Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 8924028, 26 pages
https://doi.org/10.1155/2022/8924028

Review Article

Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review

Sarah A. Alwarthan, Nida Aslam, and Irfan Ullah Khan

Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

Correspondence should be addressed to Sarah A. Alwarthan; saalwarthan@iau.edu.sa

Received 6 June 2022; Revised 5 August 2022; Accepted 8 August 2022; Published 19 September 2022

Academic Editor: Manikandan Ramachandran

Recently, educational institutions faced many challenges. One of these challenges is the huge amount of educational data that can be used to discover new insights that have a significant contribution to students, teachers, and administrators. Nowadays, researchers from numerous domains are very interested in increasing the quality of learning in educational institutions in order to improve student success and learning outcomes. Several studies have been made to predict student achievement at various levels. Most of the previous studies were focused on predicting student performance at graduation time or at the level of a specific course. The main objective of this paper is to highlight the recently published studies for predicting student academic performance in higher education. Moreover, this study aims to identify the most commonly used techniques for predicting the student's academic level. In addition, this study summarized the highest influential features used for predicting the student academic performance where identifying the most influential factors on student's performance level will help the student as well as the policymakers and will give detailed insights into the problem. Finally, the results showed that the RF and ensemble model were the most accurate models as they outperformed other models in many previous studies. In addition, researchers in previous studies did not agree on whether the admission requirements have a strong relationship with students' achievement or not, indicating the need to address this issue. Moreover, it has been noticed that there are few studies which predict the student academic performance using students' data in arts and humanities major.

1. Introduction

As the volume of stored data increases, there is a need for analyzing and discovering knowledge from large and complex stored datasets. Extracting hidden useful knowledge from datasets has become highly important in many fields during this competitive world. However, data mining, also known as Knowledge Data Discovery (KDD) enables discovering the hidden patterns and extracting useful and nontrivial information from the vast amount of stored data [1]. Data mining is one of the fast-growing areas in computer science and statistics. The strength of data mining lies in using several techniques that can be applied to different fields, including health [2–4], education [5, 6], engineering [7], marketing [8], and business [9]. More precisely, data mining can be defined as one of the Knowledge Data Discovery (KDD) process' phases that involves the following steps, as shown in Figure 1, which can be applied to discover and find interesting patterns from the stored data [10].

The data mining process goes through several stages, which start with preprocessing the targeted dataset. It is necessary to focus on the data related to the problem by applying the data selection method to obtain a set of data related to the problem from the databases. In the preprocessing stage, the data will be cleaned and processed from all of the issues it suffers from in order to make it suitable for applying the mining techniques. The next stage is processing the data, where the data mining techniques are used to extract useful hidden patterns. The final stage is focused on interpreting and evaluating the extracted patterns that were

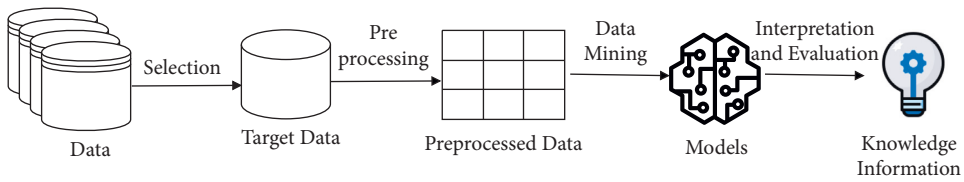

Figure 1: Data mining process.

discovered from the previous step. To identify and evaluate the interesting patterns, appropriate measures of the data mining methods will be used [10].

In recent years, researchers from different fields (such as computer, statistics, and education), have taken an interest in the study of academic problems faced by students in higher educational institutes to find potential solutions. Applying data mining techniques on educational data is known as Educational Data Mining (EDM). Educational Data Mining (EDM) [11] focused on implementing different techniques on the large amount of data collected from educational institutes for understanding the student's behavior and having a significant impact on improving the learning process and outcomes. Educational data not only include academic grades (course grade and GPA) but also data acquired from online platforms which include learning management system (LMS), demographic data (age, nationality, and gender), and admission data (entry test and high school grade) [12]. Researchers have applied different data mining techniques on educational data for educational purposes. The widely used data mining techniques are prediction (classification and regression), structure discovery (clustering), and relationship mining (association rule mining, sequential pattern mining, and correlation mining) [13, 14].

Recently, many researchers are interested in educational data mining techniques and conducted several studies that contributed to the improvement of the educational process. Several issues have been addressed to enhance the educational process, such as identifying the student at risk of failure or drop out and predicting students' academic level at an early stage to provide the necessary support for the at-risk students. This paper provides the foundation knowledge on applying educational data mining techniques in order to predict student performance. Moreover, this study identifies the gaps and conflicts in the previous studies. The main contribution of this paper is to highlight the recently published studies for predicting student academic performance. Moreover, this paper aims to answer the following questions:

(i) What are the most common techniques used in the previous studies to predict the student's academic performance

(ii) What are the highly influential features for predicting the student's academic performance

(iii) What is the most targeted student major in the previous studies

(iv) What are the existing gaps in the current published studies

The remaining part of this study is organized as follows: section 2 presents the background of the data mining process; section 3 presents the methodology used for searching, filtering, and reviewing the previous studies; section 4 highlights the previous studies which predict student academic performance; section 5 discusses the findings; sections 6 and 7 highlight the existing gap and the challenges encountered in previous studies to predict student performance; and finally, the conclusion is presented in section 8.

2. Background

This section briefly presents the background of the data mining process and the knowledge related to the techniques commonly used in educational data mining. The main objective of this section is to provide a brief explanation of the most common techniques used to predict students' academic level at higher education and to discuss the frequent evaluation methods used to evaluate the classification models.

2.1. Educational Data Mining Process. This section provides a brief description of data mining stages that are commonly used in mining educational data. The data mining process contains several stages as shown in Figure 2, where each stage involves a set of techniques that are used according to the targeted problem. The following figure illustrates the main stages of the educational data mining process that will be briefly described in this section.

2.1.1. Data Collection. Data collection is the process of gathering and collecting educational-related data that is stored in educational institutes' repositories. Higher educational institutions hold a vast amount of student-related data that have been gathered since the student is enrolled in one of the university programs. Data related to program enrollment requirements are collected about students that include high school grade, Scholastic Achievement Admission Test (SAAT) score, and General Aptitude Test (GAT) score as a prerequisite for enrollment. There is a set of data collected while students study the program which include the grades of their courses and their majors. Students Information System (SIS) is one of the student data sources where many student-related data such as demographics

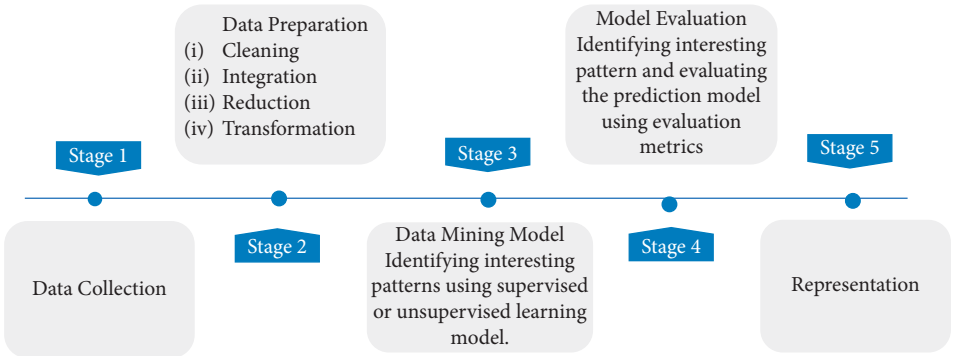

Figure 2: Stages of educational data mining.

information (age, gender, and nationality) and academic performance (university and preuniversity grades) are stored [5, 15]. On the other hand, social and economic characteristics data are not available in the SIS, as they are collected by using one of the data collection methods such as a questionnaire [16, 17]. Furthermore, information about a student can be collected while using an online learning management system (LMS), such as the Blackboard in order to obtain course materials and grades, to participate in the discussion, and to attend online exams and assignments. All the previously mentioned data are collected from the on-campus learning process. It has been observed from previous studies discussed in section 4 that most of the reviewed studies used academic performance and demographic information data to construct a prediction model that determine student's level at higher education.

2.1.2. Data Preparation. Preprocessing has a critical role in data mining. The primary goal of the data preparation stage is to make the raw data suitable for data mining techniques to be implemented. Due to the huge size of the educational databases, the data stored in the educational databases usually face some problems that affect the quality of data. In order to increase the quality of the data, it is imperative to implement data cleaning methods to handle missing, inconsistent, and outlier data. Data preprocessing improves the quality of data, thus enhancing mining results. This section introduces the essential data preprocessing methods, where data cleaning can be applied to remove noise and handle missing values. Data integration combines data from multiple sources and stores them into a single source. Data reduction can be applied to reduce the volume of the dataset by removing the redundant and irrelevant features. In contrast, data transformation can be applied to scale the data into a smaller range. During the preprocessing stage, the following tasks are performed to enhance the performance and accuracy of the prediction model [1, 18]. All the previously mentioned steps are not mandatory, and it depends upon the dataset.

(1) Data Cleaning. Usually, the stored data face some problems such as incomplete, noisy, and inconsistent data; therefore, it needs to be cleaned. Data cleaning attempts to fill the missing values, detect outliers, and remove noise. Missing data are those incomplete data that can be fixed by eliminating the tuple (where this method is used when the target label is missing, or when the tuple has many missing values), filling the missing value in the numeric attribute with the mean, median, or mode of the attribute, and filling the missing value in the nonnumeric attribute with the mode (most occurs) value of the attribute. Noisy data are also known as random errors that can be removed using the binning methods (e.g., smoothing by bin means, smoothing by bin boundaries, and smoothing by bin median). Detecting and eliminating outliers is the process of identifying an anomaly object that differs in its behavior from other objects. Identifying outliers can also be performed using graphical representation methods such as boxplots or unsupervised learning methods such as clustering.

(2) Data Integration. It is the process of merging several data from several sources such as multiple databases and files for the purpose of data analysis. Sometimes, different data sources contain the same variable with different names. Once data sources are merged, this results in a data redundancy (data duplication) problem. Data redundancy is a common problem in datasets that may have identical values for two different attributes names. Moreover, data integration leads to inconsistent data using different measurement units for the same attribute in several databases (e.g., the first database stores the weight in kilograms and the second database stores the weight in pounds). In addition, using different data encoding (e.g., the gender attribute's value in the first database ("male" or "female") and in the second database ("M" or "F")) is another issue in the data integration process. During this phase, several techniques will be applied to handle these inconsistencies and duplications.

(3) Data Reduction. The primary purpose of implementing data reduction methods is to get a smaller representation of the original dataset that produced either the improved or almost the same analytical results. Several strategies were considered in the data reduction stage, including numerosity reduction, dimensionality reduction, and data compression. Using smaller data representations to replace the original

data volume is called numerosity reduction, where parametric and nonparametric models were used. While dimensionality reduction is the method of decreasing the number of dataset features using several techniques related to data compression, feature selection, and feature construction. In data compression, the original data are transformed and compressed to obtain a reduced representation (compressed data). When the original data can be recovered from the compressed data without losing any information, the data reduction is lossless. Otherwise, it is known as lossy data reduction, where the recovered data are approximately similar to the original data.

(4) Data Transformation. In the data transformation process, the data is converted into another form that is suitable for the data mining process. Appling data transformation has a significant contribution in improving the efficiency of the mining process and understanding patterns easily. Data normalization and discretization are some of the commonly used data transformation techniques. Normalization is a widely used method that transforms original attribute's values to fall in a smaller data range. Data discretization methods are used to replace numeric attribute's values (e.g., height) with nominal values (e.g., tall, medium, or short).

2.1.3. Data Mining Model. Predictive and descriptive analyses are the two primary objectives of data mining. Descriptive analysis is used to mine data and provide valuable information about past or current events without the class attribute. It is also known as unsupervised learning techniques, including clustering and association pattern mining. However, clustering is the process of grouping objects into several groups to provide insight into data. Clustering differs from classification and regression in the learning type as it used unsupervised learning techniques. The clustering techniques used unlabeled training datasets to group the objects into classes based on increasing the similarity between the objects within the same group and decreasing the similarity between the objects within different groups. While the association rule mining technique discovers the interesting hidden relationship between the attributes where the frequent patterns are initially generated, then the rule will be extracted from the frequent itemset that fulfills the stated criteria (the value of minimum confidence is greater than a predefined threshold) [19].

Predictive analysis is commonly used to forecast unknown or future values by using historical data to make the decision [10]. Classification and regression are the main types of supervised learning techniques used for prediction which the model learns using the training labeled dataset to predict the label or the value of the unknown testing sample. Classification is a widely used supervised learning technique that maps a specific input to the categorical target class [14]. However, regression assigns a particular input to a continuous value.

2.1.4. Classification Model Evaluation. This section introduces the evaluation measures that are commonly used in assessing the classifier's performance. For the evaluation process, the test samples that were not used in model construction will be used to evaluate the prediction model. Usually, the original dataset is divided into two independent parts, training and testing, where the training set is used to fit the model. In contrast, unseen testing samples are used to evaluate the performance of the constructed model. In the cases of having large-sized datasets, researchers divide the dataset into three partitions, which are training, validation, and testing sets. The training set is used to build the prediction model, and to search for the optimal value for the hyperparameters, the validation set is used. The third set is for testing, where it is used to evaluate and compare the model's performance. Holdout, random subsampling, and cross-validation are the common methods used for data partitioning.

In the holdout method, the original dataset is randomly divided into two separate datasets: a training set and a testing set. The partitioning percentage varies based on the size of the dataset. Often, 70% of the data are used for training to derive the prediction model, and the remaining 30% is used for testing where the model's accuracy is estimated, whereas in random subsampling, the holdout method will be repeated several times, and the final accuracy of the model's estimation is the average accuracy of all the iterations [20].

Besides, cross-validation is frequently used with parameter tuning methods. In cross-validation, the dataset is randomly split into k equal folds/subsets where the training and the testing are applied k times. In each iteration, $k-1$ folds are used to fit the model, and the remaining one fold is used to estimate the performance of the prediction model. Leave-one-out is a special case of cross-validation where it is used when having a small-sized dataset. The dataset of size n splits into n folds. In each iteration, one sample is used for the testing and the remaining part for training. In general, it is recommended to use stratified partitioning where the class distribution of the samples in each partition is approximately similar to the original dataset [20].

A confusion matrix is a widely used model evaluation measure for classification problems. The confusion matrix shows how the proposed model can distinguish samples from different n classes where $n \geq 2$. True Positive (TP) and True Negative (TN) refer to the correctly classified sample, while False Positive (FP) and False Negative (FN) refer to the misclassified samples. The widely used evaluation model measures accuracy, sensitivity, specificity, precision, and F1-score [20] can be defined using the confusion matrix as follows:

(i) Accuracy is calculated by dividing the number of students who are correctly classified over the total number of entire students:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (1)$$

(ii) Sensitivity (also known as true positive rate or recall) shows how many positive class/at-risk students were correctly classified, and it is calculated using the following formula:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (2)$$

(iii) Specificity (also called as true negative rate) shows how many negative class/not at-risk students were correctly classified, and it is calculated using the following formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (3)$$

(iv) Precision shows the percentage of the positive class/at-risk student which is labeled as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (4)$$

(v) F1-score or F-measure combines the recall and the precision in a single measure defined as follows:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5)$$

The accuracy measure is preferred if the dataset is balanced (the dataset has an equal number of instances in each class). Otherwise, sensitivity, specificity, and F1-score can be used with the imbalanced dataset.

3. Survey Methodology

The adopted methodology in this paper focuses on recent studies published during the past five years until the start of April 2022. This study focused on reviewing the published studies that aim to predict student's performance in higher education using educational data mining techniques. Several studies that discussed predicting student's academic performance during distance learning were excluded as this paper focused on predicting student's academic performance during traditional (face-to-face) learning. A group of keywords was considered to search within specific databases, including IEEE Xplore, Scopus, ACM Digital Library Journals, Google Scholar, and Web of Science. The search terms used to search the databases can be represented as follows: (predict* or forecast* or identify*), (at-risk student or student or first-year student), and (machine learning or data mining). After obtaining several highly related articles, the resulting studies were grouped into different categories based on the desired research objectives. Furthermore, the previous studies within each category were arranged based on the chronological order as presented in the next section. Figure 3 presents the review methodology used in this paper.

4. Predicting Student's Academic Performance at Higher Education

Many studies related to educational data mining have been conducted during the last several years. These previous studies were targeting different objectives such as predicting the student dropout, student's success on the course level, student's achievement at the graduation time, and student's performance at the end of the academic year. Therefore, the literature review is based on study objectives. This section presents the previous studies conducted in the last five years to predict the students' academic performance using various educational data mining techniques.

4.1. Predicting Student Dropout. Authors in Refs. [21–28] predicted the students' performance and whether or not the student will drop out. Researchers in Ref. [21] discovered the patterns that contribute to prevent at-risk students from drop out. This study aimed to detect whether or not the student will be evaded by using the decision tree (DT) for classification and genetic algorithm (GA) for feature selection. The proposed model was applied on the dataset that contains 12,969 instances collected from 106 undergraduate courses. The results showed an average precision equal to 98%. Also, they found that students with a grade point average (GPA) less than 5.79 (10 scale GPA) and who have enrolled for more than a year are more likely to drop out.

Moreover, a group of researchers in Ref. [22] proposed a stacking ensemble model that combines RF, eXtreme gradient boosting (XGBoost), gradient boosting (GB), and artificial neural networks (ANN) to predict the student that may be at risk of being a drop out at an individual course. In this study, we used 261 students' samples and 12 attributes collected from 2016 to 2020 at Constantine the Philosopher University in Nitra. Different features relevant to students' academic success were considered in this study including information about access, tests, tests grade, exam, project, project grade, assignments, result points, result grade, graduate, year, and academic year. The proposed stacking ensemble model achieved the highest performance, which is 92.18% accuracy.

Similarly, the random forest (RF) has been applied in Ref. [23] to identify the relevant features that influence student drop out. Authors in Ref. [23] calculated the importance of each feature using mean decrease accuracy (MDA) and mean decrease Gini (MDG) measures. The proposed dataset includes 206 first-year informatics engineering students' records and 40 features, including students' academic achievement in the first semester, university test results, and demographics features. After applying the feature selection, only seven attributes related to the first-semester academic factors and parents' income were selected to build the final model. The proposed prediction model was built using the DT classification algorithm. The final DT classification model overfitted where it achieved 97.21% accuracy for training and 81.01% accuracy for testing.

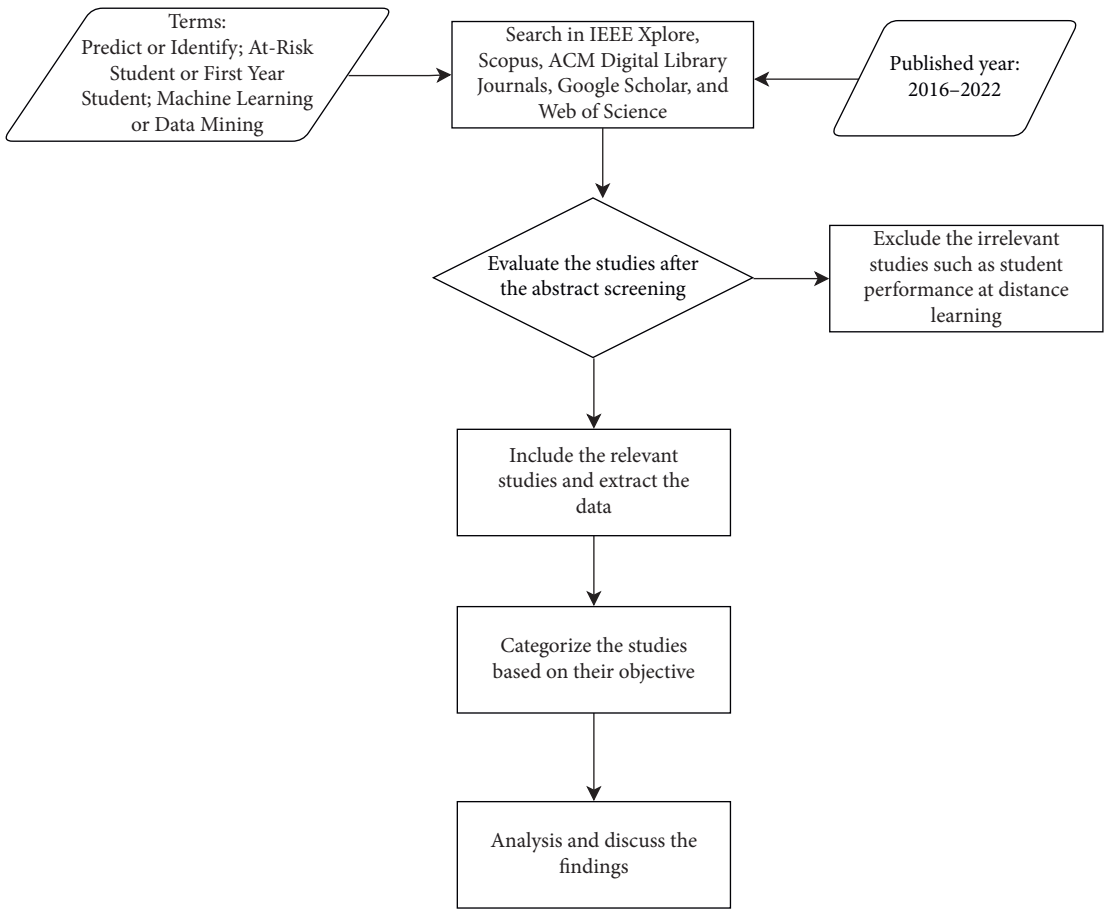

Figure 3: Review methodology.

Moreover, researchers in Ref. [24] identified students who are more likely to drop out by applying the clustering method on the data collected from 561 undergraduate students through an online survey. The authors used Bayesian profile regression (BPR) as a classification technique. They found that the highest group of students at drop out risk was characterized by facing difficulties in understanding and studying, with lack of ability to face difficulties and challenges, scored low exam grades, and had only low motivation level.

A group of researchers in Ref. [25] combined support vector machine (SVM), naïve Bayes (NB), and DT to produce an ensemble model. The proposed model identifies the student that may be at risk of being a drop out. In this study, 499 students' responses and 50 attributes collected through a questionnaire have been analyzed. Different features relevant to students' academic success, behavioral, demographic, and social issues were considered in the study. The proposed ensemble model achieved the highest performance, which is 99% accuracy. They found that the previous semester's percentage was the most influential attribute in students' academic achievement.

Another study [26] developed two models to predict student achievement and major based on the first-year courses. The first model was built to identify whether the student will complete the program or not. In contrast, the second model was introduced to predict among which of the 71 majors, the student will enroll using the first-year courses information. Logistic regression (LR) and three forms of the RF technique have been evaluated to build the prediction model. The proposed model used an available dataset collected from Toronto University for 65,000 students' grades. After preprocessing, 38,842 students' grades were used to train the first prediction model, where 26,488 students' records are labeled as completed the program and 12,294 as not completed the program (dropout). To train the second model, only 26,488 students' records marked as completed the program are used to predict the student's major. The common RF package in *R* obtained the highest output among the three RF versions as it achieved 78.84% accuracy for predicting whether or not the student will complete the program and 47.41% accuracy for predicting student's major. They found that the RF achieved higher results than LR.

A study in Ref. [27] proposed a predictive model trained and tested using 10,196 students' records. The collected dataset contains 41 features, including first-semester performance, secondary school achievement, and student demographic features. Three machine learning techniques were evaluated, namely, Gradient Boosted Trees (GBT),

eXtreme Gradient Boosting (XGB), and ANN, to predict students' status whether they drop out at the end of the academic year or not. They found that the best predictive model achieved 85.8% accuracy using ANN and considering the data related to the first-semester academic performance.

Furthermore, research [28] aims to find out the best way to predict student's performance by applying different EDM techniques on the real data of 104 students that were collected from Universitas Islam Indonesia's information system. Bayesian network (BN) and DT classification algorithms and five feature selection methods were used to predict students who are most likely to drop out. As a result of applying features selection, they found that the accuracy of the prediction model was enhanced. Moreover, university features that include attendance and GPA of the first semester have the highest impact on the student performance compared with other features such as personal information, family information, and preuniversity characteristics. The highest accuracy of 98.08% was achieved by using the Bayesian network technique.

The reviewed studies aimed to predict at-risk students of being drop out are summarized in Table 1.

4.2. Predicting Student's Achievement in the Course Level. In addition, authors in Refs. [29–46] applied several educational data mining techniques to predict student achievement in the course level. A study in Ref. [29] developed a prediction model to predict the final exam grades of undergraduate students. A group of machine learning algorithms including RF, SVM, LR, NB, ANN, and k-nearest neighbor (KNN) was compared to classify the final exam grades into four classes: "<32.5," "32.5–55," "55–77.5," and "≥ 77.5." The proposed model used a dataset collected from a state University in Turkey for 1854 students' grades of the Turkish Language-I course. Three features related to the student which are the midterm exam grades, department, and faculty have been used to predict the final exam grades. The RF and ANN obtained the highest performance compared to other classification models as they achieved 74% accuracy and an area under curve (AUC) of 86% for classifying the final exam grades.

Moreover, the students at risk of failing the course were identified at an early stage in a research conducted by Ref. [31]. Student's performance during the course was divided into different stages for the earliest prediction: 0%, 20%, 40%, 60%, 80%, and 100% of the course length. The authors used the Open University Learning Analytics Dataset (OULAD) which is a public dataset with 32,593 student records and 31 variables, such as student demographics, course assessment scores, and student online participation. They used six machine learning techniques to build the prediction model, including k-nearest neighbor (KNN), RF, SVM, ExtraTree (ET), AdaBoost, gradient boosting classifier, as well as one deep learning technique called deep feed forward neural network (DFFNN). The proposed model divides students into four categories: withdrawn, fail, pass, and distinction. Prediction model's performance was improved using the feature engineering technique. By grouping the withdrawn and fail labels to form the fail class label and combining the pass and distinction labels to form the pass class label, the multiclasses problem was transformed to a binary classes problem. Finally, the results showed that RF achieved the best performance, scoring 79% for precision, recall, F-score, and accuracy at 20% of the course duration. Furthermore, at 60% of the course duration, RF enhanced the model's performance to 88% precision, recall, F-score, and accuracy. The RF received a 92% precision and 91% for recall, F-score, and accuracy at 100% of the course length.

A study carried out by Ref. [32] used the bagging ensemble method with eight machine learning techniques including KNN, RF, SVM, LR, and NB, and three distinct topologies of ANN to determine at-risk students who would fail the course. This study used two datasets from two different courses to classify students into one of three categories: good, fair, or weak. The first dataset provides the assessment marks for a group of 52 engineering students in one course from an e-learning platform, while the second dataset contains different task grades for 486 science students, including assignments, quizzes, and exams from an e-learning platform. The results revealed that the bagging ensemble model performs the best performance, where the first dataset's bagging model achieved an accuracy of 66.7% during 20% and 50% of the coursework, while the second dataset's bagging model achieved 88.2% and 93.1% accuracy when considering 20% and 50% of the coursework.

Additionally, researchers in Ref. [33] applied several machine learning methods, including NB, DT, LR, SVM, KNN, sequential minimal optimization (SMO), and ANN to forecast student's performance at the end of the course would be pass or fail. The model was trained and tested using the dataset collected from a computer science college under the University of Basra for 499 students and using ten attributes, including student ID, gender, date of birth, employment, activity grade, and exam grade. LR outperformed other classifiers as it scored 68.7% accuracy for a passed class and 88.8% accuracy for a failed class. They found that many factors including the data cleaning, the type of features, and the dataset size influenced the accuracy of the results achieved by the final model.

The study [34] used the first two weeks of formative assessment activity (exercise and homework) grades to identify students who were at risk of failing the final exam of an introductory programming course. The RF ensemble technique was used to deploy the prediction model, which was used to identify students as at risk of failing the final exam or not. The dataset, which contains two predictors for 289 students enrolled in a programming course, was used to train and test the proposed RF model. The proposed classification model overfitted, achieving 72.73% training accuracy and 59.64% testing accuracy.

Furthermore, authors in Ref. [35] proposed a new hybrid data mining approach based on classification and clustering techniques to predict student performance. Four classification algorithms, which are SVM, NB, DT, and ANN methods, were applied to find the best subset of features from the dataset collected from Kerala educational institutions. The best subset of features was used as an input for

Table 1: Summary of the related studies on predicting drop out among students.

Ref	Techniques	Results	Study sample size	Findings
[21]	DT	Precision (98%)	12969	Student having a GPA <5.79 is more likely to drop out
[22]	Ensemble	Accuracy (92.18%)	261	The ensemble model improves the prediction performance and solves the overfitting issue
[23]	DT	Accuracy (81.01%)	206	Six features related to the first-semester academic factors and parents' income are the most influencing factors
[24]	Clustering & BPR	—	561	Low exams grades and motivation level may lead to evasion.
[25]	Ensemble	Accuracy (99%)	499	The previous semester GPA is the most influencing attribute in students' academic achievement
[26]	RF	Accuracy Model_1 (78.84%) Model_2 (47.41%)	38,842	The common RF package in *R* obtained the highest output
[27]	ANN	Accuracy (85.8%)	10,196	The best predictive model was achieved by considering the data related to the first-semester academic performance
[28]	BN	Accuracy (98.08%)	104	Student's attendance and GPA are the highest impacted features on student's performance

the K-means clustering technique to classify student performance into low, medium, or high classes. The accuracy of the proposed hybrid model outperformed other methods where it scored 75.47%. Furthermore, they found that the student's behavior and academic success are strongly related.

Moreover, researchers in Ref. [36] predicted students' GPA for a course based on the previous courses' achievement. Several classification and regression techniques were used, including the sequential minimal optimization algorithm for regression (SMOReg), RF, linear regression (LR), multilayer perceptron (MLP), KNN, Gaussian processes random tree (GPRT), DT, and simple logistic regression (SLR). The proposed dataset was collected from the student information system (SIS) at United Arab Emirate University (UAEU). The collected data include 145 student samples representing student records in secondary school grades and 32 courses' GPA. SMOReg achieved the highest accuracy of 96.98% accuracy when considering all previous grades. After applying the feature selection method, 20 features were used to build the final model, and RF achieved 96.4% accuracy when using the highly influenced courses on the GPA.

The study mentioned in Ref. [37] examined the factors affecting the final exam result for an English course. 101 freshmen records were analyzed to discover the influencing factors by applying Spearman's correlation coefficient and the back-propagation neural network (BP-NN) technique. National College Entrance Examination (NCEE) score with 0.731 correlation coefficient and learning attitude with 0.471 correlation coefficient were the factors that had the highest influence on the final exam results.

Similarly, the authors in Refs. [30, 38, 39] evaluated different machine learning techniques for the same objective. The studies used the same public dataset that was collected from the learning management system (LMS), which contains 480 student records for different courses and different school levels. Researchers in Ref. [38] applied five classification methods such as RF, NB, multilayer perceptron (MLP), SVM, and J48 (DT). They found that MLP got the highest accuracy of 76.07%, followed by SVM with 75.49%. However, researchers in Refs. [39] used three classification algorithms such as DT, NB, and ANN, and the highest accuracy of 78.1% was achieved by applying ANN. They found that student's attendance and parents' participation greatly impacted the student's performance level at the end of the semester.

Moreover, the authors in Ref. [30] compared different machine learning techniques including RF, ANN, NB, KNN, DT, bagging, and XGBoost. The study used a same Jordan public dataset that contains 16 independent variables and 480 student records for different courses and different school levels. The prediction model classified the student who may drop out of the course into three classes: high, average, and low. Finally, they found that the bagging ensemble method got the highest accuracy (99.40%), the highest precision (99.51%), and the highest recall (99.42%) compared to other classification methods.

Using a small dataset containing 38 master student records, authors in Ref. [40] tried to predict a dissertation project grade for master students. The heat map and hierarchical clustering techniques are used to visualize the relation between the features. Moreover, authors used a group of classification methods, namely, MLP-ANN, NB, SVM, KNN, and linear discriminant analysis (LDA) to build the prediction model. The SVM achieved the highest accuracy of about 76.3% for forecasting the dissertation grade. The study proved the efficiency of visualization and clustering in identifying the courses that influenced the dissertation final grade. Furthermore, they found that the preadmission information (age, bachelor GPA, and specialization) significantly impacts student grades compared to other attributes.

Authors in Ref. [41] applied an ensemble classification model that predicts student's achievement whether the student passes or fails in two courses: mathematics and Portuguese. The study applied on the educational dataset, which was obtained from UCI. The datasets contain 1044 instances and 12 features (student grades, demographic, and social and school related features) after applying an information gain-based selection algorithm. Three classification

methods, such as DT (J48), NNge, and MLP, were used with the ensemble method. The highest performance reached 95.78% accuracy by applying an ensemble method with DT on a balanced dataset with the selected features.

In addition, authors in Ref. [42] evaluated four classification techniques which include NB, LR, KNN, and RF to detect students that were more likely to fail the course. The real dataset was collected from an educational tool called Xorro-Q. The dataset contains assessment scores of 240 students for a 12-week-duration course. The result showed that the first test conducted in the fifth week did not detect the student who may fail the course. Moreover, by considering the second test in the tenth week, the final exam prediction accuracy has improved. RF achieved the highest performance with 74% F1-score.

Authors in Ref. [43] evaluated eight supervised machine learning regression methods, namely, linear regression, RF, KNN, M5 rules, M5 algorithm, SMOreg, Gaussian processes (GP), bootstrap aggregating (Bagging) to predict second-semester final exam grades based on the first-semester achievement. The proposed model uses a dataset containing 592 student records and 16 attributes, including demographic characteristics, final examination grades for the first semester courses, and the number of attempts that the student fails in the first-semester course. The Friedman Aligned Ranks Test was applied to evaluate the performance of the prediction model. The results showed that RF provides the best results followed by SMOreg and bagging. RF achieved an MAE ranging between 1.198 and 1.910 for predicting the grade of the second-semester's six courses.

In addition, the authors of Ref. [44] used ANN, DT (J48), SVM, and NB to predict whether students will fail programming classes or not. This research used two different datasets. The first dataset includes 161 student records from traditional (face-to-face) learning, while the second dataset has 262 student records from online distance learning. The SVM had the best performance with an F1-score of 92% for the distance dataset and 83% for the on-campus dataset, as a result of preprocessing the datasets and searching for optimal parameters (fine-tuning). Unlike on-campus learning, where a student had to finish at least a quarter of the course, the SVM was able to predict student performance when the student completed at least half of the distance education course time.

Furthermore, the study conducted in Ref. [45] aimed to predict the grade of the future courses by applying user-based collaborative filtering (UBCF), matrix factorization, singular value decomposition (SVD) and non-negative matrix factorization (NMF)), and restricted Boltzmann machine (RBM) techniques on a real data of 225 undergraduate students. The Pearson correlation was applied to find the relation between the preadmission requirements and student performance. The result showed that the preadmission factors such as CGPA, higher secondary school certificate (HSSC), and an entry test could predict student's performance. Also, the RBM achieved the highest performance with 0.3 RMSE, followed by NMF, SVD, and then UBCF in predicting the course GPA.

Moreover, researchers in Ref. [46] proposed a web-based application system that used a group of supervised machine learning methods, namely, DT, ANN, and NB to classify the student academic performance into four classes (excellent, good, average, and poor). The proposed model trained and tested using a dataset contains 700 students' records and 19 features including gender, parent's qualification, family income, and the 10^{th} and 12^{th} grades. The final NB classification model outperformed other techniques as it achieved the highest accuracy compared to DT and ANN models.

Table 2 summarizes the previous studies that focus on predicting student performance at the course level.

4.3. Predicting Student's Achievement at Graduation Time. Several studies were made to predict student's achievement at graduation time [5, 6, 47–51]. The authors of Ref. [5] analyzed 339 computer college students at Imam Abdulrahman Bin Faisal University to classify student's CGPA during graduation based on the preparatory year's success. To classify students into three classes including high, average, and below average, the proposed model used tree-based classification algorithms such as J48, random tree, and REPTree. When the optimal parameter's value for J48 was used and the number of features was reduced to 4 out of 14 features, the final prediction model achieved 69.3% accuracy. Authors found that the most influential variables on the graduation CGPA are the first-year CGPA, an introductory math course, a computer skills course, and a communication skills course.

Moreover, a study carried out by Ref. [47] compared two classification methods, namely, DT (C4.5) and NB to predict the student's CGPA at graduation time. This study used three datasets of 25, 50, and 79 students' samples for graduate informatic engineering students to classify students into one of four categories: with praise, very satisfactory, satisfactory, and enough. The results showed that the NB method outperformed DT (C4.5), where it achieved higher performance with an average accuracy of 73.41% and an area under curve (AUC) of 66.4%.

Furthermore, 15 classification methods were used in this study [6] to predict computer college students' final CGPA. The findings revealed that the seven classifiers such as NB, Hoeffding Tree, SMO, RF, LMT, simple logistic, and KNN outperformed the average accuracy of the 15 classifiers. A computer college dataset with 530 records and 64 features, including the final CGPA class, was used to build the proposed classification model. NB and Hoeffding Tree had the highest accuracy score of 91%. Authors found that the operating systems, statistics, general physics, computer programming, and algorithm courses have a significant impact on the CGPA.

The impact of the first-three-year GPA on the final CGPA was investigated in a study that used a data of 1841 engineering students [48]. The final CGPA was classified using several classification algorithms, including probabilistic neural network (PNN), RF, DT, NB, tree ensemble, and LR, with a high accuracy of 89.15% using LR. Furthermore, the authors found that the third-year GPA is the most influenced feature, followed by the second and first-year GPAs.

Table 2: Summary of the related studies on predicting student's level in the course.

Ref	Techniques	Results	Study sample size	Findings
[29]	RF & ANN	Accuracy (74%)	1854	The results showed that the RF and ANN outperformed SVM, LR, NB, and KNN.
[31]	RF	Accuracy (79%–91%)	32,593	The feature engineering technique improved the prediction model where it achieved more than 80% accuracy.
[32]	Ensemble (bagging)	Accuracy (66.7%) for dataset_1 (93.1%) for dataset_2	Dataset_1 : 52 Dataset_2 : 486	The highest results were achieved when considering 50% of the coursework.
[33]	LR	Accuracy (78.75%)	499	Several factors including data cleaning, the type of features, and the dataset size influenced the final model's accuracy.
[34]	RF	Accuracy (59.64%)	289	The formative assessment tasks grades were able to predict at-risk students.
[35]	Hybrid (DT and K-means)	Accuracy (75.47%)	—	Student's behavior and students' academic success are very strongly related.
[36]	RF	Accuracy (96.4%)	145	Regression methods outperformed the classification in predicting course GPA.
[37]	BP-NN	—	101	NCEE score and learning attitude have a major influence on the final exam results.
[38]	ANN (MLP)	Accuracy (76.07%)	163	—
[30]	Ensemble (bagging)	Accuracy (99.40%)	480	Ensemble models outperformed other classification algorithms.
[39]	ANN	Accuracy (78.1%)	480	Student's attendance and parents' participation have a great impact on student's performance level.
[40]	SVM	Accuracy (76.3%)	38	Preadmission information influences the student grades
[41]	Ensemble method with DT	Accuracy (95.78%)	1044	A balanced dataset and the selected features increase the model's accuracy.
[42]	RF	F1-score (74%)	240	The accuracy of the prediction model has improved when considering the 2nd test.
[43]	RF	MAE (1.198 to 1.91)	592	RF provides the best result followed by SMOreg and bagging.
[44]	SVM	F1-score (92%) for distance dataset F1-score (83%) for on-campus dataset	Online dataset: 262 On-campus dataset: 161	The accuracy of the prediction model has improved when the student performed 50% of distance education course and 25% of on-campus education course.
[45]	RBM	RMSE (0.3)	225	CGPA, HSSC, and an entry test can predict student performance.
[46]	NB	—	700	Demographic information and school grades allow predicting a college student's level.

Similarly, another study [49] evaluated multiple classification algorithms to see if the preadmission requirement and personal information had an impact on the final GPA. To predict the final GPA level, two DT (C4.5 and ID3), NB, and KNN classification algorithms were applied. The experiment was conducted on 2281 undergraduate students, and the result showed that NB is an efficient algorithm with a 43.18% accuracy. In addition, the preadmission requirement and personal student information were shown to have the greatest impact on the graduation GPA.

Another study [50] used 100 students' data collected from King Saud University (KSU) to identify the low-performance students at an early stage. This study's primary goal is to identify the courses from the first three years that impact the final graduation GPA. A decision tree (ID3) was applied to predict the final GPA based on the first three years' GPA. The highest accuracy achieved was about 80% by using the second-year courses' grades to predict the final GPA.

Moreover, authors in Ref. [51] evaluated the preuniversity exams, including SAAT, GAT, high school GPA, and their influence on predicting the graduation CGPA. Nine hundred and fifty-seven student records from a computer department and six attributes, including university CGPA, school GPA, GAT, SAAT, graduation year, and enrolled year for graduate and undergraduate students at King Saud University (KSU) in KSA, are used to fit in the proposed LR model. The regression model scored a correlation coefficient around 0.75 and MAE of 0.27 for graduate students. Also, for the undergraduate student, a correlation coefficient of 0.64 and MAE of 0.17 was obtained by the regression model. The authors found that students' graduation GPA is not affected by admission requirements such as SAAT and GAT. However, they found that the high school GPA impacts the student's graduation GPA.

Table 3 shows a summary of the previous studies that predicts student's achievement on completion of the degree program (at graduation time).

4.4. Predicting Student's Performance at the End of the Academic Year. Recently, several studies [15–17, 52–59] have focused on predicting the student's performance at the end of the academic year. A study published in Ref. [52] developed a three-predictive model that was trained and tested

Table 3: Summary of the related studies based on predicting student's achievement at graduation time.

Ref	Techniques	Results	Study sample size	Findings
[5]	DT (J48)	Accuracy (69.3%)	339	The CGPA of the first year and three courses of the first year: Introductory math, computer skills, and communication skills are the most influence factors on the graduation CGPA.
[47]	DT and NB	Accuracy (73.41), AUC (66.4%).	79	The findings showed that the NB outperformed DT in predicting the graduation CGPA.
[6]	NB and Hoeffding Tree	Accuracy (91%)	530	Four courses have a significant influence on the CGPA: operating systems, statistics, general physics, computer programming, and algorithms course.
[48]	LR	Accuracy (89.15%)	1841	Third year GPA is the highest influencing feature on the final year graduation GPA.
[49]	NB	Accuracy (43.18%)	2281	Preadmission requirement and the personal student information influence the graduation GPA
[50]	DT	Accuracy (80%)	100	Second year course grade is the highest influencing feature on the final graduation year GPA.
[51]	LR	Correlation coefficient (64%), MAE (0.17)	957	Preuniversity exams such as SAAT and GAT do not influence the student's GPA, whereas the high school GPA affects the student's GPA.

using data from 9652 students at a Portuguese higher education institution. The data were gathered three times: during entry time, the end of the first semester, and the end of the first year. Several prediction models were built using four classification algorithms such as RF, DT, ANN, and SVM, and also during various collection times. The first model was built by using 30 features that was collected at the time of enrollment, the second model was constructed by using 44 features that was collected at the end of the first semester, and the third model used 68 features which was collected at the end of the first year. The first-year students were classified into binary classes: "failure" and "success." Finally, the results showed that the SVM achieved 77% and 91% AUC for the first and second models, whereas the RF and SVM scored an equal performance which is 93% for the third model.

Family information variables were utilized in a study by Ref. [53] to predict freshmen student's performance at the end of the first semester in the first academic year. The authors analyzed a dataset from a Taiwanese university that included 2407 student records and 18 independent variables of personal information, such as demographic variables, parents' occupations, and family income. They used four machine learning algorithms, namely, DT (CART), DT (C5.0), RF, and ANN to develop the prediction model for different output scenarios. The first scenario classified the students into five categories: excellent, very good, good, average, and poor. In the second scenario, the student was divided into three classes: excellent, normal, and poor. In the last scenario, the student was categorized as either excellent or poor. Finally, the model achieved the best results by using binary class labels to predict student's performance. They also found that RF and DT (CART) performed the best, where DT (CART) scored 80% accuracy and RF scored 79.9% accuracy. Furthermore, the results showed that the mother's occupation, department, father's job, primary source of living expenses, and admission status are the most important factors for predicting students' learning achievement at the end of the first semester.

A group of researchers in Ref. [54] developed a prediction model for classifying students' academic performance as pass or fail. ANN, KNN, k-means clustering, NB, SVM, LR, DT, and voting ensemble were all used to develop the proposed model. The collected dataset from academic institutions in the UAE includes 1491 student records and 13 variables such as gender, age, school system, math level, English level, and scholarship. The voting ensemble model scored the best performance with an overall accuracy of 75.9%. Finally, the findings revealed that using the synthetic minority oversampling technique (SMOTE) to balance the proposed dataset enhanced the prediction model's accuracy.

Moreover, authors in Ref. [17] designed a new prediction model that combined numerous classifiers, such as the NB, SVM, and DT classifiers, with bagging and stacking ensemble methods to classify student achievement into one of four categories: excellent, good, average, and poor. The proposed ensemble model was trained and tested on a dataset of 233 examples and 45 variables grouped into student personal information, learning pattern, behavior, emotional, and cognitive features. The proposed ensemble model scored the best accuracy of 97%, followed by bagging, stacking ensemble, NB, then SVM, and finally DT.

The author in Ref. [15] proposed a classification model for predicting students' CGPA of the first academic year using preadmission data. The dataset was obtained from three computer departments at Princess Nourah bint Abdulrahman University (PNU) in KSA. The collected data includes 1,569 students' records, three admission criteria as independent variables, and the CGPA for the first academic year. Four machine learning techniques, namely, ANN, DT, SVM, and NB, were applied in this study. ANN outperformed the other methods, where it scored 79% accuracy and 81% precision. DT achieved the highest F1-measure and recall rates, which are 80% and 81%, respectively. The author found that the Scholastic Achievement Admission Test (SAAT) is the most influential factor on the CGPA, where it scored the highest correlation coefficient among the admission attributes.

Similarly, authors in Ref. [55] conducted a study on a dataset which contains 9,458 student records and 14 attributes, including demographics and environment features, to classify the CGPA into low, medium, or good classes. Two datasets were used in this study; the first dataset represents students' achievement in the first year, while the second one represents the students' achievement in the second year. The gain ratio feature selection method was used to decrease the number of features to be 14 instead of 22. Two stages of classification are applied to predict student's performance. The authors evaluated several classification techniques at the first stage, including NB, SMO, ANN, KNN, REPTree, partial decision trees (PART), and RF. At the second stage, the misclassified samples were eliminated from the datasets using C4.5, and then the same classification methods were applied. The authors found that the SMO outperformed other classifiers when considering the misclassified samples, while the RF scored the highest performance after eliminating the misclassified observations. RF obtained the highest efficiency in F1-score, precision, and recall, which was 97.1% after the misclassified samples were removed.

Researchers in Ref. [56] classified the student performance at the end of the first academic year to know the impact of the admission criteria on the student performance. A dataset of 1445 undergraduate students was used in this study. Different techniques were applied in this experiment, including RF, tree ensemble, DT, NB, LR, and MLP. LR received the highest accuracy of 50.23% with the KNIME platform, and ANN got the highest accuracy of 51.9%. In addition, the result showed that there is a weak relationship between the admission criteria and the student's academic success.

Another study [57] aimed to support students in the first year of higher education by applying different statistical techniques to predict at-risk students. The second objective of this work was to study the relationship between noncognitive attributes and student performance. Structural equation modeling (SEM) was applied on 519 female students to study the relation between the noncognitive features and student performance. The authors did not find a strong correlation between noncognitive attributes and student performance.

Similarly, another study [58] applied the MLP to identify students who are most likely to be at risk in the second year based on their performance in the first year. The study used 300 student samples, and the results show that MLP got 95% accuracy with the training set and 85% accuracy on the testing set. Authors found that CGPA of the first year is the most influencing attribute on the student academic performance.

Furthermore, another study [16] proposed a classification model to predict student's achievement. This study used a dataset which contains 161 students' responses from Computer Science and Information Technology (CSIT) college and 60 questions related to personal relationship, study skills, demographic, social, and academic information. Three decision tree-based classifiers, namely, J48, random tree, and REPTree, were applied in this study. Moreover, the CorrelationAttributeEva method was used to find the highly correlated attributes with the class attribute. The authors noted that the student demographic feature such as age, gender, and country are not highly correlated with the class attribute, whereas the GPA, credits, and the father's work have a significant impact on the target attribute. Finally, the J48 algorithm obtained the best performance where it achieved 62.9% precision and 63.4% recall when considering all the features. Besides, the J48 algorithm scored 61.6% precision and 62.1% recall when using the highly correlated attributes.

Moreover, authors in Ref. [59] proposed a forecasting system that could help the instructors to predict the students' academic performance for the fourth year student based on their achievement in the third year. The experiment was applied on a small dataset that contains 50 records. The highest accuracy achieved was 74% using improved ID3.

There is increasing interest especially in applying deep learning techniques (DL) in educational fields to predict students' achievement. Authors in Ref. [60] examined and compared the ability to apply machine learning (ML) classifiers and deep learning (DL) techniques to predict enrollment in university courses. They used several ML and DL models to analyze 309 instances for students' age, gender, and high school performance. The study's result presents that the performance of machine learning classifiers and deep learning are approximately similar, where the SVM achieved an accuracy of 90.60% and TensorFlow (BoostedTrees) scored 90.32% accuracy.

The previous works related to predicting student achievement at the end of the academic year are presented in Table 4.

5. Findings and Discussion

This section presents the most relevant findings of the studies that have been reviewed to achieve the main objective of this paper and addresses the questions previously stated.

5.1. Data Mining Techniques for Predicting Student Performance. Several educational data mining techniques have been used in previous studies. A total of 45 studies were reviewed to answer the aforementioned questions. Figure 4 shows that most of the reviewed studies used classification techniques (39 studies, 87%), while a small number of studies used regression techniques (4 studies, 9%) and clustering techniques (2 studies, 4%).

Table 5 summarizes the data mining methods that have been applied in the previous works. The most common classification techniques used to predict student achievement in previous studies are DT, NB, ANN, SVM, KNN, RF, LR, and the ensemble methods. DT, NB, ANN, SVM, KNN, RF, LR and ensemble were used in 24, 20, 20, 14, 11, 14, 9, and 9 studies, respectively. After we counted the number of times that common algorithms scored the highest performance, we noticed that some algorithms were used frequently, but they did not achieve the best performance in previous studies (e.g., KNN was applied in eleven studies, but did not achieve the highest result in any of these studies).

Table 4: Summary of the related studies on predicting student's achievement at the end of the academic year.

Ref	Techniques	Results	Study sample size	Findings
[52]	SVM for model 1 and model 2 RF and SVM for model 3	Model 1: AUC (77%) Model 2: AUC (91%) Model 3: AUC (93%)	9652	They found that SVM outperformed RF, DT, and ANN.
[53]	DT (CART)	Accuracy (80%)	2407	Mother's job, department, father's job, the main source of living expenseS, and the admission status are the highest influential features.
[54]	Ensemble	Accuracy (75.9%)	1491	The accuracy of the prediction model was improved after applying the SMOTE technique to balance the proposed dataset.
[17]	Ensemble	Accuracy (97%)	233	The proposed ensemble model outperformed bagging, stacking, NB, SVM, and DT.
[15]	ANN and DT	ANN model: Accuracy (79%) Precision (81%). DT model: Recall (80%) F1-Measure (81%).	1,569	The SAAT is the most influence factor on the CGPA as it scored the highest correlation coefficient among the admission attributes.
[55]	RF	F1-score, precision and recall (97.1%)	9,458	The SMO outperformed other classifiers before eliminating the misclassified samples while the RF scored the highest performance after eliminating the misclassified observations.
[56]	ANN	Accuracy (51.9%)	1445	The relation between admission criteria and student academic success is weak.
[57]	Structural equation modeling (SEM)	—	519	There is no relation between the noncognitive attributes and student performance.
[58]	MLP-ANN	Accuracy (85%)	300	First year GPA is the most influencing attribute on student's performance in the second year.
[16]	DT (J48)	Precision (62.9%), Recall (63.4%)	161	Student demographic features are not highly correlated with the class attribute, whereas GPA, credits, and father's work have a significant impact on the target attribute.
[59]	Improved ID3	Accuracy (74%)	50	—
[60]	SVM	Accuracy (90.60%)	309	The results show that the performance of machine learning and deep learning techniques are similar.

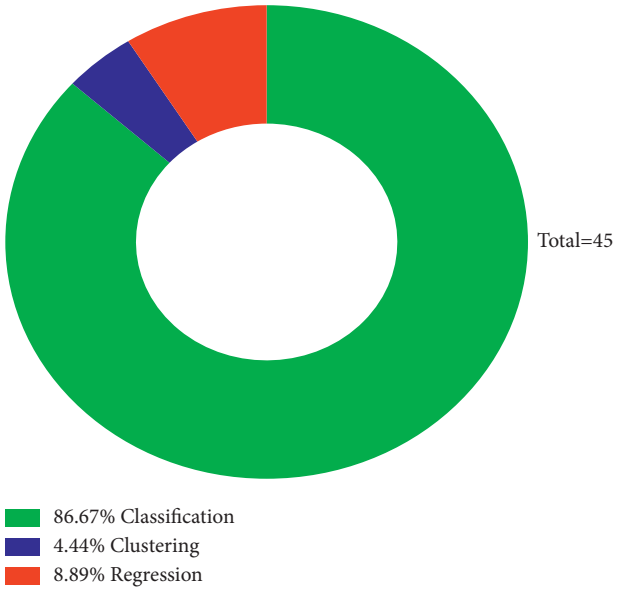

Figure 4: Educational data mining task used in the reviewed studies.

Figure 5 shows the number of times common classification methods were used versus the number of times they achieved the highest performance. It has been noted that the ensemble technique was used in nine reviewed studies to predict student's achievement, and the ensemble outperformed other techniques as it recorded promising results in seven of the aforementioned studies. Moreover, the RF algorithm was applied in fourteen previous studies and achieved high results in eight of them, indicating the efficiency of this algorithm in predicting student performance. On the other hand, DT, ANN, NB, and SVM were used frequently in previous studies as they showed high results in some studies and less promising results in others.

5.2. The Influential Factors for Predicting Student's Performance. This section presents the main student features that are commonly used for predicting student's achievement in higher education. Student features used in the reviewed studies are categorized into different groups, namely, demographic features (personal information and family information), academic features (university and preuniversity features), and social and behavioral features. Figure 6 shows that university features are the most commonly used attribute where 31 studies of the reviewed studies considered this attribute. Personal features and preuniversity features were considered in 26 studies and 18 studies, respectively. Cumulative grade point average

Table 5: Data mining techniques used in previous works.

DM task	DM technique	Studies	Number of studies (percentage of occurrence)
Classification	DT	[5, 6, 15–17, 21, 23, 25, 28, 30, 33, 35, 41, 44, 47–50, 52–56, 59]	24 (53%)
	NB	[6, 15, 17, 25, 29, 30, 32, 33, 35, 40, 42, 44, 46–49, 54–56, 60]	20 (44%)
	ANN	[6, 15, 22, 27, 29, 30, 32, 33, 35, 37, 40, 41, 44, 48, 52–56, 58]	20 (44%)
	SVM	[6, 15, 17, 25, 29, 31–33, 35, 40, 44, 52, 54, 60]	14 (31%)
	KNN	[29–33, 40, 42, 49, 54, 55, 60]	11 (24%)
	RF	[6, 22, 26, 29–32, 34, 42, 48, 52, 53, 55, 56]	14 (31%)
	LR	[6, 29, 32, 33, 42, 48, 54, 56, 60]	9 (20%)
	Ensemble	[17, 22, 25, 30, 32, 41, 48, 54, 56]	9 (20%)
	Bayesian network (BN)	[28]	1 (2%)
	eXtreme gradient boosting	[22, 27]	2 (4%)
	AdaBoost	[31]	1 (2%)
	Gradient boosted trees	[22, 27, 31]	3 (7%)
	ExtraTree	[31]	1 (2%)
	SMO	[33, 55]	2 (4%)
	Linear discriminant analysis	[40]	1 (2%)
	NNge	[41]	1 (2%)
Regression	SMOReg (SVM)	[36, 43]	2 (4%)
	Simple logistic regression (SLR)	[36]	1 (2%)
	DT	[36]	1 (2%)
	RF	[36, 43]	2 (4%)
	Linear regression	[36, 43, 51]	3 (7%)
	KNN	[36, 43]	2 (4%)
	ANN	[36]	1 (2%)
	Gaussian Processes random tree	[36, 43]	2 (4%)
	Ensemble (bagging)	[43]	1 (2%)
	M5	[43]	1 (2%)
	M5 rules	[43]	1 (2%)
	Collaborative filtering (CF)	[45]	1 (2%)
	Matrix factorization (MF)	[45]	1 (2%)
	Singular value decomposition (SVD)	[45]	1 (2%)
	Restricted Boltzmann machines (RBM)	[45]	1 (2%)
Clustering	K-means	[21, 35, 54]	3 (7%)

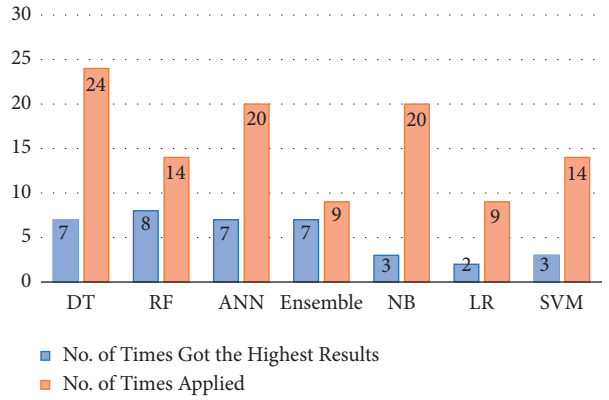

Figure 5: The most common classification techniques.

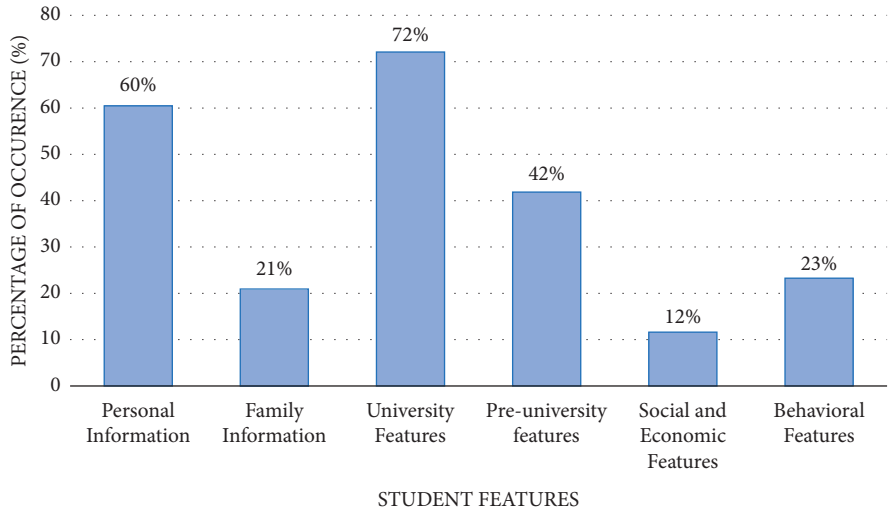

Figure 6: Influencing factors for predicting student's academic performance.

(CGPA) and assessment grades are university features that are frequently used as indicators to predict student performance. A group of studies [21, 23, 24] found that CGPA is the most influential attribute to identify at-risk students who may drop out. Furthermore, assessment grades are considered in several studies (i.e., [29, 32, 34, 42] and [44]) as influential factors to predict student achievement at the end of the course. Moreover, preuniversity features and student grades in previous courses have a high impact on studies in Refs. [6, 49] and [51] that predicted student's achievement level at graduation. Additionally, studies in Refs. [15, 58] and [16] showed that the CGPA of the previous year and preuniversity requirements have a significant impact on forecasting the student's achievement level at the end of the academic year. It has been noticed that there is no study that predicts the student's academic performance at the end of the academic year by using the assessment task grades. Table 6 lists the student attributes that are considered in the reviewed studies to predict student achievement.

5.3. The Predicted Values (Output). As mentioned earlier, the reviewed studies were classified into four categories according to the purpose of the study, which aims to classify students into two classes (binary class) or more than two classes (multiclass) based on the targeted problem. A set of classification models has been proposed to predict student performance. As shown in Figure 7, a total of 39 studies applied classification techniques to forecast student's achievement, where 90% of the models classified the student's achievement outcome into two to four classes. In contrast, the remaining 5% classified the student outcome into more than four classes, and 5% of the classification models did not specify the class labels' number. The majority of the previous classification models (18 studies, 44%) classify students into two classes, such as at-risk and not at-risk, dropout and complete, or pass and fail. On the other side, nine studies (22%) and ten studies (24%) focused on classifying students into three and four classes, respectively. All the outputs of the previously developed models are listed in Table 7.

5.4. Students' Sample Scope in Experimental Datasets. All the reviewed studies used at least one historical dataset collected from educational institutions in order to train and test the prediction model or to find the influential factors on student outcomes. Most of the reviewed studies used a dataset collected from traditional classroom learning, while some studies have an additional dataset that was collected from virtual learning environments. Figure 8 shows the number of reviewed studies that used datasets containing 30 to 40000 student samples, where most of the studies used a dataset containing less than 600 student samples. On the other hand, few studies used datasets containing more than 2000 student samples. Besides, most of the reviewed studies (18 studies, 41%) used student data from STEM (Science, Technology, Engineering, and Math) fields, whereas (9 studies, 20%) used student records from computer-related fields. Figure 9 shows that the humanities field received the least attention in the reviewed studies compared to other fields (2 studies, 5%). Table 7 summarizes the important information about the experimental datasets used in previous studies.

6. Further Discussion

A group of studies, such as in Refs. [15, 40, 45, 49, 51], has examined the relationship between the admission requirements such as the preuniversity test and the student achievements. However, researchers have not agreed on whether the admission requirements have a strong relationship with student achievement or not. There are some studies ([15, 51]) which are conducted in the Kingdom of Saudi Arabia to find out whether the admission requirements affect the student's performance or not. Both studies were conducted on computer science students. However, their findings were different. The first study [51] found that the student's GPA of the secondary school affects the student's performance in higher education, while the preuniversity tests such as SAAT and GAT do not affect the student's performance. However, Ref. [15] showed that the admission test, namely, SAAT, significantly predicts the student's CGPA. Another study such as Ref. [26] used the

Table 6: Factors for predicting student's performance.

Student features' categories		Student features values	Studies	No. of studies (percentage of occurrence)
Demographic features	Personal information	Gender, age, date of birth, nationality, motivation level, ethnicity, and employment.	[5, 6, 16, 17, 21, 25, 27, 28, 30, 31, 33, 35, 37, 40, 41, 43, 44, 46, 52–57, 59, 60]	26 (60%)
	Family information	Father's qualification, mother's qualification, father's occupation, and mother's occupation.	[16, 23, 28, 30, 41, 46, 52, 53, 55]	9 (21%)
Academic features	University features	Assessment grades (final exam, project, midterm exam, quizzes, lab grades), student attendance, and CGPA.	[5, 6, 16, 21, 23, 25–29, 31–34, 36, 40–46, 48–52, 54, 55, 58, 59]	31 (72%)
	Preuniversity features	High school courses grades, high school CGPA, preadmission requirement, and entry test.	[5, 6, 15, 27, 28, 37, 41, 45, 46, 49–52, 54–57, 60]	18 (42%)
Extra features	Social and economic features	Impact of friends and family income.	[23, 25, 30, 41, 53]	5 (12%)
	Behavioral features	Discussion participates and self-study time.	[17, 25, 30, 31, 35, 37, 44, 46, 53, 58]	10 (23%)

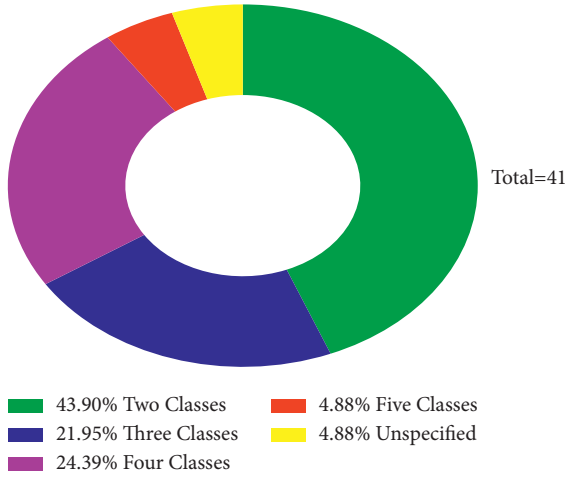

Figure 7: Classification model outputs (class labels).

Table 7: General information about the experimental dataset of the reviewed studies **(BDC: Before data cleaning, ADC: After data cleaning, DS: Dataset).

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[21]	12969	28	1. Personal information 2. University features	1.1 Age 2.1 GPA 2.2 Total course hours 2.3 Final status 2.4 Years enrolled 2.5 Language exam 2.6 Writing exam 2.7 Science exam 2.8 Humanities exam 2.9 Maths exam Withdraws 2.10 Applied year 2.11 Applied semester	Binary class: dropout/graduated	Classification	Unspecified
[23]	206	40	1. University features 2.Social & economic features	1.1 Weighted average grade 1.2 Academic index 1.3 Average success rate 1.4 Success-to-failure ratio 1.5 Efficiency 1.6 Progress factor 2.1 Family income quintile	Binary class: dropout/retention	Classification	Engineering

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[24]	ADC: 561	18	1. Student's academic competencies 2. Motivations 3. Academic resilience 4. Student satisfaction	1.1 Delay index 1.2 Average grade 1.3 Course year 1.4 Perceived delay index 1.5 Difficulty in understanding 1.6 Difficulty in memorization 2.1 Motivation level 2.2 Employment chances 3.1 ARS item1 . . . ARS item 6 4.1 Satisfaction with the study method 4.2 Satisfaction with achievements 4.3 Satisfaction with lectures 4.4 Satisfaction with learning	9 clusters	Clustering	Unspecified
[25]	499	51	1. Family information 2. University features 3. Social 4. Behavioral features	1.1 Qualification of father 1.2 Mother's income 2.1 Previous semester's percentage 2.2 Present semester attendance 3.1 Impact of friends' circle 3.2 Extracurricular activities and location of residence 4.1 Self-study time 4.2 Behavior of assignment completion	Binary class: at-risk/not at-risk	Classification	Computer
[26]	38,842	>15	1. University features	1.1 Course credit (maths, chemistry, finance, economics, first-year seminar, etc.) 1.2 Course numeric grade	Binary class: completed/dropout	Classification	Arts & science
[27]	10,196	41	1. Personal information 2. Preuniversity features 3. University features	3.1 First-semester academic achievement	Binary class: dropout/ graduation	Classification	Technology & economics
[28]	BDC: 178 ADC: 104	12	1. Personal information 2. Family information 3. Preuniversity features 4. University features	1.1 Gender 1.2 Origin 2.1 Father's occupation 2.2 Mother's education 2.3 Mother's occupation 3.1 Senior high school department 4.1 First-semester attendance 4.2 First-semester GPA	Binary class: dropout/not dropout	Classification	Engineering

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[31]	32593	32	1. Personal information 2. Behavioral features 3. Academic features	2.1 Clickstream features: mean of clicks, sum of clicks 3.1 Assessment features: average score, relative score	Multiclass: withdrawn, fail, pass, and distinction	Classification	Unspecified
[32]	ADC: 52	17	1. University features	1.1 Exercise grades (exercise 1.1 to exercise 3.5.	Multiclass: good, fair, and weak	Classification	Engineering
[32]	486	71		1.2 Tasks grade (first two assignments, first quiz, and midterm exam)	Multiclass: good, fair, and weak	Classification	Science
[33]	Adc: 499	11	1. Personal information University features	1.1 Gender 1.2 Date of birth 1.3 Employment 2.1 Course 2.2 Registration 2.3 Activity grade 2.4 Exam grade	Binary class: pass/ fail	Classification	Computer
[34]	289	3	University features	1.1 Assessment tasks (exercise and homework) grades	Binary class: at-risk/not at-risk	Classification	Computer
[35]	Unspecified	17	1. Personal information 2. University features 3. Behavioral features 4. Extra features	2.1 Student grade level 2.2 No. of days absent 2.3 Course title 3.1 Raising hands 3.2 Visiting resources 3.3 Viewing announcement 3.4 Discussion participates 4.1 Parent answering survey 4.2 Parent school satisfaction	Multiclass: low, medium, or high	Classification & clustering	Various fields
[36]	145	33	1. University features	Five courses grades (PRVT-333, Maj OP-2, PRVT-338, PUBL-220, and PUBL-2223)	—	Regression	Various fields
[37]	101	5	1. Personal information 2. Preuniversity features 3. Behavioral features	2.1 National College Entrance Examination (NCEE) 3.1 Learning attitude	Multiclass: fail, pass, good, or outstanding	Classification	Art (English)
[40]	DS1: 38	23	1. University features	Grades for Course-2, Course-1, Course-5, Course-6, and Course-3	Multiclass: A, B, C, and F	Classification	Computer & engineering
	DS2: 273	25	2. Personal information	Age, bachelor GPA, and specialization			

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[41]	1044	33	1. University features 2. Personal and family features 3. Social features 4. Preuniversity features	1.1 Grade of the second term 1.2 Grade of the first term 1.3 No. of fails 1.4 No. of absent days 2.1 Father's qualification	Binary class: pass/ fail	Classification	Mathematics & Portuguese
[42]	240	10	1. University features	1.1 Test 1 in week 5 1.2 Test 2 in week 10	Multiclass; high, medium, or low risk	Classification	Engineering
[43]	592	19	1. University features 2. Personal information	1.1 Grades for eight courses from the first semester 1.2 No. of attempts student fails the course 2.1 Gender	—	Regression	Technology
[44]	DS1: 262 DS2: 161	23 16	1. University features 2 Personal information Behavioral features	1.1 Exam 1 1.2 Activity of week 5 1.3 Activity of week 2 1.4 Activity of week 4 1.5 Activity of week 3 1.6 Blog 2.1 City 2.2 Age 3.1 Access 3.2 Exercises	Binary class: fail/ pass	Classification	Computer
[45]	225	28	1.Preuniversity features 2. University features	1.1 Entry test 1.2 HSSC	—	Regression	Engineering
[46]	700	26	1. Personal information 2. Family information 3. Preuniversity features 4. University features 5. Behavioral features	1.1 Gender 2.1 Parent's qualification 2.2 Family income 3.1 Grade in 10^{th} class 3.2 Grade in 12^{th} class 4.1 Grades in each semester in the program	Multiclass: poor, average, good, or excellent	Classification	Computer
[5]	339	15	1. Preuniversity features 2. University features 3. Personal information	2.1 Preparatory year's CGPA 2.2 Introductory math grade 2.3 Computer skills grade Communication skills grade.	Multiclass: high, average, or below average	Classification	Computer

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[6]	530	64	1. Preuniversity features 2. University features 3. Personal information	3.1 operating systems grade 3.2 Statistics grade 3.3 General physics grade 3.4 Computer programming grade 3.5 Algorithms grade	Multiclass: excellent, very good, good, or pass	Classification	Computer
[48]	1841	6	1. University features	1.1 3rd year GPA 1.2 2nd year GPA 1.3 1st year GPA 1.4 enrollment program 1.5 entry year	Multiclass: first class, second class upper, second class lower, or third class.	Classification	Engineering
[49]	2281	7	1. University features 2. Preuniversity features	1.1 Previous background 1.2 Type of admission 2.1 Loan 2.2 Gender 2.3 Talent	Multiclass: excellent, very good, good, or poor.	Classification	Unspecified
[50]	100	Unspecified	1. Preuniversity features 2. University features	Second-year courses including 2.1 Human computer interaction course 2.2 Information security 2.3 Networks 1 2.4 Data mining 2.5 Intelligent systems 2.6 Advanced human computer interaction 2.7 Operating systems 2.8 Networks 2	Binary class: weak/good GPA	Classification	Technology
[51]	957	6	1. Preuniversity features 2. University features	1.1 High school GPA	Unspecified	Classification	Technology
[54]	1491	13	1. Personal information 2. Preuniversity features 3. University features	1.1 Gender 1.2 Age 1.3 Ethnicity 2.1 School system 2.2 Math level 2.3 English level 2.4 Scholarship 2.5 Transfer status 2.6 Admitted on probation 2.7 In dorm 3.1 Program 3.2 Course load	Binary class: pass/fail	Classification	Unspecified
[17]	233	45	1. Personal information 2. Learning pattern 3. Behavioral features 4. Emotional and cognitive features	Unspecified	Multiclass: excellent, good, average, or poor	Classification	Engineering & art

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[15]	1,569	4	1. Preuniversity features	1.1 GAT 1.2 SAAT 1.3 High school GPA	Multiclass: excellent, very good, good, average, or poor.	Classification	Computer
[55]	9458	14	1. Personal information 2. Family information 3. Preuniversity features 4. University features	1.1 Gender 1.2 Student status 2.1 Parent status 2.2 Father's career 2.3 Father's income 2.4 Mother's career 2.5 Mother's income 3.1 Admission year 4.1 Course 4.1 GPA enrolled	Multiclass: low, medium, or good.	Classification	Various fields
[56]	1445	5	1. Personal information 2. Preuniversity features	1.1 Entry age 2.1 WAEC score 2.2 JAMB score 2.3 CUSAS score	Multiclass: first class, second class upper, second class lower, or third class.	Classification	Engineering
[57]	BDC:781 ADC:519	17	1. Preentry information 2. Cognitive variables 3. Noncognitive variables	1.1 Age 1.2 Gender 1.3 prior education 2.1 Math knowledge 2.2 Chemistry knowledge	—	Regression	Science
[58]	ADC:300 Bdc: 708	9	1. University features 2. Learning pattern 3. Behavioral features	1.1 Second semester's CGPA	Binary class: at-risk/not at-risk	Classification	Social sciences
[16]	161	61	1. Personal information 2. Social features 3. University features 4. Health 5. Study skills features 6. Motivation 7. Relationship 8. Money management 9. Career planning	2.1 Father's work 3.1 GPA 3.2 Number of credits 4.1 Healthy food	Binary class: pass/fail.	Classification	Computer
[59]	50	Unspecified	1. Personal information 2. University features	1.1 Application ID 1.2 Student name 1.3 Gender 2.1 Second-year GPA 2.2 Grade of entrance exam 2.3 Type of admission	Binary class: Yes/No	Classification	Engineering
[60]	309	4	1. Personal information 2. Preuniversity features	1.1 Age 1.2 Gender 2.1 High school achievement	Unspecified	Classification	Science

Table 7: Continued.

Paper	Dataset size	No. of variables	Independent variable categories	Highly influential factors	Dependent variable	DM task	Sample's scope
[29]	1854	4	1. University features	1.1 Midterm grade 1.2 Department 1.3 Faculty	Multiclass: "< 32.5," "32.5–55," "55–77.5," or "≥ 77.5."	Classification	Art
[53]	2407	18	1. Personal information	1.1 Mother's job 1.2 Department 1.3 Father's job 1.4 The main source of living expenses 1.5 The admission status	Binary class: excellent/poor	Classification	Technical

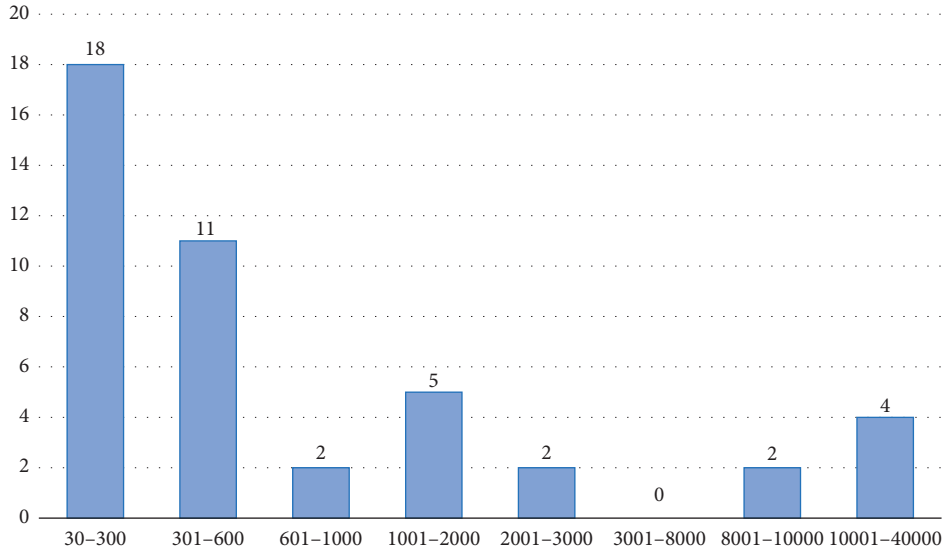

Figure 8: No. of student samples in the experimental datasets.

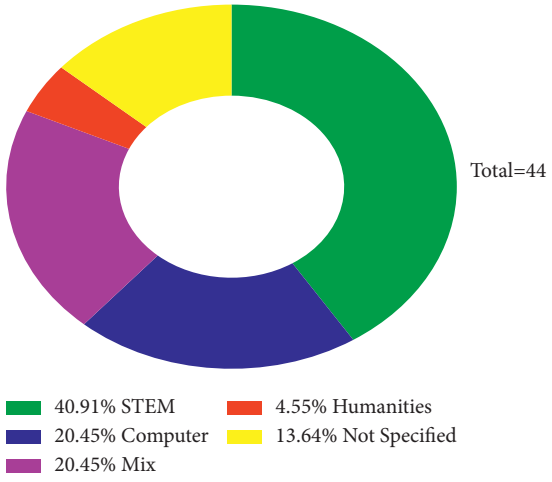

Figure 9: Student samples scope in the experimental dataset.

first-year's CGPA to predict the student who may be at risk of being drop out, indicating that knowing the student's first-year CGPA at an early stage has a significant impact on reducing student dropouts and increasing student retention. On the other hand, many researchers like in Refs. [5, 58] have demonstrated the possibility of using the first-year's CGPA to predict the student's success in the coming years.

As shown in section 5, most of the previous studies classified student's CGPA at graduation or predicted the student performance in a specific course, but unfortunately, there are only a minimal number of studies focusing on predicting student's level at the first academic year. So far, the existing studies on classifying student's achievement at the first academic year examined the nonacademic factors that may affect student achievement [57] or tried to evaluate admission criteria and their impact on student's performance at the first year [15, 56]. On the other hand, as per our knowledge, none of the studies determined the effect of using course assessment task grades to predict CGPA at the first academic year. Additionally, it has been noticed that most of the previous works analyzed student data from computer science or STEM (Science, Technology, Engineering, and Math) major. In contrast, a few studies have examined student data in arts and humanities major, indicating the need to investigate and analyze the student data in these academic disciplines.

7. Challenges and Opportunities

Most of the previous studies suffered from some common challenges, which can be summarized in the following points:

(i) *Dataset Related.*

The small size of the dataset is used to build the prediction model and the imbalanced dataset. Most of the previous studies resorted using the Synthetic Minority Oversampling Technique (SMOTE) to increase their dataset size. Additionally, the educational dataset suffers from missing data that needs to be addressed.

Most of the previous studies built the students' performance prediction model based on students' sample from one university/program major.

Most of the previous studies focussed on analyzing student data from computer science or STEM majors. In contrast, a few studies have examined student data in arts and humanities majors.

(ii) *Data Availability.*

Accessing educational data is not an easy task, as it requires numerous approvals to obtain the required data to build the prediction model.

Some factors are not available in the educational institutions' records, such as social factors, motivation factors, and behavioral factors, where it requires the use of one of the data collection methods such as an interview or questionnaire.

(iii) *Predictive Model.*

Often the predictive models are built using the default values of the algorithms' parameters rather than searching for the optimal parameters' values.

Most of the prediction models are not interpretable.

8. Conclusion

Predicting student achievement is becoming one of the most attractive research topics due to its significant impact on enhancing students' academic levels by utilizing various educational data mining methods to provide the essential support for struggling students. This study contributes to the literature in different ways where a systematic review was conducted to summarize the previous studies that aimed to predict student academic performance in higher education. Several studies have utilized data mining techniques in the educational fields with the purpose of predicting students' achievement. The previous studies have been arranged in different groups based on the targeted research objective. The main goal of this paper is to highlight the existing gap in the reviewed studies. A total of 45 studies were reviewed in this paper to identify the existing gap and answer the aforementioned questions. It was noted that most of the reviewed studies focused on studying scientific disciplines. However, the humanities disciplines received little attention, which indicates the importance of studying this major. In addition, previous studies used many techniques, but we noticed that the DT, NB, ANN, SVM, KNN, and LR were considered as the most used techniques for predicting and identifying student's level, whereas the RF and ensemble method showed the best performance and results. Many of the research studies that have been published in this field used several factors to predict the level of the student, but it turns out that the academic factors such as CGPA, assessment marks, and admission requirements are the most influencing factors in predicting student achievement, followed by the demographic factors. The main limitation of this work is that the literature review included only studies published in the past five years, and in addition, it focused on studies that predict student academic performance during traditional (face-to-face) learning and excluded studies that predict student performance during distance learning. In future work, researchers should consider studies that focus on predicting a student's academic performance during distance learning in addition to the studies that focus on predicting the students' performance using the deep learning techniques.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Vol. 4, John Wiley & Sons, Hoboken, NJ, USA, 2014.

[2] M. Singh, "Classification system via data mining algorithm: new tool to diagnose Alzheimer's disease," *Journal of the Neurological Sciences*, vol. 405, pp. 161-162, 2019.

[3] A. K. Arslan, C. Colak, and M. E. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke," *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 87–92, 2016.

[4] N. T. Southall, M. Natarajan, L. P. L. Lau et al., "The use or generation of biomedical data and existing medicines to discover and establish new treatments for patients with rare diseases-recommendations of the IRDiRC Data Mining and Repurposing Task Force," *Orphanet Journal of Rare Diseases*, vol. 14, no. 1, pp. 225–229, 2019.

[5] E. Alyahyan and D. Dusteaor, "Decision trees for very early prediction of student's achievement," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, vol. 2020, 2020.

[6] N. Alangari and R. Alturki, "Predicting students final GPA using 15 classification algorithms," *Romanian Journal of Information Science and Technology*, vol. 23, no. 3, pp. 238–249, 2020.

[7] M. Rogalewicz and R. Sika, "Methodologies of knowledge discovery from data and data mining methods in mechanical engineering," *Management and Production Engineering Review*, vol. 7, no. 4, pp. 97–108, 2016.

[8] W. Y. Chiang, "Applying data mining for online CRM marketing strategy: an empirical case of coffee shop industry in Taiwan," *British Food Journal*, vol. 120, no. 3, pp. 665–675, 2018.

[9] S. C. Huang, T. K. Wu, and N. Y. Wang, "An intelligent system for business data mining," *Glob. Bus. Financ. Rev.,* vol. 22, no. 2, pp. 1–7, 2017.

[10] M. Kantardzic, *DATA MINING: Concepts , Models , Methods , and Algorithms*, John Wiley & Sons, Hoboken, New Jersey, USA, 2011.

[11] R. B. Cristóbal Romero, S. Ventura, and M. Pechenizkiy, *Handbook of educational data mining*, CRC Press, London, UK, 2010.

[12] O. Scheuer and B. M. McLaren, "Educational Data Mining," *Encyclopedia of the Sciences of Learning*, vol. 13, 2012.

[13] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: a review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019.

[14] C. Romero and S. Ventura, "Educational data mining and learning analytics: an updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. 1–21, 2020.

[15] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[16] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, p. 26, 2018.

[17] R. Vidhya and G. Vadivu, "Retracted article: towards developing an ensemble based two-level student classification model (ESCM) using advanced learning patterns and analytics," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7095–7105, 2020.

[18] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Waltham, MA, USA, 2012.

[19] P. Rojanavasu, "Educational data analytics using association rule mining and classification," in *Proceedings of the 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, Nan, Thailand, 2019.

[20] M. Kubat, *An Introduction to Machine Learning*, Springer International Publishing, Berlin, Germany, 2015.

[21] G. A. S. Santos, K. T. Belloze, L. Tarrataca, D. B. Haddad, A. L. Bordignon, and D. N. Brandao, "EvolveDTree: analyzing student dropout in Universities," in *Proceedings of the2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niterói, Brazil, 2020.

[22] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization," *Computers & Education: Artificial Intelligence*, vol. 3, Article ID 100066, 2022.

[23] F. A. Bello, J. Kohler, K. Hinrechsen, V. Araya, L. Hidalgo, and J. L. Jara, "Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout," in *Proceedings of the 2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, Coquimbo, Chile, 2020.

[24] A. Sarra, L. Fontanella, and S. Di Zio, "Identifying students at risk of academic failure within the educational data mining framework," *Social Indicators Research*, vol. 146, no. 1–2, pp. 41–60, 2019.

[25] P. Kamal and S. Ahuja, "An ensemble-based model for prediction of academic performance of students in undergrad professional course," *Journal of Engineering, Design and Technology*, vol. 17, no. 4, pp. 769–781, 2019.

[26] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Research in Higher Education*, vol. 60, no. 7, pp. 1048–1064, 2019.

[27] B. Kiss, M. Nagy, R. Molontay, and B. Csabay, "Predicting dropout using high school and first-semester academic achievement measures," in *Proceedings of the 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pp. 383–389, Starý Smokovec, Slovakia, 2019.

[28] A. U. Khasanah, "A comparative study to predict student's performance using educational data mining techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 215, Article ID 012036, 2017.

[29] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[30] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and shapley additive explanations," *IEEE Access*, vol. 9, pp. 152688–152703, 2021.

[31] M. Adnan, "Predicting At-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[32] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemblemodel selection for multi-class educational data mining," *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, 2020.

[33] A. Salah Hashim, W. Akeel Awadh, and A. Khalaf Hamoud, "Student performance prediction model based on supervised machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, Article ID 032019, 2020.

[34] A. Kumar Veerasamy, D. D'Souza, M. V. Apiola, M. J. Laakso, and T. Salakoski, "Using early assessment performance as early warning signs to identify at-risk students in programming courses," in *Proceedings of the 2020 IEEE Frontiers in Education Conference (FIE)*, Uppsala, Sweden, 2020.

[35] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of Medical Systems*, vol. 43, pp. 162–6, 2019.

[36] B. Al Breiki, N. Zaki, and E. A. Mohamed, "Using educational data mining techniques to predict student performance," in *Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, Ras Al Khaimah, UAE, 2019.

[37] W. Liu, "An improved back-propagation neural network for the prediction of college students' English performance," *Int. J. Emerg. Technol. Learn.,* vol. 14, no. 16, pp. 130–142, 2019.

[38] C. Jalota and R. Agrawal, "Analysis of educational data mining using classification," in *Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 243–247, Faridabad, India, 2019.

[39] S. Tuaha, I. F. Siddiqui, and Q. Ali Arain, "Analyzing students' academic performance through educational data mining," *3C Tecnol. innovación Apl. a la pyme*, vol. 8, pp. 402–421, 2019.

[40] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 27–18, 2019.

[41] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, 2019.

[42] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Information and Learning Sciences*, vol. 120, no. 7/8, pp. 451–467, 2019.

[43] M. Tsiakmaki, G. Kostopoulos, G. Koutsonikos, C. Pierrakeas, S. Kotsiantis, and O. Ragos, "Predicting university students' grades based on previous academic achievements," in *Proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 9–14, Zakynthos, Greece, 2018.

[44] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, J. Rego, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247–256, 2017.

[45] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," 2017, https://arxiv.org/abs/1708.08744.

[46] T. Devasia, T. P. Vinushree, and V. Hegde, "Prediction of students performance using educational data mining," in *Proceedings of the 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp. 91–95, Ernakulam, India, 2016.

[47] J. Teguh Santoso, N. L. W. Sri Rahayu Ginantra, M. Arifin, R. Riinawati, D. Sudrajat, and R. Rahim, "Comparison of classification data mining C4.5 and Naïve Bayes algorithms of EDM dataset," *TEM Journal*, vol. 10, no. 4, pp. 1738–1744, 2021.

[48] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, Article ID e01250, 2019.

[49] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich, and T. Thamrongthanyawong, "Comparative study of prediction models for final GPA score: a case study of rajabhat rajanagarindra university," in *Proceedings of the 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 92–97, Singapore, 2018.

[50] Y. Altujjar, W. Altamimi, I. Al-Turaiki, and M. Al-Razgan, "Predicting critical courses affecting students performance: a case study," *Procedia Computer Science*, vol. 82, pp. 65–71, 2016.

[51] S. M. Hassan and M. S. Al-Razgan, "Pre-university exams effect on students GPA: a case study in it department," *Procedia Computer Science*, vol. 82, pp. 127–131, 2016.

[52] P. D. Gil, S. da Cruz Martins, S. Moro, and J. M. Costa, "A data-driven approach to predict first-year students' academic success in higher education institutions," *Education and Information Technologies*, vol. 26, no. 2, pp. 2165–2190, 2021.

[53] L. Chen, T.-T. Huynh-Cam, and H. Le, "Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance," *Algorithms*, vol. 14, 2021.

[54] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, Article ID 106903, 2021.

[55] W. Nuankaew and J. Thongkam, "Improving student academic performance prediction models using feature selection," in *Proceedings of the 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 392–395, Phuket, Thailand, 2020.

[56] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," *Education and Information Technologies*, vol. 24, no. 2, pp. 1527–1543, 2019.

[57] J. Willems, L. Coertjens, B. Tambuyzer, and V. Donche, "Identifying science students at risk in the first year of higher education: the incremental value of non-cognitive variables in predicting early academic achievement," *European Journal of Psychology of Education*, vol. 34, no. 4, pp. 847–872, 2019.

[58] Z. Ahmad and E. Shahzadi, "Prediction of students' academic performance using artificial neural network," *Bulletin of Education and Research*, vol. 40, no. 3, pp. 157–164, 2018.

[59] R. Patil, S. Salunke, M. Kalbhor, and R. Lomte, "Prediction system for student performance using data mining classification," in *Proceedings of the 2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–4, Pune, India, 2018.

[60] T. Doleck, D. J. Lemay, R. B. Basnet, and P. Bazelais, "Predictive analytics in education: a comparison of deep learning frameworks," *Education and Information Technologies*, vol. 25, no. 3, pp. 1951–1963, 2020.