DOI: 10.51153/kjcis.v7i1.212

Predicting Student Performance Using Educational Data Mining: A Review

Veena Kumari[a],Areej Fatemah Meghji[a*], Rohma Qadir[a],UroojOad[a]
[a]Department of Software Engineering, Mehran University of Engineering and Technology, Pakistan
vina.mehrani@gmail.com, areej.fatemah@faculty.muet.edu.pk, zainabqadir97@gmail.com, uroojgianchand@gmail.com

Farhan Bashir Shaikh [b]
[b]Faculty of Information and Communication Technology, UniversitiTunku Abdul Rahman, Malaysia
farhanb@utar.edu.my

*Corresponding Author: Areej Fatemah Meghji areej.fatemah@faculty.muet.edu.pk

Abstract

Educational Data Mining (EDM) strategies facilitate the efficient and in-depth analysis of student data. EDM provides useful insights into comprehending student learning patterns and identifying factors that influence academic success. This review aims to evaluate the efficacy of classification algorithms popularly explored in EDM for predicting student performance and identifying common trends in existing EDM research. The review follows a systematic approach, relevant research articles have been cited following an inclusion and exclusion criteria to ensure the selection of studies that specifically address the use of EDM techniques for predicting student academic achievement. According to the review findings, most researchers have utilized the features of cumulative grade point average, internal and external assessment, and demographic information to predict student performance. The most common techniques in EDM for predicting students' performance are Naïve Bayes and Decision Trees. The review also focuses on the potential for bias, key examination of challenges, and possible future directions in the field. In the context of student performance prediction, ethical considerations regarding privacy, data handling, and the interpretation of results are also identified.

Keywords: classification, educational data mining, decision tree, academic achievement.

1. Introduction

Educational institutions have always collected vast amounts of student data [1]. These institutes are constantly trying to find the most effective ways to store, process, and make use of this data to better understand how students learn [2]. Getting insights from this data can greatly aid educational institutes in shaping pedagogical policies and devise strategies for fostering a student-centric environment of learning [3]. Educational data mining (EDM) is an emerging field of research tasked with analyzing, developing, and employing computational techniques to uncover hidden patterns in large learning or educational data sets, which can be difficult to evaluate due to their size. This developing branch of knowledge and data mining explores real-time educational databases ranging from student admissions, registration, and examination, to course management systems like Moodle and Blackboard. The main goal of research using EDM is to analyze student academic achievement at different levels of their educational journey such as school, college, and university [2]. One key area of research within EDM is predicting student performance at an early stage during their education [3].

Classification is a popular, supervised data mining method that relies on learning from historic, labeled data; it finds patterns in the data to generate a model. The generated model is then used to make predictions for new instances of data by classifying them into pre-defined labels or classes based on the discovered patterns within the data [1]. Decision trees, Bayesian networks, neural networks, and K-nearest neighbor (KNN) are some of the often-employed approaches of classification [3].

Classification has been used to target the prediction of several facets of education including the prediction of courses that have a significant impact on final degree level performance [1], or factors that cause a fluctuation in the grade point average of a student [4]. An interesting area of research is the analysis of student data to categorize students into classes of learners [2]. These techniques concentrate on analyzing student educational data, which represents their academic performance, and developing clear rules to aid students in their future academic performance [5]. Universities have been analyzing educational data such as enrollment data, student academic data, student learning data, feedback data, and many other forms of data, using varied classification approaches to provide a university with the necessary knowledge to more effectively plan for student enrolment, avoid student dropouts, detection of students at risk of failure, and resource allocation with a precise approximation [6], [7]. The outcome of EDM is intended tohelp pedagogical decision making. An insight into the factors affecting student performance can help educational institutes mitigate those factors to boost student performance, which will directly boost the overall performance of an educational institution. This review presents a picture of the use of EDM towards the prediction of student academic performance. The goals of the systematic review are to:

1. Ascertain the features or attributes that are important for analyzing a student's performance.
2. Discover the most preferred EDM methods for predicting student performance.
3. Identify the challenges facing EDM research.

The review is structured as follows: Section 2 focuses on the research methodology including the research questions to be addressed, a search strategy to extract relevant previous studies, and a review conducting process. Section 3 provides the proposed research questions in parallel with the results analysis of the review, followed by a discussion in Section 4. The conclusion of the review is presented in Section 5.

2. Research Methodology

Figure 1 represents the research methodology followed to conduct this literature review.

1. The plan phase of the review included clearly identifying the scope or boundary of the review,

outlining the research questions that the review would address, and creating a review protocol that would outline the fundamental review procedure. This step also included establishing the sources for the collection of the studies, setting a criterion for inclusion/exclusion, and defining the process through which this criterion would be applied.

2. The second phase comprised the actual conduct of the review. This included extracting information based on the research questions defined in step-1; the gathered information was also assessed and tabulated.
3. The reporting phase involved presenting the findings of the review. The analysis was written up during this phase, ensuring the review's credibility and usefulness.

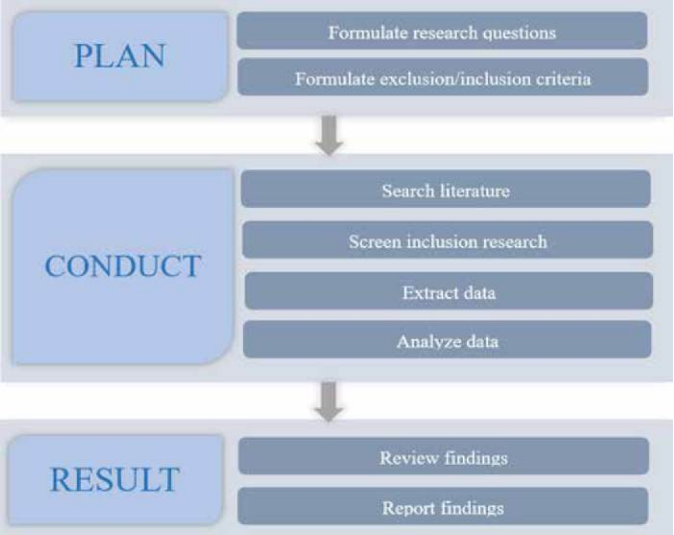
PLAN
Formulate research questions
Formulate exclusion/inclusion criteria
CONDUCT
Search literature
Screen inclusion research
Extract data
Analyze data
RESULT
Review findings
Report findings

Figure1.Research Methodology

2.1 Research Questions

The right research questions are crucial to understanding the current study on performance prediction for students. The criteria for the study have been outlined in Table 1 following the Kitchenham et al. approach of identifying the Population, Intervention, Outcome, and Context [8].

Table 1. Research question criteria

Criterion	Description
Population	University students
Context	Empirical studies including preliminary studies, case studies, comparative study
Intervention	EDM methods for student performance prediction
Outcome	Classifier accuracy, most used classification algorithms, parameters used for performance prediction

Three research questions were proposed for this review:

RQ1. What are the most essential factors explored when forecasting student performance?

RQ2. Which classification techniques or algorithms are commonly used in EDM research with the intent of predicting student performance?

RQ3. What are the limitations or challenges in using EDM for predicting student performance?

2.2 Search Strategy

It is important to have a well-thought-out search strategy for a systematic literature review. This ensures finding the most relevant work from the massive pool of research. A thorough search was made for research publications that address the suggested research questions. The search was carried out in four knowledge sources namely IEEE Xplore, ACM Digital Library, Science Direct, and Google Scholar to extract relevant research papers. The text strings used to carryout the research are mentioned in Table 2.

Table 2. Search strings executed on knowledge sources

Knowledge Source	Search Strings
IEEE Xplore, ACM Digital Library, ScienceDirect, Google Scholar	“Student performance”, “Educational data mining methods”, “Student performance prediction”, “Educational data mining techniques”, “EDM student performance”, “EDM classification”

Items extracted during the search include preliminary empirical studies such as journal articles, case studies, comparative studies, workshop papers, and conference papers. Based on the search strings outlined in Table 2, a total of 124 articles were shortlisted initially as depicted in Table 3.

Table 3. Quantity of papers acquired from knowledge sources

Knowledge Source	Number of Research Articles
IEEE	28
ACM	07
Science Direct	16
Google Scholar	73

After formulating the research questions, selecting knowledge bases, and separating the research papers based on the search strings, an important step was to execute the inclusion/exclusion criterion. The formulation of a research selection process is an important step of the review as it helps set the scope and boundary of the review. It also helps guide the review process by allowing the researcher to consider or disregard a paper based on a set of pre-established criterion. The inclusion and exclusion criterion for the conducted review have been outlined in

Table 4.

Table 4. Exclusion and inclusion criteria

S.no	Inclusion Criteria	Exclusion Criteria
1	Nature of the research should be empirical	Non-empirical studies
2	Papers must be fully accessible	Papers that are not accessible
3	Papers must be published between 2014 to 2024	Papers published before 2014
4	The paper must include EDM techniques and findings	Paper that addresses EDM techniques in general
5	Papers that address the research questions	Papers not addressing the research questions

Figure 2 outlines the detailed paper selection process and the articles dropped at each stage during the conduct of this review. To ensure the quality of the research, the article inclusion criteria have been firmly followed. After the initial selection, the studies were foremost screened based on the title of the research. After the first title screening, 42 papers were eliminated. The abstracts of the selected papers were then scrutinized. All the papers were thoroughly examined to exclude extraneous material; the papers that did not address the questions put forth in this review or lacked original findings were disregarded at this stage, leaving 28 articles for this study to analyze.

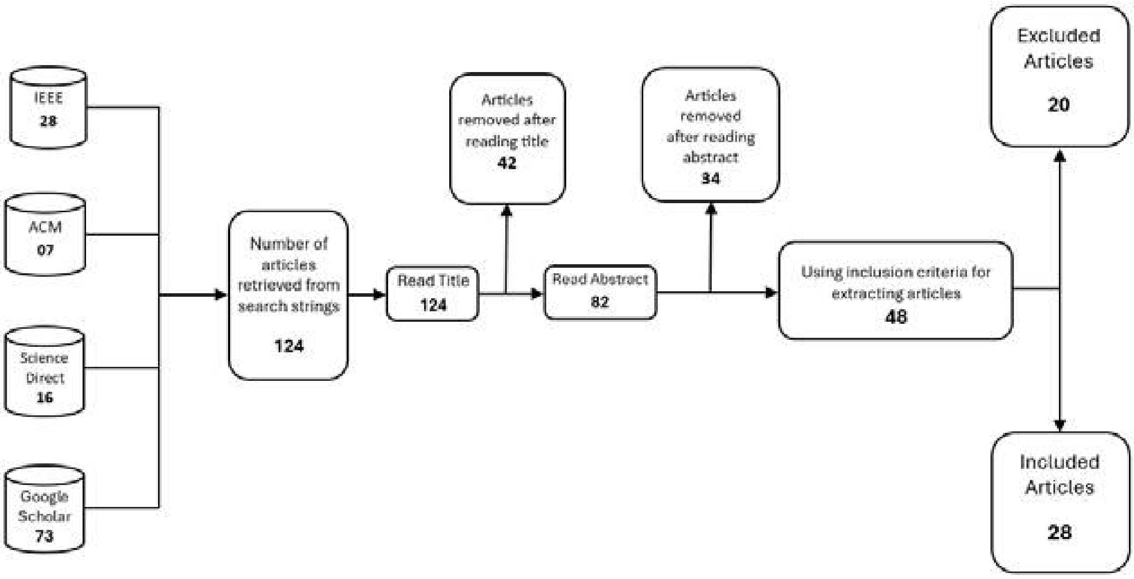
IEEE
28
ACM
07
Science Direct
16
Google Scholar
73
Number of articles retrieved from search strings
124
Read Title
124
Articles removed after reading title
42
Read Abstract
82
Articles removed after reading abstract
34
Using inclusion criteria for extracting articles
48
Excluded Articles
20
Included Articles
28

Figure 2. Paper Selection Process

Table 5 presents the summary of the extracted papers. We can observe that most of the covered literature has been acquired from journal publications, 20 out of the 28 covered articles in the review have been published in reputed journals, whereas 8 articles have been published in conference proceedings.The overall coverage of the knowledge sources has been depicted in Fig. 3.

Table 5. Summary of extracted articles

Knowledge Source	Number of Journal papers	Number of Conference papers	Total Number of Articles
IEEE	4	-	4
ACM	2	1	3
Science Direct	5	3	8
Google Scholar	9	4	13

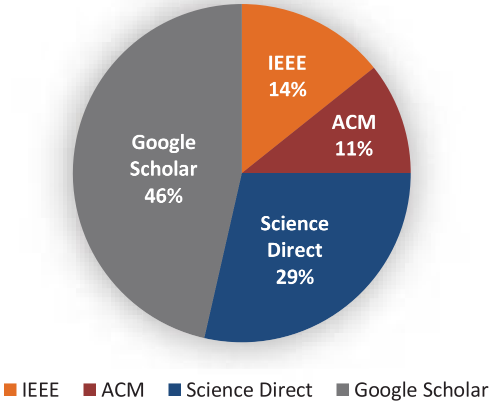

Figure 3. Coverage of knowledge sources

The literature review was performed on research papers published from 2014 to 2024 as shown in Figure 4. From the temporal view of the papers, we can see an increase in research focusing on the use of EDM towards student performance prediction in 2016 and 2021, closely followed by 2023.

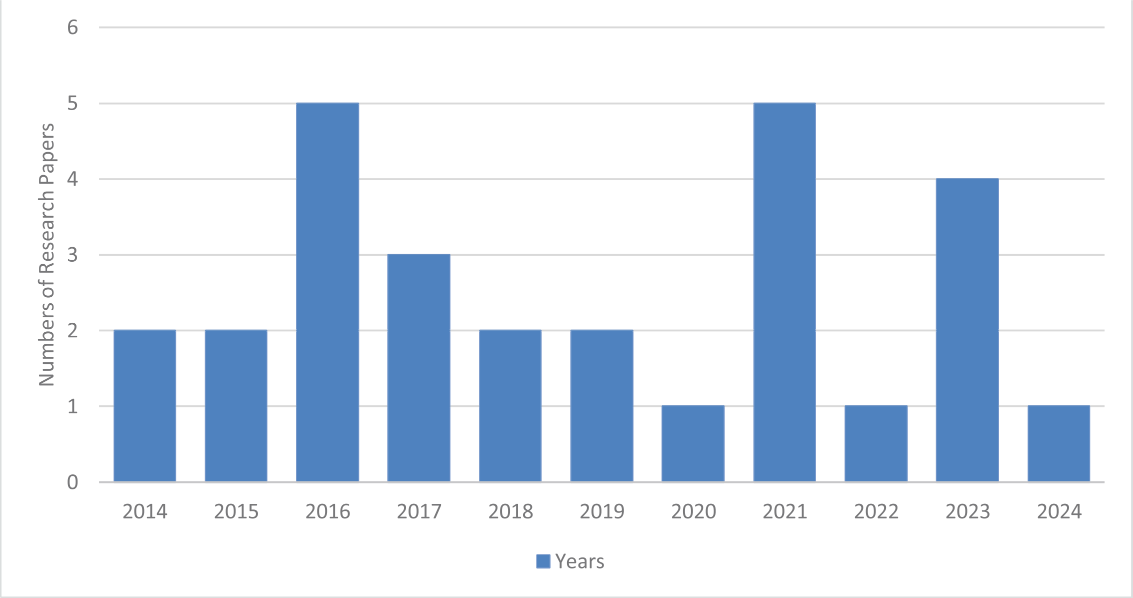

Figure 4. Temporal view of publications

3. Results &Analysis

This section of the literature review presents the findings of the conducted review. The findings have been categorized based on the research questions proposed in section 2.1. We have also attempted to provide a tabular summary of the most used EDM techniques, the features most commonly utilized in the research, and the experimental accuracy achieved using the various EDM classification approaches.

3.1 RQ1: What are the most essential factors explored when forecasting student performance?

Various factors(features/attributes) have been explored in the reviewed research for predicting student performance with the cumulative grade point average (CGPA) being the most frequently used [9], [10], [11], [12], [13], [14], [15].A primary reason why most researchers use CGPA could be because it has a concrete value that can be used to address future academic mobility; CGPA shows the exact output variable as compared to other elements, therefore, it is the most influential in deciding the survival of students in their studies, and whether they can finish their course. Apart from CGPA, another important feature considered for student performance prediction was internal assessment. Internal assessment estimates student's academic performance through various means such as quizzes, class participation, attendance, assignments, and term grades. According to some studies, the timely analysis of student performance in these activities plays a major role in student overall performance in the course [2], [10], [12], [14], [15], [16]. Also, analysis of these factors can help generate early warning systems and intervention channels which can help the instructor monitor and improve the performance of students in their course [17], [18]. The influence of the previous semester (external assessment) on the outcome of the current semester is a dynamic interplay of academic performance, and knowledge retention [12], [13]. Prequalification ensures that applicants to the university have met certain academic requirements as it suggests that they possess the fundamental knowledge and skills necessary to succeed in courses at the university level; this can have a positive effect on student performance [13].

Apart from features pertaining to academics or learning, studies have also taken into consideration factors such as demographic information including age, nationality, and gender, which may affect student performance [12], [15]. The demographics of age and gender are frequently explored for performance prediction as they are viewed as inner elements of an individual, which are easy to characterize and quantify. Studies have contradictory findings on the significance of gender in student performance prediction. However, it has been seen that female, and male students follow different learning patterns [19].

Some studies also provide an analysis that parents' educational background, job, family structure, and family income influence the performance of a student. However, this must be kept in mind that all these comprehensive factors contribute to qualitative analysis of the student's performance [9]. Research has focused on extracting and examining factors from both face-to-face as well as online forms of learning. Some researchers have also obtained additional factors using surveys or questionnaires. However, most of the research has focused on the use of features obtained from physical or face-to-face forms of education. The most common factors explored in the reviewed papers have been categorized in Table 6.

Table 6. Factors explored when forecasting student performance

Factor	Description	Reference
Internal Assessment	Student performance evaluation through quizzes, class participation, attendance, and term results, shaping overall aggregate.	[3], [6], [10], [11], [12], [14], [15], [16], [20], [21], [23], [24], [26], [29]
Prequalification	Ensuring the students meet academic pre-requirements before university enrollment, indicating ease towards university-level courses.	[6], [13], [15], [16], [24], [25], [26]
External Assessment	Performance of the previous semester or academic year influencing performance in the current semester.	[12], [13], [16], [20], [26], [27]
CGPA	A grade point average value indicating overall score summarizing the academic performance of all courses of the semester, year, or academic duration.	[1], [9], [10], [11], [12], [13], [14], [15], [20], [21], [23], [26], [28]
Demographic Information	Considering age, nationality, ethnicity, and gender of students for characterization, identifying potential learning patterns.	[12], [15], [18], [20], [21], [24], [25]
Parental Background	Educational background, family structure, and family income creates an impact on student performance, contributing to a qualitative academic understanding.	[9], [12], [25], [27]

Aptitude Score	A quantitative metric to identify student’s potential for success in upcoming academia, mostly used in the institute’s entrance testing and assessments.	[1], [4]
E-Learning	Data obtained through a learning management or e-learning system	[4], [9], [16], [22]

3.2 RQ2: Which classification techniques or algorithms for predicting student performance are commonly used in EDM research?

To better understand student learning patterns and design effective pedagogical policies, EDM has been utilized to predict various facets of student academic performance. Regression models have been used to estimate continuous values such as final percentage [7], whereas classification, a supervised machine learning technique, is often considered the better fit for predicting student academic success in terms of performance. Classification models have been used to predict student performance outcomes in a course (pass/fail) [2], and even in a degree program (pass/fail or achieved grade) [6]. Common EDM algorithms are decision trees that create tree-like structures for easy interpretation [3]; Random Forests, an ensemble machine learning approach that combines multiple decision trees to reduce overfitting [2], [6]; Naive Bayes, which works on the probability of features belonging to each class [10]; KNN, is a non-parametric and instance-based learning algorithm[20]; and Multi-layer Perceptron (MLP), a powerful architecture for artificial neural networks that can deal with complex data relationships [16]. These algorithms facilitate the prediction of student academic results, help identify learning patterns, and classify student’s performance based on student data.

3.2.1 Naïve Bayes

Naive Bayes is a probabilistic algorithm used for classification. The algorithm determines the probability of a hypothesis given the evidence and selects the class with the maximum probability to classify input data. In the reviewed literature, sixteen studies have utilized the Naïve Bayes techniques to predict student performance. Analysis of their findings has been provided in Table 7.

Table 7. Review Summary - Naïve Bayes

Classifier	Reference	Elements/Features	Accuracy
Naïve Bayes	[6]	Internal assessment, school and college results	83.65%
	[9]	CGPA, parental survey, absent days, resources used, participation in discussion	76.05%
	[10]	CGPA	73.60%
	[12]	CGPA, internal assessment, external, student demographic, parents educational background	75.00%
	[13]	CGPA, pre-qualification, previous semester	73.50%
	[14]	CGPA, internal assessment, lab marks, attendance marks	61.00%
	[15]	CGPA, student demographic, pre-qualification	86.20%

	[16]	Student demographic, pre-qualification, Institutional assessment	67.60%
	[19]	GPA, gender, ethnicity, financial status	67.00%
	[20]	CGPA, assessments, student demographic	73.00%
	[21]	CGPA, internal assessment, student demographic, parents' educational background	86.00%
	[23]	Internal assessment	71.30%
	[24]	Internal assessment, enrolment data, school and college results, demographics	79.72%
	[25]	Student demographics, parents' educational background, qualification, travel time, scholarship	69.70%
	[27]	External assessment, father qualification, social features	91.79%
	[29]	Student performance data	74.00%

3.2.2 K-Nearest neighbors

Among all data mining techniques, KNN is a simple, non-parametric, and instance-based learning algorithm preferred to perform classification and regression. Although this classifier is deemed lazy as it does not readily generate a classification model but rather scans the incoming instances at runtime to make classification decisions, it is still widely adopted as it does not make any prior assumptions regarding the data it has to classify. KNN has been used in eleven of the reviewed studies to analyze and predict student performance as depicted in Table 8.

Table 8. Review Summary - KNN

Classifier	Reference	Elements/Features	Accuracy
KNN	[3]	Internal assessment marks, final exam marks	83.00%
	[4]	Spatial reasoning test, critical thinking, online course data	77.80%
	[5]	Semester grades, gender, age, parents’ education	92.25%
	[6]	Internal assessment, school and college results	74.00%
	[9]	CGPA, parental survey, absent days, resources used, participation in discussion	76.70%
	[12]	CGPA, internal assessment, external, student demographic, parents educational background	83.00%
	[16]	Student demographic, pre-qualification, Institutional assessment	82.00%
	[20]	CGPA, assessments, student demographic	74.00%
	[23]	Internal assessment	69.90%
	[24]	Internal assessment, enrolment data, school and college results, demographics	76.28%
	[27]	External assessment, father qualification, social features	88.86%

DOI: 10.51153/kjcis.v7i1.212

3.2.3 Decision tree

The decision tree is a non-linear, predictive modeling approach used in machine learning and data mining. The algorithm based on this approach involves recursively partitioning the data hinged on the most informative attributes to create subsets with homogeneous target values [13]. To predict a student's performance, eighteen of the reviewed papers have employed the decision tree method. An analysis of their findings has been listed in Table 9.

Table 9. Review Summary - Decision Tree

Classifier	Reference	Elements/Features	Accuracy
Decision Tree	[1]	Test scores, general aptitude test score, GPA, school GPA, gender.	87.17%
	[4]	Spatial reasoning test, critical thinking, online course data	87.60%
	[5]	Semester grades, gender, age, parents' education	94.55%
	[6]	Internal assessment, school and college results	71.15%
	[9]	CGPA, parental survey, absent days, resources used, participation in discussion	95.20%
	[10]	CGPA	75.90%
	[11]	CGPA, internal assessment	55.50%
	[12]	CGPA, internal assessment, external, student demographic, parents educational background	88.00%
	[14]	CGPA, internal assessment	56.10%
	[16]	Student demographic, pre-qualification, Institutional assessment	72.30%
	[18]	Student Demographic	70.80%
	[19]	GPA, gender, ethnicity, financial status	68.80%
	[23]	Internal assessment	74.60%
	[24]	Internal assessment, enrolment data, school and college results, demographics	80.75%
	[25]	Student demographics, parents' educational background, qualification, travel time, scholarship	73.20%
	[26]	Internal assessment, school and college score, education attainment test, CGPA	80.00%
	[27]	External assessment, father qualification, social features	92.96%
	[28]	CGPA, social features, external assessment	67.37%

3.2.4 Multi-layer perceptron

The classification model constructed by MLP takes the form of an artificial neural network. Such models have demonstrated to be effective in a wide range of tasks, including classification, regression, and other pattern recognition. It is a popular neural networks architecture used to estimate student's academic performance [21]. An analysis of the findings of the reviewed papers based on MLP has been provided in Table 10.

Table 10. Review Summary - MLP

Classifier	Reference	Elements/Features	Accuracy
MLP	[6]	Internal assessment, school and college results	71.15%
	[9]	CGPA, parental survey, absent days, resources used, participation in discussion	94.50%
	[15]	CGPA, Student demographic, pre-qualification	81.30%
	[16]	Student demographic, pre-qualification, Institutional assessment	78.70%
	[20]	CGPA, assessments, student demographic	50.00%
	[21]	CGPA, internal assessment, student demographic, parents' educational background	82.70%
	[23]	Internal assessment	74.60%
	[25]	Student demographics, parents' educational background, qualification, travel time, scholarship	69.30%

Figure 5 depicts the frequency of the classifiers' explored in the reviewed literature. It can be observed that decision tree along with Naïve Bayes are the more preferred classifiers explored in the reviewed literature.

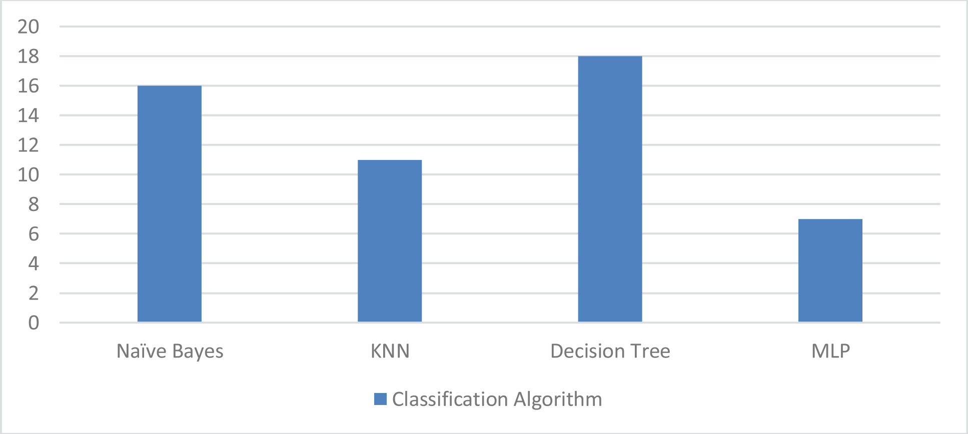

Figure 5. Frequency of classifiers' explored

3.3 RQ3: What are the limitations or challenges in using EDM for predicting student performance?

While EDM can be a useful tool for instructors and institutions for predicting student performance, it does have some limitations and difficulties. Some of the highlighted challenges in the reviewed literature include:

1. Quality and availability of the data: The accuracy of forecasts vigorously depends upon the quality and amount of the information accessible. On the off chance that the information is missing, inadequate, mistaken, or one-sided, it can prompt temperamental prediction and frustrate the adequacy of the EDM model [22].

2. Privacy and ethical concerns: Educational data frequently contains delicate information about students. Analyzing and sharing students' information should be carried out delicately, while strictly complying with severe security guidelines and guaranteeing that no singular student information is compromised [7].

3. Interpretability: Some EDM methods, particularly complex AI models, could need interpretability. Understanding why a model makes a particular prediction can be difficult, and at times impossible. This can be worrisome in educational analysis where straightforwardness and an understanding of the classification logic are fundamental for pursuing informed choices [24].

4. Logical Elements: Predicting students' execution is affected by different relevant elements, for example, financial status, parental expectations, gender comparisons, family foundation, and outer impacts. EDM models could battle to catch the intricacy of these elements [23].

5. Data Preprocessing: Frequently, extensive data preprocessing and feature engineering are required when preparing educational data for analysis. Time-consuming and requiring domain expertise are the requirements for cleaning, transforming, and selecting relevant features [2].

4. Discussion

This meta-analysis relies on the efficiency of the EDM techniques (measured using prediction accuracy) as well as the critically important factors that could affect students' performance. Taking into consideration the accuracy achieved by the models generated in the reviewed papers, the overall highest accuracy achieved by the model generated by Naïve Bayes was 91.79% [27], the model based on the decision tree recorded the highest accuracy of 95.20% [9], for the model based on MLP, the accuracy was 94.50% [9], and the model based on KNN exhibited the highest accuracy of 92.25% [5].

As observed, different researchers have investigated various blends of educational elements or features to accomplish better accuracy of their generated models. For instance, when working with the features of CGPA, internal assessment, external scores, student demographic information, and parents' educational backgrounds, the model based on Naïve Bayes achieved an accuracy of 75%. Working with the same data and features, the model based on KNN achieved an accuracy of 83%, whereas the decision tree outperformed working on the same data with an accuracy of 88% [12]. Another study used student demographic data, pre-qualification information, and institutional assessments to predict student performance using the approaches of MLP, KNN, Naïve Bayes, and decision tree; with the model based on KNN outperforming with an accuracy of 82%[16].When working with the features of internal assessment focusing on marks obtained in subjects being studied, school and college results, the model based on the decision tree and MLP achieved an accuracy of 71.15%, while the model based on KNN achieved an accuracy of 74%, and the model based on Naïve Bayes had an accuracy of 83.65% [6]. The difference in classifier performance signifies that there is no 'one fit all' approach when it comes to using classification algorithms. Classifier performance greatly depends on the size and quality of the data as well as the features being considered [6], [9], [24].

Based on the reviewed literature, and the uncovered challenges and limitations, some areas that need attention, possible solutions to the identified challenges, and potential future research directions can be categorized as:

Identification of Factors and Interpretability: The main motivation behind the field of EDM is the discovery of patterns and facets of student learning that make a student perform well or poorly in a course, semester, or across a degree program and then use these discovered patterns to fashion, among others, better learning infrastructures, early warning systems, and intervention mechanisms. In order to accomplish this, it is essential to not just focus on using varied classification algorithms to predict student academic achievement but rather take the research process a step further and try to identify the features that most influence student performance. The uncovered factors should then be communicated to the administrators and instructors so that, where required, interventions can be timely planned and pedagogical policies be reshaped to add quality to the system of education. While several of the reviewed papers have undertaken this exploration and provided conclusions based on factors that help a student excel or hinder academic achievement [6], [9], [10], [24], many papers have solely focused on a comparative analysis of classification performance, limiting their research to simply finding the best performing classifier [1], [3], [4], [13], [14], [15], [17], [18], [19], [27], [29]. Also, although Naïve Bayes has been utilized in the majority of studies, in most studies where both the approaches of Naïve Bayes and decision trees have been utilized, we can observe that the model based on the decision tree has almost consistently outperformed [9], [10], [12], [16], [19], [23], [24], [25], [27]. A benefit of using the decision tree approach is the resulting interpretability of the model [6], [24]. Due to the generated tree-based model, educators can readily understand the prediction logic and can use the logic to create early warning and intervention systems.

Feature Selection: Predictive models such as decision trees, Naïve Bayes, KNN, and MLP offer a promising roadmap for predicting student performance and improving results. Nonetheless, they are not without their limitations and difficulties. The accuracy and dependability of these models vigorously rely upon the nature of the information, making information quality a significant concern [6]. Researchers analyzing, exploring, and experimenting on educational data need to understand the significant impact of low quality, imbalanced, small datasets on their final results. EDM is a data driven field, and as such low quality data will always result in low quality results. Additionally, choosing the right arrangement of elements is fundamental, as superfluous, or inadequately picked highlights can obstruct accuracy. Few of the reviewed studies explored feature selection approaches in their experiments [1], [6], [7], [24]. The use of feature selection approaches such as t-distributed Stochastic Neighbor Embedding [1], Attribute Evaluator Wrapper with Best First Search [7], Filter with Ranker search [7], and Correlation-based Feature Subset Selection Evaluator [24] is recommended to ensure only features that play a part in the final prediction of student performance are used in the experiments. This will ensure the integrity of the results.

Data Imbalance: Overfitting can likewise be an issue, prompting unfortunate speculation of inconspicuous information. This can again cause biased results towards the majority class. A limited number of studies have focused on balancing student data before the generation of the predictive models [7], [9], [14], [24]. The application of oversampling using the Synthetic Minority Oversampling Technique (SMOTE) has shown promising results in generating synthetic samples of the minority class, thus reducing the imbalance between the majority and minority classes. The comparative findings between the use of balanced and unbalanced data further iterates the need for researchers to conduct a proper statistical analysis of the data at their disposal and ensure that the classes are balanced before moving on to the classification process [24].

Availability, Privacy, and Ethical Concerns: As far as the availability, privacy, and ethical concerns regarding student data, with the increased interest in exploring educational data, educational institutes have implemented stringent policies regarding the access and use of student data. Researchers are allowed to collect and experiment with this data after a thorough ethical review board allows them access to the data ensuring that the data is anonymized before it is used and published [6], [10], [24]. Reputed journals do not proceed with publication unless researchers provide ethical board approvals. This ensures the privacy of student data. Moral contemplations are central while managing delicate instructive information to guarantee understudy security and information morals.

Model Optimization, Ensemble, and Hybrid Models: Hyper-parameter tuning or model optimization is an issue that has not been the focus of a majority of the covered research. The same goes for performing classification using ensemble approaches. Although promising results have been uncovered using ensemble methods, this remains a less explored area [9], [24]. Also, during the COVID pandemic, educational institutes, to a large extent, shifted to an online format of education causing a surge in the generation of e-learning and online heuristic data [5]. Although many institutes have switched back to a face-to-face form of education, they supplement the use of online sources for assignments, tests, and notes. This has also opened avenues for research focusing on additional factors. Features such as course resources accessed, participation in forums or class discussion groups, number of interactions, time of joining a class, attendance, assignment or homework submission record, and a myriad of other factors that were previously unavailable can now be used in EDM based research. Hybrid datasets containing data compiled from physical as well as online mediums have the potential to shape an in-depth understanding of student learning patterns. The increase in the number of parameters to explore along with the increase in the size of the data can also allow exploration of EDM to move from tradition machine learning to ensemble and deep learning approaches.

An interesting observation during the review was the reliance onmodels must be regularly updated to remain relevant because educational environments are always changing. Also, it needs to be underscored that educational data is hierarchical in nature; data at different levels correlates and affects the overall picture. Ultimately, an absence of space mastery and powerful mediation systems can restrict the commonsense accuracy of these models.

5. Conclusion

This paper reviewed research between 2014 and 2024 on using various classification techniques to predict students' performance. Most researchers have utilized internal assessments such as assignments, lab work, quizzes, class participation, attendance, assignments, terms grades, and CGPA as their primary elements. The most common techniques in EDM for predicting students' performance are Naïve Bayes and decision trees. Out of the reviewed papers, eighteen papers used decision tree techniques to predict student performance, evaluating the model performance on the metric of accuracy. We can thus conclude that classifiers based on the decision tree are popular and preferred classification methods in EDM research. The decision tree is followed by the Naïve Bayes classifier as the second most used approach. However, the results of the studies based on decision trees have constantly outperformed, and with the advantage of generating a readily interpretable model, the decision tree appears to be an attractive approach for student performance prediction. Further work in this area of research can focus on the utilization of feature selection and data balancing techniques and how their use influences the overall classification outcome. Although there are still concerns and difficulties in the implementation of EDM, it is a useful asset for recognizing patterns, acquiring knowledge, and giving early mediation to further understand student learning behavior. By tending these limitations and using EDM, instructors can leverage data-driven approaches to enhance the learning experience and academic performance of their students.

References

1. Alhazmi, E., &Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access, 11*, 27579-27589.
2. Abu Saa, A., Al-Emran, M., &Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. Technology, Knowledge and Learning, 24, 567-598.
3. Karthikeyan, K., &Kavipriya, P. (2017). On improving student performance prediction in education systems using enhanced data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 7*(5).
4. Çetinkaya, A., Baykan, Ö. K., & Kırgız, H. (2023). Analysis of Machine Learning Classification Approaches for Predicting Students' Programming Aptitude. *Sustainability, 15*(17), 12917.
5. Daligcon, A. G., Priyadarshini, J., &Decena, L. R. (2024). Unveiling the Best-fit Model: A Comparative Analysis of Classification Methods in Predicting Student Success. *International Journal of Information Technology, Research and Applications, 3*(1), 12-19.
6. Asif, R., Merceron, A., Ali, S. A., &Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & education, 113*, 177-194.
7. Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access, 9*, 95608-95621.
8. Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15.
9. Sahlaoui, H., Nayyar, A., Agoujil, S., &Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. *IEEE Access, 9*, 152688-152703.
10. Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462-55470.
11. Mehboob, B., Liaqat, R. M., &Saqib, N. A. (2016). Predicting student performance and risk analysis by using data mining approach. *International Journal of Computer Science and Information Security (IJCSIS), 14*(7), 69-76.
12. Ashraf, A., Anwer, S., & Khan, M. G. (2018). A Comparative study of predicting student's

performance by use of data mining techniques. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 44*(1), 122-136.

13. Shruthi, P., & Chaitra, B. P. (2016). Student performance prediction in education sector using data mining. *International Journal of Advanced Research in Computer Science and Software Engineering, 6*(3), 212–218.
14. Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics, 2*, 1-25.
15. Pavithra, A., &Dhanaraj, S.(2018). Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors.*International Journal of Scientific Research & Management Studies, 7*(5).
16. Rafique, A., Khan, M. S., Jamal, M. H., Tasadduq, M., Rustam, F., Lee, E., ... & Ashraf, I. (2021). Integrating learning analytics and collaborative learning for improving student's academic performance. *IEEE Access*, *9*, 167812-167826.
17. Durairaj, M., &Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies, 5*(4), 5987-5991.
18. Khudhur, M. E., Ahmed, M. S., & Maher, S. M. (2021). Prediction of the Academic Achievement of Pupils Using Data Mining Techniques. *Webology, 18*(2), 1355-1364.
19. Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied mathematical sciences, 9*(129), 6415-6426.
20. Zohair, A., & Mahmoud, L. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education, 16*(1), 1-18.
21. Mueen, A., Zafar, B., &Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science, 8(*11), 36.
22. Wong, J., Khalil, M., Baars, M., de Koning, B. B., &Paas, F. (2019). Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. *Computers & Education*, 140, 103595.
23. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11.
24. Meghji, A. F., Mahoto, N. A., Asiri, Y., Alshahrani, H., Sulaiman, A., & Shaikh, A. (2023). Early detection of student degree-level academic performance using educational data mining. *PeerJ Computer Science, 9*, e1294.
25. Osmanbegović, E., Suljić, M., &Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija, 16*(34), 147-158.
26. Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting critical courses affecting students performance: a case study. *Procedia Computer Science*, 82, 65-71.
27. Mousa, H., &Maghari, A. (2017). School student's performance prediction using data mining classification. *International Journal of Advanced Research in Computer and Communication Engineering, 6*(8), 136-141.
28. Singh, W., & Kaur, P. (2016). Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance. *International Journal of Advanced Research in Computer Science, 7(*6).
29. Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., &Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials today: proceedings, 80*, 3782-3785.