regions are identified to discover stimulating overall distribution patterns and correlations among data attributes. Two of the most used clustering methods are Hierarchical and Partitioning, whereas k-means and hierarchical algorithms have been used simultaneously. Clustering techniques work in a concept of separating data based on their similarities where each cluster consists of similar data and is different from the other clusters [53]. K-means algorithm explores trends and patterns in a dataset to find a suitable output [37].

Despite the clustering technique being one of the highly used data mining, there have been limitations found as the approaches of clustering techniques are leaning more towards educational datasets where these datasets contain both quantitative and qualitative attributes, which could not be too suitable to use with traditional clustering approaches [54].

#### 2.4.6 Regression

One of the most basic statistical analysis data mining techniques is regression, as it investigates the functionality of relationships between the attributes [55]. The regression analysis equation is the relationship between a dependent variable and one or more independent variables. The equation is one of the most critical outputs here. The application of regression to establish or disapprove the impact of the quality of teaching is vital in educational data mining.

As mentioned before, researchers need to investigate how EDM could be effectively adapted for students' performance analyses by educators. This would allow educators to adopt this technique and achieve great results in identifying weak students, recognizing high potential ones, and supporting the mediocre to reach their full potential

[56]. Therefore, this study intends to find the most appropriate educational data mining techniques used in different learning methods to analyze students' academic performance.

### 3. RESEARCH METHODOLOGY

The review follows the procedure of [57, 58]. The overall process is organized into three significant steps (i) planning, (ii) conducting the review, and (iii) reporting the result, as summarized in Figure 2. The planning phase identified the review objectives, defined the research questions, developed a review protocol, and validated the review protocol. The conducting stage that has been adopted by [59] identifies relevant studies, selects primary studies, assesses study quality, extracts required data, and synthesizes data. The last phase reports the review results.

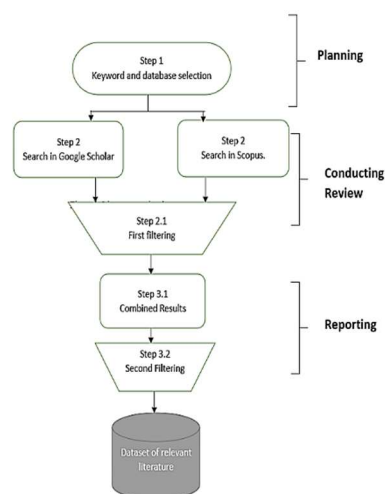


Figure 2: The retrieval process of methodology

#### 3.1 Planning Phase

This review aims to answer the following research questions.

**Q1:** What type of data mining techniques are used based on the different learning methods?

**Q2:** What data mining techniques are frequently used to analyze students' academic performance in online and blended learning methods?

To achieve the first question, this study reviewed the articles analyzing students' academic performance using data mining online and blended learning from 2015 to 2020. In each step of the search development, results were filtered according to seven selection criteria. In step 1, keywords describing the subject matter are identified. Different unique search keywords are required to ensure all relevant literature is captured through the retrieval process. During the trial search, the keywords such

as "learning +method, data mining + techniques, and students + performance" did not result in the relevant literature. Therefore, a combination of keywords is necessary to rank well in Google Scholar and Scopus. The following Table 1 shows the chosen keywords used to search for the study topics.

Table 1: List of search keywords used in the search

No	Selected search keywords
1	Learning methods AND students' academic performance
2	Data Mining Technique AND Learning methods
3	Educational data mining AND students' academic performance
4	Online Learning OR Blended Learning AND Data mining technique
5	Educational data mining technique and Students' academic performance

In step 2, as shown in figure 2, the keyword search was performed in Scopus and Google Scholar databases. In step 2.1, this step is where the study filtered the search results according to the four selection criteria indicated in Table 2. Doing so would require a lot of time and resources. This criterion was then established to limit the scope of study to only reviewing published articles.

Table 2: Criteria used in the first phase

Criterion	Description
C1	The article to be written in the English Language
C2	Studies indexed in a database other than Scopus or google scholar are out of scope.
C3	Journal and Conference papers are both considered
C4	Only articles using data mining techniques for students' performance are selected.

In step 3.1 in figure 2, the database obtained from the first filtering process is combined, and duplicates from the combination of Scopus and Google Scholar datasets are deleted. The second filtering process chose two criteria. These two are C5 and C6. The filtered database from step 3.2 yielded a few articles analyzing students' academic performance using various data mining techniques.

Table 3: Criteria used in the second phase

Criterion	Description
C5	Focuses on learning method identification, analysis of students' academic performance
C6	Focuses on data mining techniques used to analyze each learning method or type of attribute

### 3.2 Conducting Review Phase

In this phase, the results of the literature review process are presented in the following section as it yields 57 articles. Some of the published articles did not discuss the phases of students' academic

performance or the methods used in data mining using the criteria C5 and C6 as listed in table 3. We eliminated the duplicates based on the abstracts as some articles were indexed by more than one database. Of the 57 articles, 20 published articles that satisfied the requirements were selected.

The articles identified current educational data mining techniques used to analyze students' academic performance and the review according to the domain category. Based on the study, several fields that applied data mining techniques were identified and categorized into several domains: healthcare and medical data analysis, prediction, statistics, machine learning, patterns recognition, algorithms, and high-performance computing. The articles are further discussed in the following section.

### 3. FINDINGS AND DISCUSSION

The following sections discuss the answers to our research questions.

Q1. *What type of data mining techniques are used based on the different learning methods?*  
Figure 3 shows the articles related to the research topics published from 2015- to 2020. Interestingly, we found that reports on the analysis of students' academic performance using educational data mining techniques were discussed the most by researchers, comprising 35% of the total articles (57 articles). In contrast, Traditional Learning only represents 6% of entire articles, mainly on proposing of framework to analyze students' academic performance in a traditional setting.

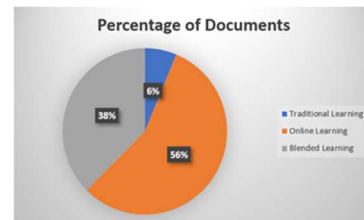


Figure 3: Documents using EDM to analyze Students' Academic Performance in Different Learning Methods

Q2. *Data mining techniques are frequently used to analyze students' academic performance in online and blended learning methods?*

We identified six methods: clustering, classification, decision tree, Naïve Bayes, Support Vector Model (SVM), and Regression. As shown in Table 4 (page 7), the most analysis done in analyzing students' academic performance using EDM techniques is in online learning. The type of EDM

techniques used by researchers to analyze students' academic performance in online and blended learning is discussed below.

### 3.3 Usage of Data Mining in Online Learning Method

Learning is a unique and challenging subject that is researched through various techniques. In 2015, a study compared the findings of two different approaches for finding salient factors that influence learning [60]. The researchers found that the two approaches can provide complementary insights into the most likely factors affecting learning.

The usage of data mining methods helped researchers predict the academic success of over 5000 students during their undergraduate years [61]. They then combined these predictions with their four-year bachelor's degree course results. The study focused on identifying low and high-achieving students. The courses that can be used to identify these individuals were selected using decision trees and clustering.

In 2018, the authors of this study noted that constructing predictive models based on student

heterogeneity could be challenging since different student characteristics might affect their learning motivation [56]. However, the experiments revealed that identifying weak students earlier helps identify them more accurately. They also noted that tree-based and rule-based methods have higher accuracy rates.

Due to the benefits of machine learning, research studies on education technology are on the rise. This study focused on fifth-year students at a university in Nigeria [62]. The fifth-year students' grade point average (GPA) is computed using the program of study, year of entry, and the grade point average during their last year of studies. In addition, a comparison of six data mining algorithms was performed, where the most accurate algorithm was the logistic regression algorithm and the least precise algorithm was the PNN algorithm.

Several studies have been conducted on the effectiveness of teaching and learning through online platforms. Assessing the learning achievements was mainly focused on students. In 2015, the authors of a study [63] revealed that outspoken students during an online course were

Table 4: Published work utilizing EDM techniques and attributes yielding interesting findings in Online Learning

Authors	Educational Data Mining Technique	Attributes	Findings	Cited by
[60]	Hierarchical Linear Modelling (HLM) Feature Selection Algorithms	Survey, Demographic, Major, GPA, Grade Level, Assessment Time & Clicks, Previous Courses	The HLM approach was most appropriate for analysing the relationships between students and learning. The Lasso approach was also used for analysing the relationships between students and instruction.	28
[61]	Decision Trees, Clustering	High School Marks, Examination marks for five courses	The policy allows teachers to automatically flag students who show signs of struggle or opportunity.	395
[56]	Naïve Bayes, Sequential Minimal Optimizer, J48	Demographic, economic status, online viewing, visiting, attempting, attendance type	Unique characteristics of different students were identified to aid in identifying the weak students.	81
[62]	KNIME (Konstanz Information Miner); PNN, Random Forest, Decision Tree, Naïve Bayes, Regression	GPA, CGPA, seven engineering departments	Engineering students in Nigeria can be reasonably expected to finish with a good graduation performance based on their academic performance in the first three years.	61
[63]	Decision Tree, Weka Data Mining	Log Files of test score	The online collaborative learning environment was hypothesized to have a positive effect on learning outcomes.	44
[37]	Clustering (K-Means)	Demographic, Academic, Parental Involvement, Interactional	The various attributes of Parental involvement and Interactional attributes vary depending on the demographic attribute of the school. They can also affect the students' learning behaviour and performance.	62

more likely to perform well in future academic pursuits. The decision tree data mining concluded that an online learning course affected learning achievements.

materials and educational data mining techniques, as shown in Table 5 (on page 10).

The use of multiple linear regression to

*Table 5: Published work utilizing EDM techniques and attributes yielding stimulating findings in Blended Learning*

Authors	Educational Data Mining Technique	Attributes	Findings	Cited by
[67]	Regression	User, Grade, Time online, activities, online session	Out of 29 usage variables analysed, 14 of them were significant. Four of them predicted that the variance in final grade would be 52%.	250
[68]	Artificial Neural Network	Notes, assignments, quizzes and other collaborative communication	The model was trained to predict the performance of students in blended learning environment with the correct classification rate of 98.3%.	54
[65]	Vector Space Model (SVM), Data Mining algorithm	Demographic data	SVM techniques can be used to collect and visualize the data related to student activity levels. They can also generate reports that can inform parents about the level of student activity.	57
[66]	Classification (Ensemble classifier)	School and students related factors	The hybrid feature selection framework for early prediction of academic failure students can filter and wrap approaches for feature selection.	1
[69]	Classification, Regression, Tree	Online Activity	Based on four learning activities, the model learned that 99.1% of the correct classifications were derived from the content creation, group wiki and online quiz taking.	23
[64]	Classification (Multiple Linear Regression)	Video-viewing exercises, quizzes, homework	They focused on seven critical factors that affect a students' academic performance. They identified four online & three traditional factors that can affect a students' performance.	40

Numerous studies have acknowledged the increasing use of analytical tools in educational data mining techniques. Among others, [37] study analyzed the effects of various learning features on students' performance; the student's dispositions are linked to their academic performance. These attributes are also dependent on their demographic and parental involvement.

The literature review findings revealed that many researchers analyzed students' academic performance in an online environment. Table 4 concludes the studies analyzing students' academic performance in online learning.

### 3.4 Usage of Data Mining in Blended Learning Method

Researchers analyzed the various aspects of blended learning during their studies, especially since the pandemic caused the Covid-19 pandemic. Many studies explore the potential of blended learning. It involves using online educational

predict academic performance was studied by [64]. The model was combined with principal component analysis to improve its accuracy. The proposed model was formulated using six components that were obtained from PCA. The components were predictive mean square error, predictive mean absolute percentage, predictive mean square error, predictive mean square error. As a result, MLR is a robust and flexible framework that can estimate a students' academic performance prediction.

For the study by [65], the researchers developed algorithms to extract data from student action logs. They can then generate a value equal to the level of activity in the students' profile. The objective of this study was to analyze the various attributes of students enrolled in an online learning environment and compare them to their activities. The data collected during the course visually represented the course performance. This paper aimed to help instructors identify students struggling and then modify their educational strategies.

A hybrid feature selection framework for the early prediction model was proposed in 2017 by [66]. A total of 45 attributes were identified and listed down to 25. Most of these attributes were related to school, institution, life, and academic status. The ensemble classifier combines multiple models and predicts a class variable using the voting technique, and the accuracy of this technique is used for the evaluation of fitness. Another study investigated the relationship between various data variables and a blended learning course [67]. The study analyzed the log files of the Moodle learning management system to identify course-related factors that have significant correlations with the final course grades.

After performing a multiple regression analysis, the study revealed that only four of the 14 variables predicted that the final grade would have a 52% variance. The model was then tested to identify which students were at risk of failing and which were not. The results indicated that the model was adequate. Aside from that, it was also studied how artificial neural networks could predict students' success based on their online activities [67]. A neural network was trained to predict students' success in passing a course. It was able to classify the students into three categories: failure, success, and overall. The study revealed that students' number of messages on social media was the most potent predictor of their course outcome. The study's goal [68] was to test the capabilities of classification and regression analysis (CART), which are commonly used in web-based learning environments. The procedure accurately classified students into those who passed the test and those who did not. The number of text messages a student sent to their teammates and instructors was most valuable in predicting course success. Table 5 concludes the studies analyzing students' academic performance in blended learning.

#### 4. CRITICAL DISCUSSION AND COMPARISON OF LITERATURE

This paper presents a review that aims to identify the attributes that affect the students' academic performance in higher education. It also explores the various data mining techniques used to analyze these attributes. The study examined 12 research articles on the subject. The main findings indicated that students' grades, class performance, e-learning activity, and demographic information influence their academic performance. Table 6 summarizes educational data mining technique attributes employed in the three learning methods

face-to-face (traditional), online, and blended learning methods.

Table 6: Summary of similar EDM attributes used in traditional, online and blended learning methods.

Authors	Educational Data Mining Technique	Attributes	Learning Methods
[70]	Classification	CGPA, quiz marks, midterm, attendance, finals	Traditional
[63]	Decision Tree	Log files of a test, school marks, examination marks	Online
[64]	Regression	Grade, time online, online session, activities, quizzes, video viewing	Blended
[68]	Artificial Neural Network	Notes, assignments, collaborative communication	Blended
[56]	Naïve Bayes	Attendance LMS activity, demographic, viewing	Online
[37]	Clustering	Demographic, parental involvement, interactional, academic	Online
[64, 69]	Classification	Quizzes, exercises, video viewing, school & students' factors and online activity	Blended

Based on Table 6, online learning uses various attributes such as log files and timestamps. These two attributes are helpful for training data and predicting the course or label of unknown data. In addition, researchers believe that developing indicators to predict their future success can improve students' academic performance. In terms of traditional learning, many of the attributes are related to the assessment and the CGPA. The techniques suitable for this type of data would be classification and clustering techniques is especially useful for these data outputs and analyses. In addition, various methods can be utilized to detect blended learning prediction patterns. These include Classification, Regression, and AI Neural networks.

Based on the literature review for this study, three key points are depicted in Table 6 and summarised below:

1. Decision on the type of EDM technique to analyze student academic performance data: No specific pattern emerges between Educational Data Mining techniques and the learning methods employed. Therefore, this study has a reason to suggest that one can use any EDM technique to analyze the data collected without focusing on the types of learning methods employed.
2. Data and attributes: This study found a pattern in the data collected based on the learning methods employed. Firstly, it was found that institutions should collect data from the learning management systems for online learning methods. The data type includes activities performed, interaction with the system, and demographic details. Secondly, educators can collect historical and numerical data for the traditional learning method to analyze students' academic performance throughout the years of



study, including demographic details or exam marks. Finally, the data can be exported historically or from a learning management system in blended learning. This data can enable researchers to analyze activities taken from both traditional and online learning methods.

3. The choice of Educational Data Mining Techniques: Based on the review analysis carried out, it is suggested that one use the clustering or classification technique for the numerical and historical type of data and attributes. The methods will depend on the kinds of attributes discovered from the published data. For example, from the learning management system, the Naïve Bayes technique can be utilized to analyze attributes such as an activity time stamp, demographic, log files, interaction with the system, or cumulative grade point average. As reported in the published literature [56, 62].

One of the most critical tasks is identifying a predictor for modeling. The study prevalent in predicting student performance [56]. Research has been conducted on the various attributes that influence a student's performance. These include their quality, social background, and academic background. However, there has been conflicting evidence about the effects of gender on student performance. It is widely believed that the quality of a student's experience and academic achievements significantly impact their performance. Other attributes such as the quality of teaching and the previous domain knowledge can also affect student motivation.

The various attributes that influence a student's academic performance relate to students' previous grades and class performance. The findings in the literature [69-72] suggest the same factors can affect a student's performance in different areas. For example, having good grades at the start of one's studies can help a student maintain an excellent academic record throughout the rest of their academic life. However, having poor grades can also affect a student's academic performance.

The social and demographic attributes are mainly responsible for the student's academic performance. Therefore, the students should take care of themselves and avoid any social activities affecting their academic performance. Moreover, the policies made by the institutions can help the students improve their academic performance by creating focus groups.

The study results revealed the most frequently used data collection techniques related to the e-learning system and the student information system. However, future work should also consider other data collection techniques, such as course evaluation or surveys, as they tend to be less used in the published literature.

The literature review discusses the various techniques used for data mining and predictive analysis, including classifications and regression. It has been observed that this technique can predict the course content and student academic performance. The results indicated that the Decision Tree, Naïve Bayes, Regression, and Artificial Neural Networks are the most used data mining techniques. These findings are supported by [73, 74]. Learning analytics help institutions identify areas of improvement and develop strategies that can improve student academic performance, engagement, and satisfaction. There are many studies examining the various advantages and challenges of this new education model, i.e., online. However, there is limited research on analyzing students' performance in a blended learning environment. Based on the literature reviewed, further research is suggested to consider other data mining algorithms, to focus on why EDM techniques are restricted to analyzing the selected attributes.

The findings of this study could help educational institutions improve their students' academic performance in either learning method by assisting them in identifying areas of weakness in their students' academic life. Therefore, the current study proposes to use demographic and educational attributes to analyze students' academic performance operating EDM.

## 5. CONCLUSION

This paper has analyzed the various attributes that affect students' performance in higher education institutions. We also identified the most used data mining techniques and algorithms to explore their academic performance over the past five years. This study carried out extensive literature, and the literature review collected necessary details for the study, including the list of questions and the exclusion criteria

Various attributes and techniques have a relationship between them that is used to evaluate students' academic performance in face-to-face, online, and blended learning environments. Different attributes would help educators improve

the quality of their teaching and learning. It would also introduce the various aspects of an e-learning management system to improve student performance. The analysis of educators can be enhanced by training in EDM.

This study would be helpful for higher education institutions to identify suitable EDM techniques to be used for the analysis of educational data of students being retrieved.

Qualitative studies such as surveys can be conducted to compare the data analytics and education case study for future work. These studies would help future works further enhance understanding of the types of educational data mining used to analyze academic attributes for students in higher education institutions. In addition, this would improve the graduates of higher education institutions to enhance their grades during their studies. Furthermore, they would be aware of their performance from the start of their semester of study.

#### ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Higher Education and the Computer & Information Science Department, Universiti Teknologi PETRONAS (UTP), for providing the facilities and the financial support to conduct this study. This work was supported by the Yayasan Universiti Teknologi (YUTP) under grant agreement no 015LC0-298.

#### REFERENCES:

- [1] Graham, C.R., W. Woodfield, and J.B. Harrison, *A framework for institutional adoption and implementation of blended learning in higher education*. The internet and higher education, 2013. **18**: p. 4-14.
- [2] Chavarría-Bolaños, D., et al., *E-Learning in Dental Schools in the Times of COVID-19: A Review and Analysis of an Educational Resource in Times of the COVID-19 Pandemic*. Odovtos International Journal of Dental Sciences, 2020. **22**(3): p. 69-86.
- [3] Brown, S., S. White, and N. Power, *Tracking undergraduate student achievement in a first-year physiology course using a cluster analysis approach*. Advances in physiology education, 2015. **39**(4): p. 278-282.
- [4] Baradwaj, B.K. and S. Pal, *Mining educational data to analyze students' performance*. arXiv preprint arXiv: 2012.
- [5] Nithya, P., B. Umamaheswari, and A. Umadevi, *A survey on educational data mining*. International Journal of Advanced Research in Computer Engineering & Technology, 2016. **5**(1): p. 69-78.
- [6] Durairaj, M. and C. Vijitha, *Educational data mining for prediction of student performance using clustering algorithms*. International Journal of Computer Science and Information Technologies, 2014. **5**(4): p. 5987-5991.
- [7] Cohen, L., L. Manion, and K. Morrison, *Action research, in Research methods in education*. 2017, Routledge. p. 440-456.
- [8] Romero, C. and S. Ventura, *Educational data mining: A survey from 1995 to 2005*. Expert systems with applications, 2007. **33**(1): p. 135-146.
- [9] Picciano, A.G., *The evolution of big data and learning analytics in American higher education*. Journal of asynchronous learning networks, 2012. **16**(3): p. 9-20.
- [10] Oyelade, O., O.O. Oladipupo, and I.C. Obagbuwa, *Application of k Means Clustering algorithm for prediction of Students Academic Performance*. arXiv preprint arXiv: 2010.
- [11] Driscoll, A., et al., *Can online courses deliver in-class results? A comparison of student performance and satisfaction in an online versus a face-to-face introductory sociology course*. Teaching Sociology, 2012. **40**(4): p. 312-331.
- [12] Huang, S. and N. Fang. *Work in progress—Prediction of students' academic performance in an introductory engineering course*. in 2011 Frontiers in Education Conference (FIE). 2011. IEEE.
- [13] Akram, A., et al., *Predicting students' academic procrastination in blended learning course using homework submission data*. IEEE Access, 2019. **7**: p. 102487-102498.
- [14] Iyanda, A.R., et al., *Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models*. International Journal of Modern Education & Computer Science, 2018. **10**(6).
- [15] Kaur, H. and, E.G. Bathla, *Student performance prediction using educational data mining techniques*. International Journal on Future Revolution in Computer Science & Communication Engineering, 2018. **4**(12): p. 93-97-93-97.
- [16] Ramanathan, L., et al., *Cluster-based distributed architecture for prediction of student's performance in higher education*. Cluster Computing, 2019. **22**(1): p. 1329-1344.

- [17] Saa, AA, M. Al-Emran, and K. Shaalan, *Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques*. Technology, Knowledge and Learning, 2019. **24**(4): p. 567-598.
- [18] Saa, A.A., M. Al-Emran, and K. Shaalan. *Mining student information system records to predict students' academic performance*. in *International conference on advanced machine learning technologies and applications*. 2019. Springer.
- [19] Wiyono, S. and T. Abidin, *Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm To Predict Student's Performance*. International Journal of Research-Granthaalayah, 2019. **7**(1): p. 190-196.
- [20] Zohair, L.M.A., *Prediction of Student's performance by modelling small dataset size*. International Journal of Educational Technology in Higher Education, 2019. **16**(1): p. 1-18.
- [21] Huang, S. and N. Fang, *Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models*. Computers & Education, 2013. **61**: p. 133-145.
- [22] Pandey, M. and S. Taruna, *Towards the integration of multiple classifier pertaining to the student's performance prediction*. Perspectives in Science, 2016. **8**: p. 364-366.
- [23] Quille, K. and S. Bergin. *Programming: predicting student success early in CS1. a re-validation and replication study*. in *Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education*. 2018.
- [24] Hong, B., Z. Wei, and Y. Yang. *Online education performance prediction via time-related features*. in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. 2017. IEEE.
- [25] Brocato, B.R., A. Bonanno, and S. Ulbig, *Student perceptions and instructional evaluations: A multivariate analysis of online and face-to-face classroom settings*. Education and Information Technologies, 2015. **20**(1): p. 37-55.
- [26] Johnson, L., et al., *The 2010 Horizon Report*. 2010: ERIC.
- [27] Dráždilová, P., et al., *Computational intelligence methods for data analysis and mining learning activities*, in *Computational intelligence for technology-enhanced learning*. 2010, Springer. p. 195-224.
- [28] Liu, C.-H. *The comparison of learning effectiveness between traditional face-to-face learning and e-learning among goal-oriented users*. in *6th International Conference on Digital Content, Multimedia Technology and its Applications*. 2010. IEEE.
- [29] Bakhshinategh, B., et al., *Educational data mining applications and tasks: A survey of the last 10 years*. Education and Information Technologies, 2018. **23**(1): p. 537-553.
- [30] Singh, H. and C. Reed, *A white paper: Achieving success with blended learning*. Centra Software, 2001. **1**: p. 1-11.
- [31] Poon, J., *Blended learning: An institutional approach for enhancing students' learning experiences*. Journal of online learning and teaching, 2013. **9**(2): p. 271-288.
- [32] Brown, R., *Blended learning: rich experiences from a rich picture*. Training and Development in Australia, 2003. **30**(3): p. 14-17.
- [33] Higgins, D. and A. Gomez, *Teaching English studies through blended learning*. New York: The Higher Education Academy, 2014.
- [34] Dakduk, S., Z. Santalla-Banderalli, and D. van der Woude, *Acceptance of blended learning in executive education*. SAGE Open, 2018. **8**(3): p. 2158244018800647.
- [35] Ian, H.W. and F. Eibe, *Data Mining: Practical machine learning tools and techniques*. 2005, Morgan Kaufmann Publishers.
- [36] Agaoglu, M., *Predicting instructor performance using data mining techniques in higher education*. IEEE Access, 2016. **4**: p. 2379-2387.
- [37] Bharara, S., S. Sabitha, and A. Bansal, *Application of learning analytics using clustering data Mining for Students' disposition analysis*. Education and Information Technologies, 2018. **23**(2): p. 957-984.
- [38] Aher, S.B. and L. Lobo. *Data mining in educational system using weka*. in *International Conference on Emerging Technology Trends (ICETT)*. 2011.
- [39] Samuel, M., M.A. Samuel-soma, and F.F. Moveh, *AI-Driven Thermal People Counting for Smart Window Facade Using Portable Low-Cost Miniature Thermal Imaging Sensors*. 2020.
- [40] Scheuer, O. and BM. McLaren, *Educational data mining*. Encyclopedia of the Sciences of Learning, 2012: p. 1075-1079.

- [41] Chalaris, M., et al., *Improving quality of educational processes providing new knowledge using data mining techniques*. Procedia-Social and Behavioral Sciences, 2014. **147**: p. 390-397.
- [42] Mohamad, S.K. and Z. Tasir, *Educational data mining: A review*. Procedia-Social and Behavioral Sciences, 2013. **97**: p. 320-324.
- [43] Ajibade, S.-S.M., NB. Ahmad, and SM Shamsuddin. *An heuristic feature selection algorithm to evaluate the academic performance of students*. in *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*. 2019. IEEE.
- [44] Kabra, R. and R. Bichkar, *Performance prediction of engineering students using decision trees*. International Journal of computer applications, 2011. **36**(11): p. 8-12.
- [45] Romero, C., S. Ventura, and E. García, *Data mining in course management systems: Moodle case study and tutorial*. Computers & Education, 2008. **51**(1): p. 368-384.
- [46] Kamavisdar, P., S. Saluja, and S. Agrawal, *A survey on image classification approaches and techniques*. International Journal of Advanced Research in Computer and Communication Engineering, 2013. **2**(1): p. 1005-1009.
- [47] Shahiri, A.M. and W. Husain, *A review on predicting student's performance using data mining techniques*. Procedia Computer Science, 2015. **72**: p. 414-422.
- [48] Ahmed, A. and I.S. Elaraby, *Data mining: A prediction for student's performance using classification method*. World Journal of Computer Application and Technology, 2014. **2**(2): p. 43-47.
- [49] Yadav, S.K., B. Bharadwaj, and S. Pal, *Data mining applications: A comparative study for predicting student's performance*. arXiv preprint arXiv: 2012.
- [50] Angeli, C., et al., *Data mining in educational technology classroom research: Can it make a contribution?* Computers & Education, 2017. **113**: p. 226-242.
- [51] Khan, A. and SK. Ghosh, *Student performance analysis and prediction in classroom learning: A review of educational data mining studies*. Education and information technologies, 2021. **26**(1): p. 205-240.
- [52] Zorilla, M.E., D. Garcia-Saiz, and J.L. Balcázar. *Towards parameter-free data mining: Mining educational data with yacaree*. in *Educational Data Mining 2011*. 2010. Citeseer.
- [53] Berkhin, P., *A survey of clustering data mining techniques*, in *Grouping multidimensional data*. 2006, Springer. p. 25-71.
- [54] Chen, L., et al., *Soft subspace clustering of categorical data with probabilistic distance*. Pattern Recognition, 2016. **51**: p. 322-332.
- [55] Kotsiantis, S.B. and P.E. Pintelas. *Predicting students marks in hellenic open university*. in *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)*. 2005. IEEE.
- [56] Helal, S., et al., *Predicting academic performance by considering student heterogeneity*. Knowledge-Based Systems, 2018. **161**: p. 134-146.
- [57] Brereton, P., et al., *Lessons from applying the systematic literature review process within the software engineering domain*. Journal of systems and software, 2007. **80**(4): p. 571-583.
- [58] Kitchenham, B. and S. Charters, *Guidelines for performing systematic literature reviews in software engineering*. 2007.
- [59] NORDIN, AAM and NMA RODZIAHLATIH, *USING SAAS TO ENHANCE PRODUCTIVITY FOR SOFTWARE DEVELOPERS: A SYSTEMATIC LITERATURE REVIEW*. Journal of Theoretical and Applied Information Technology, 2020. **98**(24).
- [60] Miller, L.D., et al., *A Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics That Impact Online Learning*. Journal of Educational Data Mining, 2015. **7**(3): p. 117-150.
- [61] Asif, R., et al., *Analyzing undergraduate students' performance using educational data mining*. Computers & Education, 2017. **113**: p. 177-194.
- [62] Adekitan, A.I. and O. Salau, *The impact of engineering students' performance in the first three years on their graduation result using educational data mining*. Heliyon, 2019. **5**(2): p. e01250.
- [63] Shukor, N.A., Z. Tasir, and H. Van der Meijden, *An examination of online learning effectiveness using data mining*. Procedia-Social and Behavioral Sciences, 2015. **172**: p. 555-562.
- [64] Yang, S.J., et al., *Predicting students' academic performance using multiple linear regression and principal component analysis*. Journal of Information Processing, 2018. **26**: p. 170-176.
- [65] Estacio, R.R. and RC Raga Jr, *Analyzing students online learning behavior in blended*

- courses using Moodle*. Asian Association of Open Universities Journal, 2017.
- [66] Krithiga, R. and D.E. Ilavarasan, *A Hybrid Feature Selection Based Framework for Early Prediction of Rural College Students Academic Failure in the Introductory Courses*. International Journal of Pure and Applied Mathematics, 2017. **115**(6): p. 157-163.
- [67] Zacharis, N.Z., *Predicting student academic performance in blended learning using artificial neural networks*. International Journal of Artificial Intelligence and Applications, 2016. **7**(5): p. 17-29.
- [68] Zacharis, N.Z., *Classification and regression trees (CART) for predictive modeling in blended learning*. IJ Intelligent Systems and Applications, 2018. **3**: p. 1-9.
- [69] Asif, R., et al., *Analyzing undergraduate students' performance using educational data mining*. Computers and Education, 2017. **113**: p. 177-194.
- [70] Burgos, C., et al., *Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout*. Computers and Electrical Engineering, 2018. **66**: p. 541-556.
- [71] Gómez-Rey, P., F. Fernández-Navarro, and E. Barberà, *Ordinal regression by a gravitational model in the field of educational data mining*. Expert Systems, 2016. **33**(2): p. 161-175.
- [72] Márquez-Vera, C., et al., *Early dropout prediction using data mining: a case study with high school students*. Expert Systems, 2016. **33**(1): p. 107-124.
- [73] Peña-Ayala, A., *Educational data mining: A survey and a data mining-based analysis of recent works*. Expert systems with applications, 2014. **41**(4): p. 1432-1462.
- [74] Vickers, N.J., *Animal communication: when i'm calling you, will you answer too?* Current Biology, 2017. **27**(14): p. R713-R715.