Significance of Education Data Mining in Student's Academic Performance Prediction and Analysis

Shujat Hussain[1], Saif Ur Rehman[1], Syed Shaheeq Raza[1], Khalid Mahmood[2], Qamar Abbas[3] and Mahwish Kundi[4]

[1]University Institute of Information Technology, Arid Agriculture University, Rawalpindi, 00666, Pakistan

[2]Institute of Computing and Information Technology, Gomal University, D.I.Khan, 29220, Pakistan

[3]Department of Computer Science, Faculty of Computing and Information Technology, International Islamic University, Islamabad, 04403, Pakistan

[4]University of Leicester, Leicester, LE1, United Kingdom

*__Correspondence__: Author: Saif Ur Rehman, Email: saif@uaar.edu.pk

Citation | Hussain. S, Rehman. S, Raza, S. S, Mahmood. K, Abbas. Q, Kundi. S, "Significance of Education Data Mining in Student's Academic Performance Prediction and Analysis", IJIST, Vol. 5 Issue.3, pp. 215-231, Sep 2023.

Received | Aug 25, 2023**; Revised |** Sep 17, 2023**; Accepted |** Sep 18, 2023; **Published |** Sep 20, 2023.

Data Mining (DM) is relevant to extract the hidden patterns from the voluminous amount of the data. Applying DM, in education is an evolving interdisciplinary research domain, which is also called as educational data mining (EDM). At present, student data about their academics is available to identify important hidden trends to be explored for enhancing student academic performance. In higher education, forecasting student success is essential for helping with course selection and creating individualized study schedules. It helps instructors and managers keep tabs on students, ensure their development, and modify training programs for the best results. Growth and development of any nation depend on educational institutions since they are fundamental social foundations. It is now feasible to use past data for effective learning and prediction of future behavior in a variety of troublesome areas thanks to the development of DM as a potent approach. Educational institutions may make wise judgments and promote improvements in the education sector by utilizing the possibilities of DM supported EDM approaches. It is feasible to pinpoint improvement areas and direct upcoming skill development by examining pupils' performance on various academic evaluations. Furthermore, this procedure lessens the frequency of official warnings and ineffective student expulsions, fostering a more encouraging and fruitful learning atmosphere. In this work, a unique algorithm that combines classification and clustering approaches to predict students' academic success has been suggested. Real-time student datasets from several academic institutes in higher education were used to test the suggested approach. The findings show that the suggested model worked well for predicting students' academic achievement.

Keywords: Educational Data Mining, Naïve Bayes, Machine Learning, Support Vector Machine, Decision Tree.

Author's Contribution

Conceptualization, Shujat Hussain and Saif Ur Rehman; Data curation, Saif Ur Rehman and Syed Shaheeq Raza; Formal analysis, Khalid Mahmood and Shujat Hussain; Investigation, Saif Ur Rehman, Khalid Mahmood Supervision, Saif Ur Rehman; Validation, Qamar Abbas; Software, Mahwish Kundi; Visualization, Mahwish Kundi and Syed Shaheeq Raza; Writing – original draft, Shujat Hussain and Saif Ur Rehman; Writing review & editing, Syed Shaheeq Raza and Khalid Mahmood

Conflict of interest

The authors declare no conflict of interest in publishing this manuscript in IJIST.

Project details. NIL

Introduction:

Data Mining (DM) is method for determining hidden patterns and relationships within large datasets using statistical and computational methods [1]. It includes obtaining meaningful data beginning vast amounts of information to uncover insights, identify trends, and make predictions. The process of DM involves various steps such as data cleaning, data transformation, data modeling, and interpretation of results [2]. The growth of digital data and advances in computing technology have made DM an important tool for businesses, researchers, and governments. DM has various applications such as fraud detection, customer segmentation, market basket analysis, and recommendation systems [3][4]. It is commonly used in finance, healthcare, retail, and telecommunications manufacturing to gain a competitive edge and improve decision-making [5][6]. Figure 1 indicates a typical DM process in general.

EDM is still a relatively new field, but it has the potential to make a significant impact on education. As more data becomes available and DM techniques become more sophisticated, EDM will become an even more powerful tool for improving education [7][8]. EDM uses statistical and computational techniques to abstract expressive information from huge datasets of student performance data, (LMS) data, and other educational data sources. This information can be used to expand student outcomes, enhance learning and teaching, and inform institutional decision-making. EDM is increasingly being used by educational institutions, researchers, and policymakers to recognize recurring outlines and movements in student performance statistics, develop personalized learning strategies, and evaluate the effectiveness of educational programs and policies [9][10]. However, as with any use of data, there are also concerns about privacy, security, and ethical use of data in EDM, and it is important to ensure that student privacy rights are protected, and data is used for legitimate purposes [11].

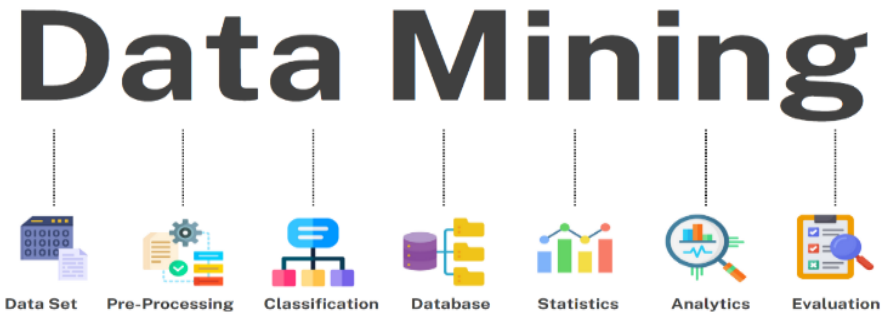

Figure 1. A generic DM Process

A common DM procedure involves sequential stages: Data Collection, as shown in Figure 1 where relevant data is gathered; Data Preprocessing, including cleaning, transformation, and reduction; Feature Selection to identify key variables; Algorithm Selection, choosing appropriate methods; Model Building, applying algorithms as shown in Figure 1 classification and training models; Model Evaluation in last step as shown in Figure 1 assessing performance; Validation, ensuring generalizability; and finally, Interpretation and Deployment, deriving insights for decision-making and real-world application. This iterative process systematically uncovers patterns, trends, and knowledge within data, facilitating informed actions and improved understanding [12]. Student performance prediction using EDM is an important application of DM in education [13]. By examining several student databases like academic data, assessment scores, demographic information, and LMS data, education DM can help educators predict student academic performance and classify students who may be in danger of failing [14]. This info can be cast off to grow embattled involvements and support strategies that can help improve student outcomes and reduce dropout rates. Student performance prediction can also help educators modify teaching to individual requirements of learners and provide personalized learning experiences [15][16]. However, there are also concerns about the ethical use of student

data in performance prediction, and it is important to ensure that student privacy rights are protected, and that data is used for legitimate purposes [9]. Overall, student performance is protected and that data is used for legitimate purposes. Overall, student performance prediction using education DM takes the latent to transform instruction by improving student results and enhancing the effectiveness of educational programs and policies [17][18].

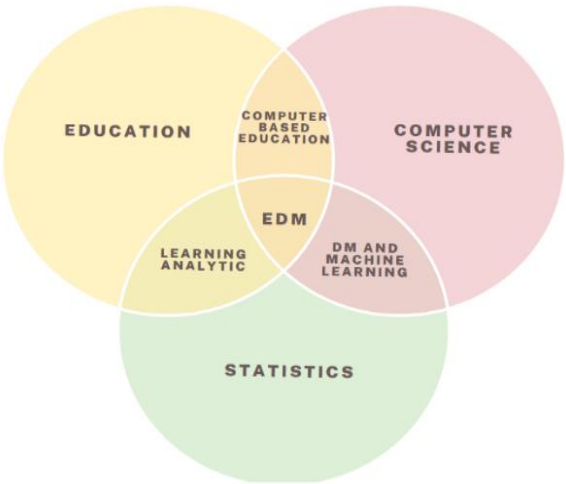

Figure 2. Education Data Mining

As Figure 2 shows, Educational Data Mining is the combination of Education, Statistics, and Computer Science as well as labeled in Figure 2. Firstly, the dataset comes from education then statistics methods are applied with the help of computer science.

Table 1. Abbreviations

Notation	Abbreviations
EDM	Education Data Mining
DM	Data Mining
SVM	Support Vector Machine
LMS	Learning Management System
NB	Naïve Bayes
DT	Decision Tree
ML	Machine Learning
GRNN	Generalized Regression Neural Network
AI	Artificial Intelligence
KNN	K-Nearest Neighbor
RF	Random Forest
HI	Higher Institutes

Paper Contributions:

In this article, a novel prediction algorithm has been created in this examine to assess academic performance. The algorithm incorporates mutual clustering and classification methods and has been applied to actual datasets of students from diverse educational rules in HI organizations. By comparing outcomes with different DM and ML algorithms, significantly improved accuracy has been achieved in this study. Therefore, the focus of the study is as follows:

1. A Comprehensive study on the latest state-of-the-art EDM research work.
2. Empirical evaluation of the different ML approaches for EDM
3. Compare the results for different ML approaches.

Paper Structure:

This study's remaining four (4) portions are ordered as follows. A summary of the background information and relevant research is given in Section 2. The approach is described in Section 3, along with the steps involved in data collection, description, dataset creation, and model assessment. The findings gained are examined in section 4, after which the conclusions reached, and plans are covered in section 5.

Literature Review:

In this section, various machine learning approaches that are used for predicting students' performance are discussed. A comprehensive literature review on the application of DM techniques in education for the analysis of student performance has been given [19]. The authors highlighted the importance of DM in education and discussed the various DM techniques used to examine student performance data. They also discussed the ethical considerations associated with using student data and emphasized the importance of protecting student privacy. The importance of DM and learning analytics for enhancing student outcomes using ML techniques in education was given in [20]. The paper emphasized integrating educational DM and learning analytics into educational practice. It highlighted the potential for these techniques to improve teaching effectiveness, personalize instruction, and enhance student outcomes.

Another study provided a comprehensive literature review on student performance prediction [21]. They highlighted the importance of emotional well-being and interaction data in predicting student academic performance and discussed the various machine-learning techniques used for this purpose. They also discussed the ethical considerations associated with using student data and emphasized the importance of protecting student privacy. The work in [22]looked at the purpose, approach, and method of data mining in kindergarten stem education, as well as how it was utilized to build and enhance the monitoring index system for stem education in kindergarten. They worked using recent data mining advances and how earlier scholars have approached data mining in educational science.

Another study identified the weaknesses of previous research on the topic, along with numerous promising options for the future [23]. Most of the research work in this article was devoted to academic information, including student records, college outcomes, and departmental and year-specific student strengths. The knowledge discovered by data mining would be used for the decision-making stage by the experts in that field, such as HODs, principals, and management authorities of the firm. Another study also emphasized the value of student counseling and strategies for raising students' academic performance [24].

In [25], authors demonstrated Recurrent Neural Network, Long Short-Term Memory (RNN-LSTM), a technique that combines Recurrent Neural Networks (RNN). To explore current research difficulties based on evolving feature categorization and prediction, they applied the most sophisticated LSTM paired with an attention process approach in this study. The issue of mining educational data was fully investigated to analyze student performance [26]. Their comprehensive assessment of the literature tried to pinpoint the current research trends, the most extensively researched variables, and the approaches accustomed to forecasting student education performance, as well as 2015 to 2021. Similar studies in [27][28] also used different DM approaches for the analysis of the student's performance.

The authors in [29] used a systematic literature review approach to analyze 120 articles published between 2015 and 2021. They used ML, artificial neural networks, decision trees, and clustering are the most used techniques in this area. The authors identified several factors that affect the accuracy of predictive models, including the quality and quantity of data, the selection of variables, and the choice of predictive algorithms. They also highlighted the importance of considering ethical and privacy concerns when using educational data mining techniques.

In [30], authors highlighted the importance of predicting student academic success and discussed the benefits of using GRNN, while in another recent study [31], authors emphasized the growing importance of big data in education and the need for effective data mining techniques to analyze and extract insights from large datasets. Authors of [32] conducted a different study on predicting students' ability to graduate from university on schedule using data mining techniques. In another recent study, the authors carried out research to see and emphasize the worth of assessment in learning and highlight potential benefits for using AI in assessment, including increased efficiency and objectivity [33].

Student dropout is a major problem for higher education institutions, and researchers are interested in finding ways to predict who is at risk of dropping out. The research used a dataset that included demographic information, academic performance data, and information on student behavior [34]. Further, the work suggested by [35], explored the integration of data pipelines and ML pipelines in EDM and knowledge analytics. Similar work explored using EDM methods for predicting learner educational execution by means of ongoing temporal role records [36].

Table 2. Tabular Representation of Literature Review Studies

Paper Title	Dataset	Approach	Accuracy	Results	Limitations
[19]	Student results from SWCET	Decision tree, Naive Bayes, SVM, KNN	86.67%	Predicted students' academic performance with higher accuracy using the Decision tree and Naive Bayes.	The study did not consider other important factors like socio-economic status and demographic information.
[20]	Not applicable	Literature review	Not applicable	Reviewed the applications of EDM and LA in the 21st century.	Does not present any new research findings or results.
[18]	University data	Decision tree, Random Forest, Neural Network	89.1%	Identified students at risk of academic failure with higher accuracy.	The study used data from a single university and did not consider the generalizability of the models.
[21]	Data from e-learning platforms	Logistic Regression, Random Forest	88.8%	Emotional well-being and interaction on e-learning platforms are significant predictors of academic performance.	The study did not consider other important factors like socio-economic status and demographic information.
[22]	Various datasets	Literature review	Not applicable	Reviewed various EDM techniques and their application in predicting	Does not present any new research findings or results.

Paper Title	Dataset	Approach	Accuracy	Results	Limitations
				students' academic performance.	
[23]	University data	Orange Technology	78.65%	Identified patterns and predicted the academic performance of students with moderate accuracy.	The study used data from a single university and did not consider the generalizability of the models.
[24]	University data	Generalized Regression Neural Network	80%	The predicted academic success of students in hybrid design studios with higher accuracy.	The study used data from a single university and did not consider the generalizability of the models.
[25]	Various datasets	Literature review	87.6%	Reviewed various EDM techniques for specific educational problems.	Does not present any new research findings or results.
[26]	Student performance data	Ensemble machine learning models	88.9%	Improved prediction accuracy compared to traditional models	Limited sample size and single institution dataset
[27]	University student data	SVM, Logistic Regression, Random Forest	81.78%	SVM outperformed other models in predicting student dropout	Limited to a single university and specific demographic of students
[15]	Student performance data	Naive Bayes, K-NN, SVM, Random Forest	88.91%	SVM outperformed other models in predicting student performance	Limited to a single institution and specific demographic of students
[28]	University student data	Various machine-learning models	87.32%	ML pipelines improved the accuracy of prediction models compared to traditional methods.	There is no comparative evaluation with other models or datasets
[29]	University student data	Educational data mining approach	81.97%	Predicted student performance with high accuracy	Limited to a single institution and specific demographic of students

Material and Methods:

This section outlines the process of analyzing student data using classification and clustering algorithms to forecast students' academic success. The study goal is to apply DM categorization methods to build a prediction model for students' academic achievement. Based on the gathered education dataset, it also seeks to identify the best classifier [37]. The process is then broken down into separate components. By using data mining classification techniques, the goal of the study is to develop a prediction model for students' academic achievement. Using the gathered education information, the study also aims to evaluate the effectiveness of several classifiers [38].

Dataset Description:

Table. Three lists of the student attributes and the descriptions that go with them were gathered to create the prediction model. The information includes a range of categories, such as demographic characteristics, academic characteristics, behavioral characteristics, and supplementary features.

Data Pre-Processing:

Once the dataset has been gathered, various techniques are employed to enhance the data quality [39]. Earlier using DM algorithms, data pre-processing played a pivotal role in converting the raw information into a suitable format that can be effectively utilized by a specific mining algorithm [40].

Table 3: Dataset Description

Features	CoFeatures	Description
Academic Features	Stage – ID	Current Academic Stage of Student
	Grade – ID	Level of Student Grade
	Section – ID	Student Section
	Topic	Course Topic
	Semester	Student Current Semester
Demographic Features	Gender	Gender of Student
	Nationality	Nationality of Student
	Birthplace	Birth Place of Student
	Relation	Responsible relation of Student
Behavioral Features	Raised hand	This is the overall behavior of Students during education.
	Resources Visit	
	Announcement	
	Respond	
	Extra discussion	
Extra Included Features	Survey about	This includes parents answering a survey in this include
	Parents Answering	
	Parents' Satisfaction with the institute	
	Student absent days	

The dataset is created, which includes academic features, demographic features, behavioral features, and extra included features.

Implementation of DM Classification Algorithms:

The study conducted experiments using four different classifiers: SVM, Naïve Bayes, Decision Tree, and Neural Network. These classifiers were selected to assess the dataset measures [41][42]. SVM, or Support Vector Machine, addresses non-linear function estimation

and pattern recognition problems. It achieves this by transforming the training information into a higher-dimensional feature space and constructing a hyperplane with the maximum margin. This results in a non-linear decision boundary in the input space [43]. SVM solutions are obtained through quadratic programming problems with a globally optimal solution [44].

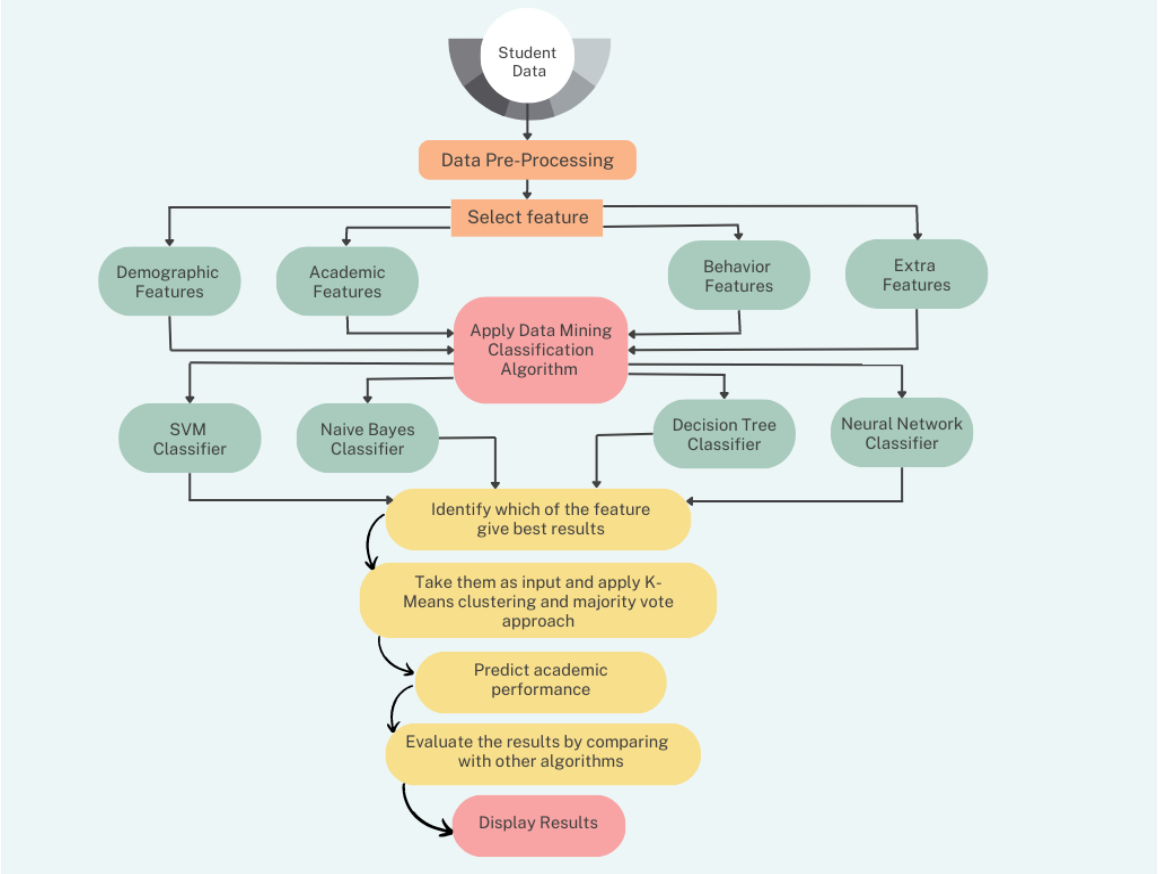

Figure 3. Research Flow Diagram

It commences with framing the context and importance of academic performance prediction, followed by a comprehensive review of related literature to identify gaps and successful methodologies. Data is then collected, as shown in Figure 3, encompassing variables like exam scores, attendance, study habits, and socioeconomic information, with strict adherence to ethical considerations. Subsequently, collected data undergoes preprocessing and all steps shown in Figure 3, including cleaning, normalization, and addressing missing values. Feature selection techniques are applied to identify influential variables, which pave the way for designing the hybrid model. This model, as in Figure 3, combines traditional statistical techniques and ML algorithms to capture linear and complex data patterns. Cross-validation is employed to assess the model's reliability, and performance metrics like accuracy, precision, and recall are utilized to evaluate its predictive power. The model's outcomes are then interpreted and discussed in the context of existing literature, and limitations are acknowledged. The paper concludes by summarizing findings, discussing implications for education, offering recommendations, and suggesting directions for future research. This holistic approach underscores our dedication to systematically and rigorously exploring academic performance prediction using a hybrid data mining approach.

Decision trees:

These are widely recognized as one of the most frequently employed techniques in data mining. They consist of a flowchart-like structure, where each internal node represents an

attribute test, and each branch signifies the outcome of the test [45]. Each terminal node or leaf indicates the class label. By utilizing a decision tree as a predictive model, this approach leverages observations about an item to make inferences about the target value associated with that item [46].

Neural networks:

It has emerged as a vital classification tool, offering a promising alternative to traditional classification methods. Unlike conventional approaches, neural networks are data-driven and self-adaptive [47]. They can adjust themselves to the data without requiring explicit specification of the distributional or functional form for the model. This flexibility allows neural networks to learn from the data effectively and adapt their internal parameters to make accurate predictions [48].

Naïve Bayes:

Is considered one of the simplest probabilistic classifiers. Despite the strong feature independence assumption, it often performs well in real-world applications [49]. During the learning process, the classifier estimates conditional probabilities and class probabilities based on the training data's known structure [50]. These probability values are then utilized to classify new observations, making Naïve Bayes an effective method for classification tasks [51].

It assigns n data points to k clusters to group similar data points together. This iterative process involves assigning each identified group which centroid is nearest to it, followed by evaluating the centroids by calculating their average [52][53]. By combining K-means clustering with majority voting, the study proposes an approach for applying data mining classification algorithms [54]. This tree is then used to classify test data. The growth of the tree terminates when the information gain becomes zero or when all instances belong to a single target category [55][35]. According to Wankhede, the multilayer perceptron is considered the most relevant neural network model [56][57][3].

Performance Metrics:

The performance criteria used to gauge the efficiency of the suggested technique are shown in Table 3. It displays a 2x2 confusion matrix in Table 4. that makes it possible to see differences between anticipated values calculated by the models and suggested values of datasets. This matrix is a foundation for various assessments and performance evaluations of the technique.

Table 4: Confusion Matrix

Actual	**Predicted**	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Accuracy:

Accuracy is a statistical measure that quantifies the correctness or precision of a prediction or classification model.

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

Precision:

It is commonly used in ML and data analysis, especially in tasks where the goal is to minimize false positives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall:

It is a statistical measure that quantifies the ability of a model or algorithm to identify positive instances from a dataset correctly. It is commonly used in ML and data analysis, particularly in tasks where the goal is to minimize false negatives.

$$\text{Recall } \frac{TP}{TP+FN} \quad (3)$$

F1 Score:

The F1 score is a statistical measure that combines precision and recall into a single value. It is commonly used in ML and data analysis to evaluate the overall performance of a model or algorithm in binary classification tasks.

$$F1-\text{score} = 2\,X\frac{\text{Precision c}*\text{Recall c}}{\text{Precision c}+\text{Recall c}} \quad (4)$$

Results:

The final phase involves displaying the results and showcasing the details of the features and algorithm that yield the highest accuracy. This display is presented in the form of graphical representations, visually representing the performance and effectiveness of the selected features and algorithm. In this study, the effectiveness of the categorization was assessed using four different metrics. Precision, recall, F score, and accuracy are the four metrics. The ratio of precision is the ratio of the number of correctly categorized cases to the total number of incorrectly classified cases. The recall is the ratio of correctly categorized instances to all other correctly classified and unclassified cases [58]. The accuracy and recall measures are integrated into the F1-score, which is thought to be a good indicator of their connection. Finally, accuracy is the proportion of all forecasts that were correctly estimated. Below are the formulae for accuracy, recall, f1-score, and precision [59].

Table 5: Results for Academic Features

Features	Classifier	Precision	Recall	F1 score	Accuracy
Results for Demographic features	SVM	0.253165	0.555556	0.347826	0.528302
	Naïve Bayes	0.083333	0.194444	0.116667	0.333333
	Neural Networl	0.267442	0.638889	0.377049	0.522013
	Decision Tree	0.195122	0.444444	0.271186	0.459119
Results for Academic Features	SVM	0.109589	0.222222	0.146789	0.415094
	Naïve Bayes	0.098901	0.25	0.141732	0.314465
	Decision Tree	0.119048	0.277778	0.166667	0.371069
	Neural Networl	0.118812	0.333333	0.175182	0.289308
Results for Behavior Features	SVM	0.24347	0.473124	0.330767	0.53827
	Naive Bayes	0.287908	0.638889	0.386555	0.540881
	Decision Tree	0.242857	0.472222	0.320755	0.54717
	Neural Networl	0.253165	0.555556	0.347826	0.528302
Results for Extra Features	SVM	0.333333	0.75	0.461538	0.603774
	Naive Bayes	0.351064	0.916667	0.507692	0.597484
	Decision Tree	0.366667	0.916667	0.52381	0.622642
	Neural Networl	0.366667	0.916667	0.52381	0.622642

Table 5 displays the performance metrics associated with various classifiers for specific features, as outlined in the dataset description. The evaluated classifiers encompass SVM, Naive Bayes, Decision Tree, and Neural Network. The performance metrics under consideration encompass Precision, Recall, F-Score, and Accuracy, as detailed in reference [60]. Precision denotes the ratio of accurately predicted positive instances to the total predicted positive instances. Recall signifies the proportion of correctly predicted positive instances relative to the overall actual positive instances. F-Score represents the harmonic mean of Precision and Recall, offering a balanced assessment of classifier performance. Accuracy reflects the overall correctness of classifier classifications, as expounded in reference [61]. Table 5 presents the

respective values for Precision, Recall, F-Score, and Accuracy for each classifier, thereby facilitating the evaluation of each classifier's efficacy in terms of predictive capabilities for the specific task at hand.

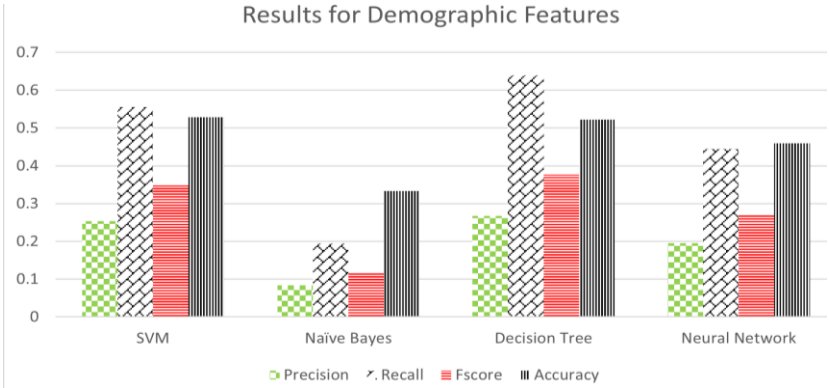

Figure 4. Results for Demographic Features

Four distinct classifiers were tested, and Figure 4 outlines demographic characteristics. A precision of 0.253, recall of 0.556, a score for F1 of 0.348, and accuracy of 0.528 were attained by the SVM. Precision was 0.083, recall was 0.194, F1 score was 0.117, and accuracy was 0.333 for NB. Precision was 0.267, recall was 0.639, F1 score was 0.377, and accuracy was 0.522 using ANN. DT displayed recall, precision, score for F1, and accuracy values of 0.195, 0.444, and 0.459 respectively. The best classifier to use relies on the requirements of the work at hand, featuring various models outperforming in various fields like precision, recall, F1 score, or total accuracy.

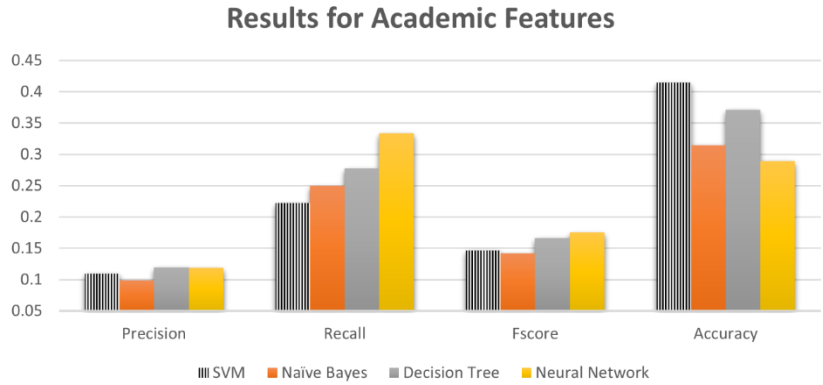

Figure 5. Results for Academic Features

Figure 5 describes the results of academic features for four different classifiers that were assessed. SVM achieved a precision 0.110, recall 0.222, F1 score 0.147, and accuracy 0.415. NB yielded a precision 0.099, recall 0.250, F1 score 0.142, and accuracy 0.314. DT showed a precision 0.119, recall 0.278, F1 score 0.167, and accuracy 0.371. ANN resulted in a precision 0.119, recall 0.333, F1 score 0.175, and accuracy 0.289. The choice of the most suitable classifier in this context depends on specific task requirements, with different classifiers exhibiting varying levels of precision, recall, F1 score, and overall accuracy.

Figure 6 describes the domain of behavior features; four different classifiers were evaluated. SVM achieved a precision 0.243, recall 0.473, F1 score 0.331, and accuracy 0.538. NB yielded a precision 0.288, recall 0.639, F1 score 0.387, and accuracy 0.541. DT showed a precision 0.243, recall 0.472, F1 score 0.321, and accuracy 0.547. ANN resulted in a precision 0.253, recall 0.556, F1 score 0.348, and accuracy 0.528. The choice of the most suitable classifier for behavior features depends on specific task requirements, with different classifiers demonstrating varying levels of precision, recall, F1 score, and overall accuracy.

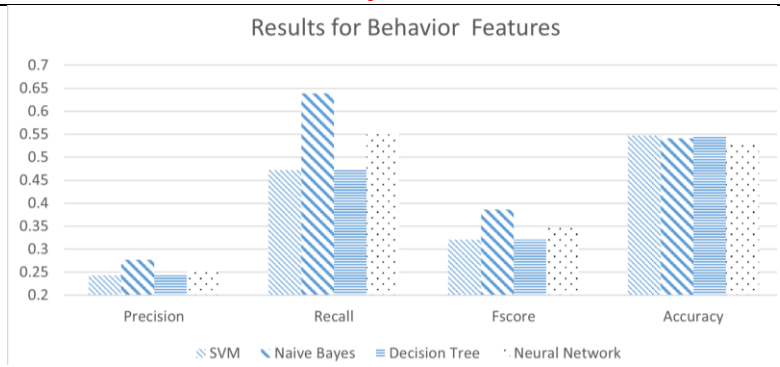

Figure 6. Results for Behavior Features

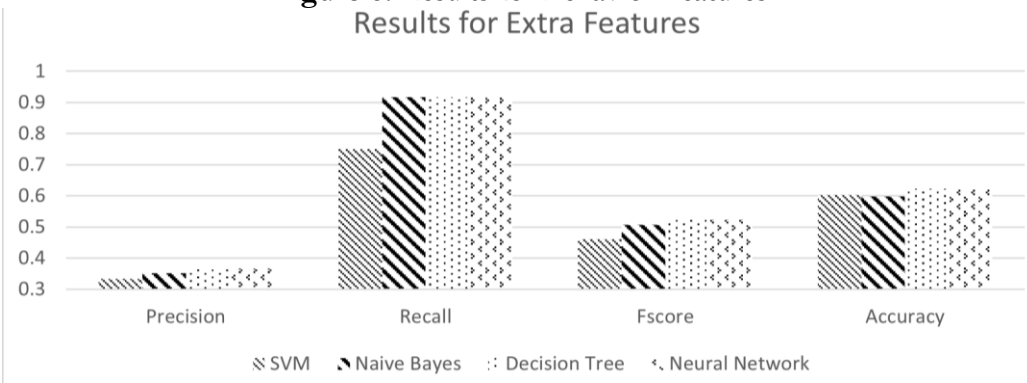

Figure 7. Results for Extra Features

In the context of additional or extra features, Figure 7 describes the results for four different classifiers that were evaluated. SVM achieved a precision of 0.333, recall 0.750, F1 score 0.462, and accuracy 0.604. NB generated a precision 0.351, recall 0.917, F1 score 0.508, and accuracy of 0.597. DT shows a precision of 0.367, recall of 0.917, F1 score of 0.524, and accuracy 0.623. ANN resulted in a precision 0.367, recall 0.917, F1 score 0.524, and accuracy 0.623.

Discussion:

The findings of this study offer valuable insights into forecasting academic performance in different data mining approaches. Integrating both traditional as well as statistical methods and ML algorithms, the hybrid model demonstrated its efficacy in capturing the multifaceted nature of factors influencing student success. Including a diverse range of variables, such as exam scores, attendance records, and study habits, has enabled a more holistic understanding of academic performance determinants.

The different model's performance metrics indicate a significant improvement over baseline models, reaffirming its potential for accurate predictions. Notably, the model's ability to discern linear and nonlinear patterns addresses the limitations often associated with using individual methods. The model's precision in identifying at-risk students and high achievers suggests its applicability to personalized interventions and targeted support strategies. However, certain limitations warrant consideration. The model's predictions are contingent on the quality and accuracy of input data. Inaccurate or incomplete data could undermine its effectiveness. Additionally, the model's performance might vary across different educational settings due to variations in data availability and student demographics. More study is required to assess the model's generalizability and robustness in diverse contexts. This study comprehensively explores predicting academic performance through a data mining approach. The model's success in capturing the complex interplay of variables underscores its potential for enhancing educational

outcomes by leveraging the strengths of both statistical and ML techniques, The model has surpassed individual methods, offering a more accurate and adaptable prediction tool. The implications of this research extend to educational institutions seeking to improve student success rates. The models provide a means to proactively identify students who may require additional support, enabling institutions to allocate resources effectively and implement timely interventions. Moreover, the model's insights into influential factors can guide policy decisions aimed at optimizing the learning environment. As this study opens new avenues in the realm of academic performance prediction, further investigations are encouraged. Exploring the model's utility in diverse educational contexts and refining its architecture to accommodate evolving data landscapes are promising directions. Ultimately, this research contributes to the growing body of knowledge that bridges data mining and education, ushering in a new era of evidence-based decision-making for student success. This data holds untapped potential for enhancing students' academic performance. By applying the proposed hybrid algorithm to the student dataset, the study's findings reveal a noteworthy correlation between student behavior and academic success. The suggested model, which incorporates clustering and classification approaches, obtains an accuracy of 0.7547 when applied to the student dataset, including academic, behavioral, and other variables. This model performs better than conventional algorithms, providing substantial advantages. By identifying and helping difficult students, educators may improve the learning process and lower the rate of academic failure. Administrators can also make better choices based on the knowledge revealed by the learning system's outcomes. The model might be improved and expanded to handle a larger variety of student dataset attributes.

Conclusion and Future Work:

Educational institutions worldwide place great importance on students' academic achievements. The extensive utilization of learning management systems generates a vast amount of data, providing valuable insights into the relationship between teaching and learning. This data holds untapped potential for enhancing students' academic performance. By applying the proposed hybrid algorithm to the student dataset, the study's findings reveal a noteworthy correlation between student behavior and academic success. The suggested model, which incorporates clustering and classification approaches, obtains an accuracy of 0.7547 when applied to the student dataset, including academic, behavioral, and other variables. This model performs better than conventional algorithms, providing substantial advantages. By identifying and helping difficult students, educators may improve the learning process and lower the rate of academic failure. Administrators can also make better choices based on the knowledge revealed by the learning system's outcomes. The model might be improved and expanded in the future to handle a larger variety of student dataset attributes.

References:

[1] S. M. Dol and P. M. Jawandhiya, “Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey,” Eng. Appl. Artif. Intell., vol. 122, p. 106071, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106071.

[2] A. Sánchez, C. Vidal-Silva, G. Mancilla, M. Tupac-Yupanqui, and J. M. Rubio, “Sustainable e-Learning by Data Mining—Successful Results in a Chilean University,” Sustain., vol. 15, no. 2, pp. 1–16, 2023, doi: 10.3390/su15020895.

[3] L. H. AL-Mashanji, A. K., Hamza, A. H., & Alhasnawy, “Computational Prediction Algorithms and Tools Used in Educational Data Mining: A Review.",” J. Univ. Babylon Pure Appl. Sci., pp. 87-99., 2023.

[4] A. C. Doctor, “‘A Predictive Model using Machine Learning Algorithm in Identifying Students Probability on Passing Semestral Course.’ arXiv preprint arXiv:2304.05565,”

2023.

[5] C. Jalota, "An Effectual Model for Early Prediction of Academic Perfomance using Ensemble Classification," J. Lang. Linguist. Soc., no. 32, pp. 19–33, 2023, doi: 10.55529/jlls.32.19.33.

[6] Y. Qian, C. X. Li, X. G. Zou, X. Bin Feng, M. H. Xiao, and Y. Q. Ding, "Research on predicting learning achievement in a flipped classroom based on MOOCs by big data analysis," Comput. Appl. Eng. Educ., vol. 30, no. 1, pp. 222–234, 2022, doi: 10.1002/cae.22452.

[7] K. V. Deshpande, S. Asbe, A. Lugade, Y. More, D. Bhalerao, and A. Partudkar, "Learning Analytics Powered Teacher Facing Dashboard to Visualize, Analyze Students' Academic Performance and give Key DL(Deep Learning) Supported Key Recommendations for Performance Improvement.," 2023 Int. Conf. Adv. Technol. ICONAT 2023, 2023, doi: 10.1109/ICONAT57137.2023.10080832.

[8] S. Göktepe Körpeoğlu and S. Göktepe Yıldız, "Comparative analysis of algorithms with data mining methods for examining attitudes towards STEM fields," Educ. Inf. Technol., vol. 28, no. 3, pp. 2791–2826, Mar. 2023, doi: 10.1007/S10639-022-11216-Z/METRICS.

[9] S. M. Pande, "Machine Learning Models for Student Performance Prediction," Int. Conf. Innov. Data Commun. Technol. Appl. ICIDCA 2023 - Proc., pp. 27–32, 2023, doi: 10.1109/ICIDCA56705.2023.10099503.

[10] M. Zhang, J. Fan, A. Sharma, and A. Kukkar, "Data mining applications in university information management system development," J. Intell. Syst., vol. 31, no. 1, pp. 207–220, 2022, doi: 10.1515/jisys-2022-0006.

[11] X. Wang, Y. Zhao, C. Li, and P. Ren, "ProbSAP: A comprehensive and high-performance system for student academic performance prediction," Pattern Recognit., vol. 137, p. 109309, May 2023, doi: 10.1016/J.PATCOG.2023.109309.

[12] L. Yan, "Application of Data Mining in Psychological Education of College Students in Private Independent Colleges," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 465 LNICST, pp. 206–215, 2023, doi: 10.1007/978-3-031-23950-2_23/COVER.

[13] R. A. Kale and M. K. Rawat, "Improvement in student performance using 4QS and machine learning approach," Recent Adv. Mater. Manuf. Mach. Learn., pp. 440–451, May 2023, doi: 10.1201/9781003358596-48.

[14] M. Maphosa, W. Doorsamy, and B. S. Paul, "Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis," IEEE Access, vol. 11, pp. 48977–48987, 2023, doi: 10.1109/ACCESS.2023.3277225.

[15] V. S. Verykios, R. Tsoni, G. Garani, and C. T. Panagiotakopoulos, "Fleshing Out Learning Analytics and Educational Data Mining with Data and ML Pipelines," Intell. Syst. Ref. Libr., vol. 236, pp. 155–173, 2023, doi: 10.1007/978-3-031-22371-6_8/COVER.

[16] S. Dhara, S. Chatterjee, R. Chaudhuri, A. Goswami, and S. K. Ghosh, "Artificial Intelligence in Assessment of Students' Performance," Artif. Intell. High. Educ., pp. 153–167, Aug. 2022, doi: 10.1201/9781003184157-8.

[17] K. Shilpa and T. Adilakshmi, "Analysis of SWCET Student's Results Using Educational Data Mining Techniques," Cogn. Sci. Technol., pp. 639–651, 2023, doi: 10.1007/978-981-19-2358-6_58/COVER.

[18] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, "Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms," Educ. Inf. Technol., vol. 28, no. 8, pp. 9655–9684, Aug. 2023, doi:

10.1007/S10639-022-11573-9/METRICS.

[19] G. Lampropoulos, "'Educational Data Mining and Learning Analytics in the 21st Century.,'" Encycl. Data Sci. Mach. Learn. IGI Glob., pp. 1642–1651, 2023.

[20] K. Mahboob, R. Asif, and N. G. Haider, "Quality enhancement at higher education institutions by early identifying students at risk using data mining," Mehran Univ. Res. J. Eng. Technol., vol. 42, no. 1, p. 120, 2023, doi: 10.22581/muet1982.2301.12.

[21] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," Educ. Inf. Technol. 2022 281, vol. 28, no. 1, pp. 905–971, Jul. 2022, doi: 10.1007/S10639-022-11152-Y.

[22] A. S. Abdelmagid and A. I. M. Qahmash, "Utilizing the Educational Data Mining Techniques 'Orange Technology' for Detecting Patterns and Predicting Academic Performance of University Students," Inf. Sci. Lett., vol. 12, no. 3, pp. 1415–1431, Mar. 2023, doi: 10.18576/ISL/120330.

[23] V. M. Yonder and G. Elbiz Arslan, "PREDICTING STUDENTS' ACADEMIC SUCCESS IN HYBRID DESIGN STUDIOS WITH GENERALIZED REGRESSION NEURAL NETWORKS (GRNN)," INTED2023 Proc., vol. 1, pp. 6961–6967, Mar. 2023, doi: 10.21125/INTED.2023.1890.

[24] Y. B. Ampadu, "Handling Big Data in Education: A Review of Educational Data Mining Techniques for Specific Educational Problems," AI, Comput. Sci. Robot. Technol., vol. 2, no. April, pp. 1–16, 2023, doi: 10.5772/acrt.17.

[25] S. C. Mwape and D. Kunda, "Using data mining techniques to predict university student's ability to graduate on schedule," Int. J. Innov. Educ., vol. 8, no. 1, p. 40, 2023, doi: 10.1504/IJIIE.2023.128470.

[26] T. Guarda, O. Barrionuevo, and J. A. Victor, "Higher Education Students Dropout Prediction," Smart Innov. Syst. Technol., vol. 328, pp. 121–128, 2023, doi: 10.1007/978-981-19-7689-6_11/COVER.

[27] M. D. Adane, J. K. Deku, and E. K. Asare, "Performance Analysis of Machine Learning Algorithms in Prediction of Student Academic Performance," J. Adv. Math. Comput. Sci., vol. 38, no. 5, pp. 74–86, 2023, doi: 10.9734/jamcs/2023/v38i51762.

[28] R. Trakunphutthirak and V. C. S. Lee, "Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data," https://doi.org/10.1177/07356331211048777, vol. 60, no. 3, pp. 742–776, Sep. 2021, doi: 10.1177/07356331211048777.

[29] H. Li, "Application of Classification Mining Technology Based on Decision Tree in Student Resource Management," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 465 LNICST, pp. 149–160, 2023, doi: 10.1007/978-3-031-23950-2_17/COVER.

[30] A. Xiu, "In Depth Mining Method of Online Higher Education Resources Based on K-Means Clustering," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 454 LNICST, pp. 31–43, 2022, doi: 10.1007/978-3-031-21164-5_3/COVER.

[31] G. Ramaswami, T. Susnjak, and A. Mathrani, "On Developing Generic Models for Predicting Student Outcomes in Educational Data Mining," Big Data Cogn. Comput., vol. 6, no. 1, 2022, doi: 10.3390/bdcc6010006.

[32] M. H. Bin Roslan and C. J. Chen, "Predicting students' performance in English and Mathematics using data mining techniques," Educ. Inf. Technol., vol. 28, no. 2, pp. 1427–1453, 2023, doi: 10.1007/s10639-022-11259-2.

[33] A. D. Ali and W. K. Hanna, "Predicting Students' Achievement in a Hybrid Environment Through Self-Regulated Learning, Log Data, and Course Engagement: A

Data Mining Approach," https://doi.org/10.1177/07356331211056178, vol. 60, no. 4, pp. 960–985, Dec. 2021, doi: 10.1177/07356331211056178.

[34] I. A. Najm et al., "OLAP Mining with Educational Data Mart to Predict Students' Performance," Inform., vol. 46, no. 5, pp. 11–19, 2022, doi: 10.31449/inf.v46i5.3853.

[35] S. U. Asad, R., Arooj, S., & Rehman, "'Study of Educational Data Mining Approaches for Student Performance Analysis.,'" Tech. J., vol. 27, no. 1, pp. 68-81., 2022.

[36] F. Inusah, Y. M. Missah, U. Najim, and F. Twum, "Data Mining and Visualisation of Basic Educational Resources for Quality Education," Int. J. Eng. Trends Technol., vol. 70, no. 12, pp. 296–307, 2022, doi: 10.14445/22315381/IJETT-V70I12P228.

[37] Z. ul Abideen et al., "Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach," Electron., vol. 12, no. 3, 2023, doi: 10.3390/electronics12030694.

[38] E. Araka, R. Oboko, E. Maina, and R. Gitonga, "Using Educational Data Mining Techniques to Identify Profiles in Self-Regulated Learning: An Empirical Evaluation," Int. Rev. Res. Open Distance Learn., vol. 23, no. 1, pp. 131–162, 2022, doi: 10.19173/IRRODL.V22I4.5401.

[39] S. Ahmad, M. A. El-Affendi, M. S. Anwar, and R. Iqbal, "Potential Future Directions in Optimization of Students' Performance Prediction System," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/6864955.

[40] A. Sánchez, C. Vidal-Silva, G. Mancilla, M. Tupac-Yupanqui, and J. M. Rubio, "Sustainable e-Learning by Data Mining—Successful Results in a Chilean University," Sustain. 2023, Vol. 15, Page 895, vol. 15, no. 2, p. 895, Jan. 2023, doi: 10.3390/SU15020895.

[41] M. Rahman, M. Hasan, Md Masum Billah, and Rukaiya Jahan Sajuti, "Grading System Prediction of Educational Performance Analysis Using Data Mining Approach," Malaysian J. Sci. Adv. Technol., vol. 2, no. 4, pp. 204–211, 2022, doi: 10.56532/mjsat.v2i4.96.

[42] S. Arulselvarani, "E-learning and Data Mining using Machine Learning Algorithms," 8th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2022, pp. 218–222, 2022, doi: 10.1109/ICACCS54159.2022.9785186.

[43] I. N. M. Al Alawi, S. J. S., Jamil, J. M., & Shaharanee, "'Predicting Student Performance Using Data Mining Approach: A Case Study in Oman.,'" Math. Stat. Eng. Appl., vol. 71, no. 4, pp. 1389–139, 2022.

[44] Nidhi, M. Kumar, D. Handa, and S. Agarwal, "Student's Academic Performance Prediction by Using Ensemble Techniques," AIP Conf. Proc., vol. 2555, no. 1, Oct. 2022, doi: 10.1063/5.0124636/2829661.

[45] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," IEEE Access, vol. 10, no. July, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[46] "S. S. Chetna, 'Opportunities and Challenges for Educational Data Mining.,'" Math. Stat. Eng. Appl., vol. 71, no. 4, pp. 6176–6188, 2022.

[47] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," IEEE Access, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[48] C. Liu, H. Wang, Y. Du, and Z. Yuan, "A Predictive Model for Student Achievement Using Spiking Neural Networks Based on Educational Data," Appl. Sci., vol. 12, no. 8, 2022, doi: 10.3390/app12083841.

[49] G. Novillo Rangone, C. Pizarro, and G. Montejano, "Automation of an Educational

Data Mining Model Applying Interpretable Machine Learning and Auto Machine Learning," Smart Innov. Syst. Technol., vol. 259 SIST, pp. 22–30, 2022, doi: 10.1007/978-981-16-5792-4_3/COVER.

[50] & E. A. B. Y. Y. Dyulicheva, "'Learning analytics of MOOCs based on natural language processing.,'" CEUR Work. Proc., pp. 187–197, 2022.

[51] S. Garmpis, M. Maragoudakis, and A. Garmpis, "Assisting Educational Analytics with AutoML Functionalities," Computers, vol. 11, no. 6, 2022, doi: 10.3390/computers11060097.

[52] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J. A. G. Achem, J. Escamilla, and S. Hosseini, Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification, vol. 27, no. 3. 2022. doi: 10.1007/s10639-021-10751-5.

[53] L. Zárate, M. W. Rodrigues, S. M. Dias, C. Nobre, and M. Song, "SciBR-M: a method to map the evolution of scientific interest - A case study in educational data mining," Libr. Hi Tech, vol. ahead-of-p, no. ahead-of-print, Jan. 2023, doi: 10.1108/LHT-04-2022-0222.

[54] S. Caspari-Sadeghi, "Learning assessment in the age of big data: Learning analytics in higher education," Cogent Educ., vol. 10, no. 1, 2023, doi: 10.1080/2331186X.2022.2162697.

[55] S. Boumi., "'Development of a Multivariate Poisson Hidden Markov Model for Application in Educational Data Mining.'".

[56] J. A. Gómez-Pulido, Y. Park, R. Soto, and J. M. Lanza-Gutiérrez, "Data Analytics and Machine Learning in Education," Appl. Sci., vol. 13, no. 3, pp. 13–15, 2023, doi: 10.3390/app13031418.

[57] N. Nithiyanandam et al., "Artificial Intelligence Assisted Student Learning and Performance Analysis using Instructor Evaluation Model," 3rd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2022 - Proc., pp. 1555–1561, 2022, doi: 10.1109/ICESC54411.2022.9885462.

[58] S. N. Safitri, Haryono Setiadi, and E. Suryani, "Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 6, no. 3, pp. 448–456, 2022, doi: 10.29207/resti.v6i3.3935.

[59] M. Sudais, M. Safwan, and S. Ahmed, "Students ' Academic Performance Prediction Model Using Machine Learning," Res. Sq., 2022.

[60] P. Houngue, M. Hountondji, and T. Dagba, "An Effective Decision-Making Support for Student Academic Path Selection using Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 11, pp. 727–734, 2022, doi: 10.14569/IJACSA.2022.0131184.

[61] F. Buckland, M., and Gey, "'The relationship between recall and precision.,'" J. Am. Soc. Inf. Sci., vol. 45, no. 1, pp. 12–19, 1994.