Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 06
June 2025

Early Prediction on Student Performance Using Machine Learning

Priyanka Madhav Achat, Poonam G. Kanade

Department of Computer Engineering

S.N.D College of Engineering & Research Center Nashik, Maharashtra

Department of Computer Engineering

S.N.D College of Engineering & Research Center Nashik, Maharashtra

Abstract: Predictingstudentacademicoutcomesthroughautomatedsystemshasbecomeincreas- ingly crucial due to the growing volume of data maintained by educational institutions This challenge is being tackled by the field of educational data mining (EDM), This issue is being addressed through educational data mining (EDM), a field that develops techniques to derive meaningful insights from such data to better understand students and their learning environments. Educational institutions frequently aim to estimate the number of students who will pass or fail to make informed preparations. Although many previous studies have focused on choosing the best classification algorithms, they often neglect the practical challenges of the data mining process, including high dimensionality, class imbalance, and misclassification issues, which can hinder model accuracy. Anticipating student academic outcomes is essential in designing effective educational strategies. Machine learning (ML) models can analyze historical academic data to identify patterns that predict student performance. This paper explores the application of various ML algorithms—such as Decision Trees, Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbors (KNN)—to forecast student success rates. The goal is to enable educational institutions to intervene early, providing support where necessary to improve learning outcomes.

Keywords—Educational Data mining, Educational Data Mining, Predicting Stu- dent Performance, Decision Tree, Ensemble.

1. INTRODUCTION

Graduate unemployment, particularly in technical fields such as Computer Science and Information Technology, is increasingly linked to inadequate academic performance. By analyzing academic records using machine learning, educators can detect early warning signs in student behavior or academic metrics. For example, in countries like Malaysia, CGPA (Cumulative Grade Point Average) serves as a primary performance indicator. Since performance is influenced by both educational and environmental factors, predictive systems that integrate historical data offer valuable insight for early interventions. This study uses Data Mining (DM) tools to analyze available data from past batches of students at the College of Computer Science IT (authors' institution) and extract useful information to explain the phenomena of low performance. Leveraging machine learning allows for the identification of at-risk students early in the academic process, facilitating timely intervention Student's performance at university level various courses has been evaluated by the many researcher in Malaysia. Most of researchers used CGPA as a the key performance index to analyse student's academic a performance. There are several factors have been considered to give a significant effect on student's academic achievement. Studies conducted in [3–5] presented the influence of ethnic and gender in the academic performance at university level. Unemployment among graduates, particularly in technical fields like Computer Science and IT, is partly attributed to poor academic performance. To address this, data mining techniques are used to analyze historical academic data and uncover performance-related patterns. In countries like Malaysia, CGPA is a common metric for evaluating student achievement, influenced by both academic and environmental factors. Leveraging machine learning allows for the identification of at-risk students early in the academic process, facilitating timely intervention. As a result, the instructors can focus more on such weak students to make them ready by the time summative assessments are scheduled. In this paper, we have applied different machine learning algorithms on the historic results of a course being taught in bachelor's in computer information systems program to find out the prediction accuracy.

2. RELATED WORKS

1. **Jiang et al. – Modeling Uncertainty in Student Data**

Jiang and colleagues introduced a sophisticated approach combining **Student's t-distribution Hidden Markov Models (t-HMM)** with **Nuisance Attribute Projection (NAP)** to model the uncertainties and fluctuations in student performance data. This hybrid technique addresses the challenges of noisy and heavy-tailed data distributions. The use of NAP allows for filtering out irrelevant or nuisance factors that could negatively affect predictive accuracy, while t-HMM facilitates the temporal modeling of student learning patterns. Their approach presents a more reliable and adaptable prediction framework suitable for real-world educational data environments.

2. **Evawaty Tanuar et al. – Machine Learning for Early Performance Prediction**

In their research, Evawaty Tanuar, Yaya Heryadi, Lukas, Bahtiar Saleh Abbas, and Ford Lumban Gaol applied **Generalized Linear Models**, **Decision Trees**, and **Deep Learning techniques** to predict student performance at early academic stages. Despite challenges such as the wide scope of educational factors and data quality issues, the authors explored how machine learning can enhance prediction capabilities. Their study emphasized the need for incorporating student feedback as a performance benchmark and called for more refined models grounded in frequently encountered real-world educational scenarios.

3. Ching-Chieh Kiu – Educational Data Mining for Social and Academic Insights

Ching-Chieh Kiu focused on using **Educational Data Mining (EDM)** methods, including **Decision Trees**, **Naïve Bayesian classifiers**, and **Neural Networks**, to analyze the impact of both academic and social factors on student success. The study highlighted that variables such as socioeconomic background, family education level, peer relationships, and involvement in social activities significantly influence academic outcomes. By applying data mining, the study uncovers hidden patterns that can aid educators in designing interventions and enhancing the overall learning experience.

4. Rosa M. Vasconcelos – Gender Differences in Engineering Education

Rosa M. Vasconcelos, President of the Pedagogical Council at the School of Engineering, University of Minho, examined gender-based differences in academic success among first-year engineering students. Her study explored variables including study habits, classroom participation, motivation, and learning styles. The analysis highlighted how gender can influence engagement and performance in technically demanding programs such as engineering. These insights aim to help educators and institutions implement more equitable teaching strategies and create inclusive academic environments.

5. Cory Brozina – Academic Integration and Commuter Student Success

Cory Brozina investigated the academic experiences of **commuter computer science students** compared to their residential peers. His research centered on the concept of **academic integration**, which encompasses students' connection to faculty, coursework, and their peer network. Brozina pointed out that commuter students often face barriers such as limited access to on-campus resources and social isolation. However, his findings suggest that with intentional academic support programs focused on integration, commuter students can perform at levels comparable to residential students. The study proposes targeted intervention strategies to enhance student engagement and academic achievement.

In this concept that refers to how well students connect with the academic community and engage with the academic experience, including faculty, coursework, and peer relatioships. Understanding the relationship between academic integration and the computer students success in this guide the development of the targeted interventions programs aimed at improving engagement, and overall the academic performance.

3. LITERATURE SURVEY

- **Essa Alhazmi**: In a blended Calculus course, Alhazmi applied educational big data and learning analytics to predict students' final grades using Principal Component Regression. Seven critical factors influencing academic performance were identified, including three traditional and four online-related factors.

- **Kethatireaddy & Kotagoda Sahithi**: This study emphasizes the importance of evaluating student learning outcomes as a key component of assessing educational institutions. It highlights the significance of student performance in identifying and addressing learning process challenges.

- **Yu et al.**: Yu and colleagues utilized sentiment analysis to extract affective information from student interactions, aiming to enhance predictive accuracy for early identification of students at risk of failing a subject.

- **Mohammed Afzal Ahmed, Prayuk Chianti & Prof. Mahesh T. R.**: This research proposed a model to predict student performance based on various input parameters using Bayes' Algorithm. Cluster techniques were employed to categorize students into distinct groups, aiding in performance prediction.

- **Hanan Abdullah Mengash**: Mengash developed a model to predict final-year performance of graduate-level students using previous academic marks. The study did not consider socio-economic or demographic parameters in its analysis.

- **Harry B. Zainal & A. Hasibuan**: This study applied cluster analysis to analyze student results, identifying key characteristics that affect performance. These characteristics were used to inform future predictions.

- **Mustafa Yağcı**: Yağcı utilized the ID3 decision tree algorithm to predict graduation grades of students based on university entrance examination and Ordinary Level results. The model demonstrated effectiveness in grade prediction.

- **Ching-Chieh Kiu**: Kiu conducted a data mining analysis to explore the relationship between students' backgrounds, social activities, and academic performance. The study employed Decision Trees, Naïve Bayes, and Neural Networks to analyze the data.

- **Rosa M. Vasconcelos**: Vasconcelos compared learning and academic success among first-year engineering students, focusing on gender differences. The study aimed to understand how gender influences academic expectations and outcomes.

- **Cory Brozina**: Brozina examined the support systems available to engineering commuter students, analyzing multiple constructs to determine if differences exist between residential and commuter students in terms of academic integration and success.

Proposed System& System Architecture

In This presents a comprehensive assessment based on a real prototype implementation and performance evaluation. In our configuration, an edge server serves a dual purpose: acting as the administrative controller of the IoT infrastructure while also meeting the application’s latency and privacy requirements. The system proposed in this study features a hybrid architecture utilizing both edge and cloud technologies:

Edge Server: Handles latency-sensitive tasks such as real-time data preprocessing and local model predictions. It ensures low-latency user interaction and privacy-sensitive operations.

Cloud Component: Manages complex operations, including model training, storage, and system-wide analytics. Advanced security frameworks are integrated to protect student data and maintain system integrity.

Modules include authentication, classification, data cleaning, and training pipelines, providing a full-stack predictive platform.

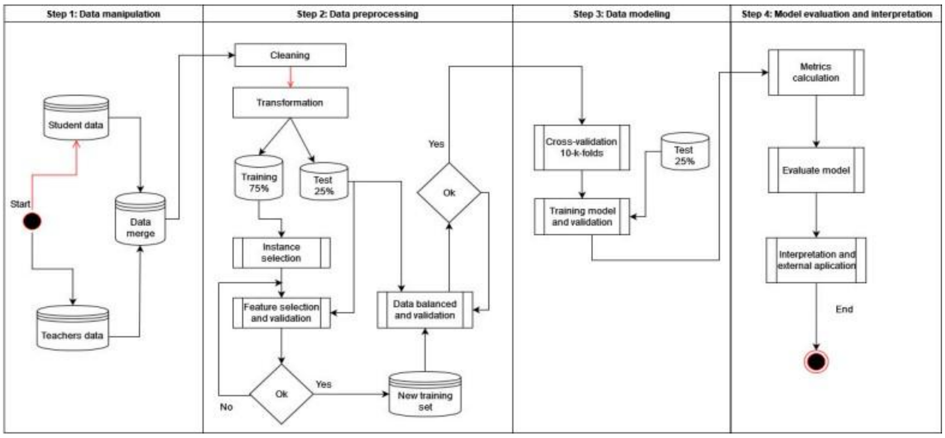

Step 1: Data manipulation
Step 2: Data preprocessing
Step 3: Data modeling
Step 4: Model evaluation and interpretation
Student data
Start
Data merge
Teachers data
Cleaning
Transformation
Training 75%
Test 25%
Instance selection
Feature selection and validation
Ok
Yes
No
New training set
Data balanced and validation
Ok
Yes
Cross-validation 10-k-folds
Test 25%
Training model and validation
Metrics calculation
Evaluate model
Interpretation and external aplication
End

Fig 1. Workflow of System

this architecture through the independent implementation of various micro services, which are then interconnected to form an IoT application. Additionally, we explore the potential for sharing these micro services across different IoT applications operating concurrently to improve interoperability. Ultimately, we conduct an in-depth performance analysis that emphasizes application latency, as well as CPU and memory usage. We operate under the assumption that the data owner is trustworthy and that data users have been authorized by the owner. The communication channels between the owner an users are secured using established security protocols such as SSL and TLS.

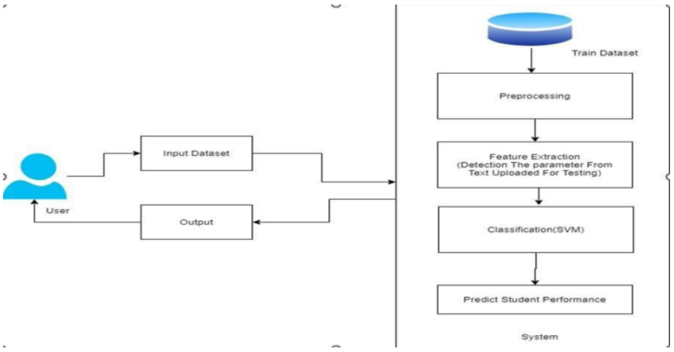

Train Dataset
Preprocessing
Feature Extraction (Detection The parameter From Text Uploaded For Testing)
Classification(SVM)
Predict Student Performance
System
Input Dataset
Output
User

Fig 2. System Architecture

Concerning the cloud server, ourthis approach addresses a more complex security model and that goes the ”semi- honest server” framework typically employed in the other secure semantic search schemes. In our model, a dishonest cloud server may attempt to provide incorrect or fabricated search results and gain access to sensitive information, but it will not engage in malicious activities such as the deleting and altering outsourced documents. Consequently, our secure semantic scheme is designed to ensure verifiability and confidentiality within this security framework.

1.User: Provides secure login interfaces for users. Users must authenticate using their username and password to connect to the server. New users can register by providing their details, allowing the server to create an account and manage upload/download rates.. If the user already exits directly can login the server else user must register their details such as username, password and Email id, into the server. Server will be create the account for this information is the entire user to maintain upload and download rate. Name will be set as user id. Logging in this is the beyond usually used to enter a specific page

2.Pre processing: After login, data users can search for files by name and download encrypted data. Users can also send trapdoor requests to the server, which, upon receiving permission from the data owner, allows the file to be downloaded in plain text.. Data user can also have a download file it will show an encrypted data. Data user can also send this is the trapdoor request to the server. Server can accept the request and then data user can takes permissions from the owner then the file it will downloaded in plain text

3.Classifier : Data owners register and log in to upload files into the database. They can also send requests to data users, facilitating data sharing and collaborationData Owner should register and Login. Data Owner will Uploads the files into the database. Data owner can also send request to the data user.

4.Training Data The cloud server logs in to access all data owners' information and stored data files. It can request keys from users and analyze file information, supporting model training and evaluation.. In this module Cloud Server can login. After login all data owners' information. Cloud server can see all users' information. Cloud server can see an all stored data files. Cloud server can give keys request to the user. Cloud server can attacker information of file .

4. METHODOLOGY

A. Artificial Neural Network (ANN)

ANN is widely utilized method in Educational Data Mining (EDM) that aims to replicate the human brain's architecture to address intricate issues. It comprises a collection of units that process a weighted array of inputs and generate an output. Numerous studies have employed ANN to forecast student performance. We also selected this technique for its capability to identify all potential interactions among variables and its proficiency in learning from a limited number of examples. Furthermore, previous research indicated that ANN models surpassed other classification methods in accurately categorizing applicants as either accepted or not accepted. In this study, we opted for the Multilayer Perceptron(MLP) architecture for the ANN model, given that the datasets were not sufficiently Artificial Neural Networks (ANN): Suitable for detecting complex interactions among variables with limited data using Multilayer Perceptron (MLP). large to necessitate more intricate topologies.

B. Decision Tree

A Decision Tree consists of a series of nodes organized vertically, with each node symbolizing a feature of an instance and its branches indicating potential values. This method is favored by researchers for its straightforwardness (e.g.[1, 6, 7, 9]). It offers a clear and uncomplicated approach to value prediction. Additionally, it presents several benefits, such as the ability to articulate rules that are easily comprehensible to users, effective performance with both categorical and numerical data, and Popular for its simplicity and ability to handle mixed data types minimal requirements for intricate data preparation .

C Support Vector Machine

This classification method constructs a hyperplane that distinguishes objects according to their respective classes. As the distance from the hyperplane to the nearest object increases, the generalization error of the SVM technique decreases. SVM has been utilized in a limited number of studies (e.g., [7, 23, 24]) and is employed in this research due to its effectiveness with small datasets. Moreover, it demonstrates faster performance compared to other techniques .Effective for small datasets due to its strong generalization capabilities.

D Naïve Bayes

Is a straightforward probabilistic method that utilizes Bayes' theorem while assuming independence among variables. It calculates probabilities for each object across all potential classes. In this research, we opted for this method due to its ease of use, strong performance in practical applications, computational efficiency and its widespread recognition in the academic literature. Efficient and effective in practical scenarios despite assuming variable independence.

E. Educational Data Mining

The educational process encompasses various aspects of student performance. To identify students who may face challenges in achieving academic success, it is essential to predict their future performance. When data is effectively transformed into actionable knowledge, it can be utilized for making informed predictions.

Consequently, this information has the potential to enhance the quality of education and facilitate students in reaching their academic goals. The field of educational data mining (EDM) employs data mining techniques to analyze information obtained from educational contexts. The application of EDM also supports the development of strategies aimed at improving student performance. Ultimately this will lead to enhanced teaching Provide insights into educational outcomes by analyzing academic and behavioral data.

5. RESULT AND EXPERIMENTAL EVALUATION

Dataset

data mining classification techniques: Artificial Neural Network (ANN), Decision Tree, Support Vector Machine (SVM), and Naive Bayes. Each model was developed using a 10-fold cross-validation method, where nine subsets of data were employed for training and one subset was reserved for testing. The dataset comprises various academic and non-academic attributes relevant to student performance analysis. The features include both numerical and categorical data types as This procedure was executed ten times once for each distinct subset, thereby optimizing the total number of observations used for testing. All models. The models were validated using 10-fold cross-validation.

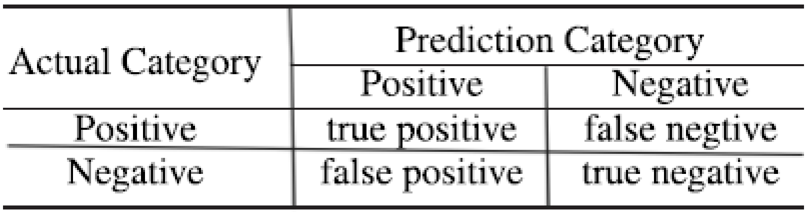
A Linear Regression model trained on 70% of the data yielded an accuracy of 98%, with low MAE and MSE scores.

Five models were tested with various target labels; models achieved accuracies ranging from 96% to 98%.

Comparative analysis with recent literature shows that this approach matches or exceeds existing accuracy benchmarks.

Evaluation Metrics

Table 1: Confusion matrix



Actual Category	Prediction Category	
	Positive	Negative
Positive	true positive	false negtive
Negative	false positive	true negative

Tabel2 :Dataset Definition

AttributeName	Attribute code	Type
Stud name	C1	char
RegistrationNumber	C2	Int
Gender	C3	Object
ClubActivities	C4	Object
TotalODcount	C5	Int
SportsParticipationScore	C6	Int
TestPreparation	C7	object
Subject1	C8	Int
Subject2	C9	Int
Subject3	C10	Int
Subject4	C11	Int

Subject1time	C12	Int
Subject2time	C13	Int
Subject3time	C14	Int
Subject4time	C15	Int
Subject5time	C16	Int
PrivateClass	C17	object
PhysicalFitness	C18	object
MentalFitness	C19	object

A predictive model is developed utilizing the Linear Regression Algorithm. The model is trained on 70% of the dataset, while the remaining 30% is randomly selected for testing purposes. The implementation is carried out using the Panda tool. The prediction score reflects the anticipated values generated by the linear regression models for various target data. The same features are evaluated against different target classes. For the initial model training, the features from the first semester are employed. When the target class corresponds to the first semester marks(FSM), the predictive model achieves an accuracy of 0.98, with a Mean Absolute Error of 1.85, a Mean Squared Error of 6.27, and an R² score of 0.98, indicating that the model is nearly flawless. However, the primary objective is to forecast the future performance of students. Consequently, the second model is trained using the second semester marks (SSM)while maintaining the same selected features.

Accuracy represents the proportion of correctly predicted outcomes and is calculated The formula for calculating the overall score is represented as follows: first semester mark(FSM), To verify the accuracy of the models, the average accuracy of all five models is calculated as.

Accuracy=(TP+TN)/TP+FP+TN+FN

To assess the validity of this prediction score, Models 4 and 5 have been developed. Model 4 establishes the Target Class by utilizing the highest scores from both the first and second semesters, while Model 5 determines the Target Class based on the lowest scores from the same semesters. Model 4 achieved an accuracy of 0.96, with a meanabsoluteerrorof2.89,ameansquarederrorof12. 98, and an R² score of 0.96. In contrast, Model 5 attained an accuracyof0.97,withameanabsoluteerrorof2.24,ame an squared error of 9.01, and an R² score of 0.97. To further validate the models' accuracy, the average accuracy across all five models is computed.

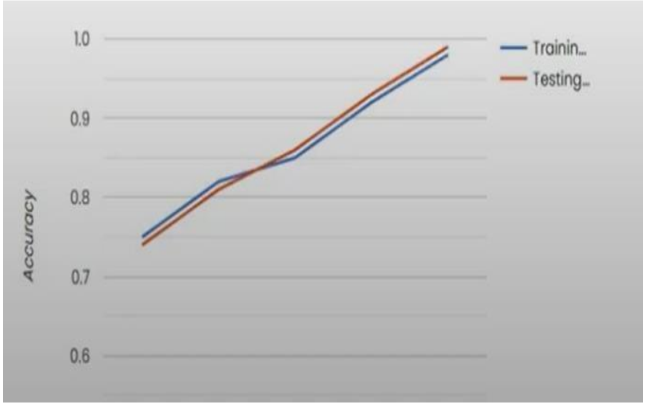


Figure 3:Accuracy of Prediction

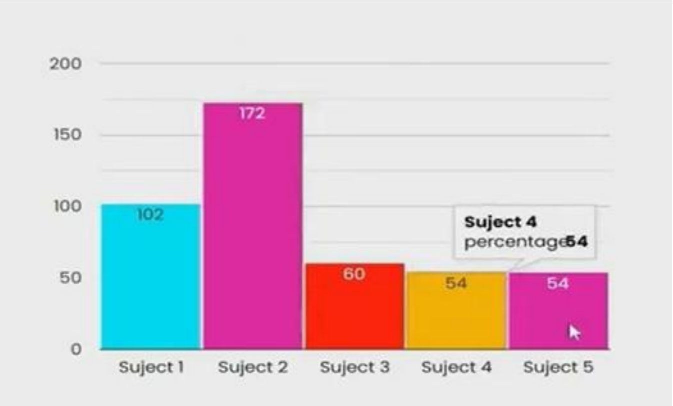


Figure 4: study time

Tableno3Comparisonofthestudywithrecentwork

PaperName	AccuracyRate	Algorithm
Proposed Paper: Predicting Educational Performance of Students Using ML	Up to 96.05%	SVM, J48, ANN, Naive Bayes
Student Academic Performance Prediction Using Supervised Learning	Up to 95.78% with Real AdaBoost	J48 Decision Tree
Predicting Admission Test Outcomes Using ML & Resampling	Up to 92%	SVM, AdaBoost, SMOTE-ENN
Early Prediction of Academic Achievement in Online Learning	85.03%	RandomForest,SMOTE
Using Data Mining to Predict Admissions Performance	Above 79%	ArtificialNeural Network (ANN)
StudentPerformancePredictionUsingMachineLearning	Upto79%	Decision Tree,ANN

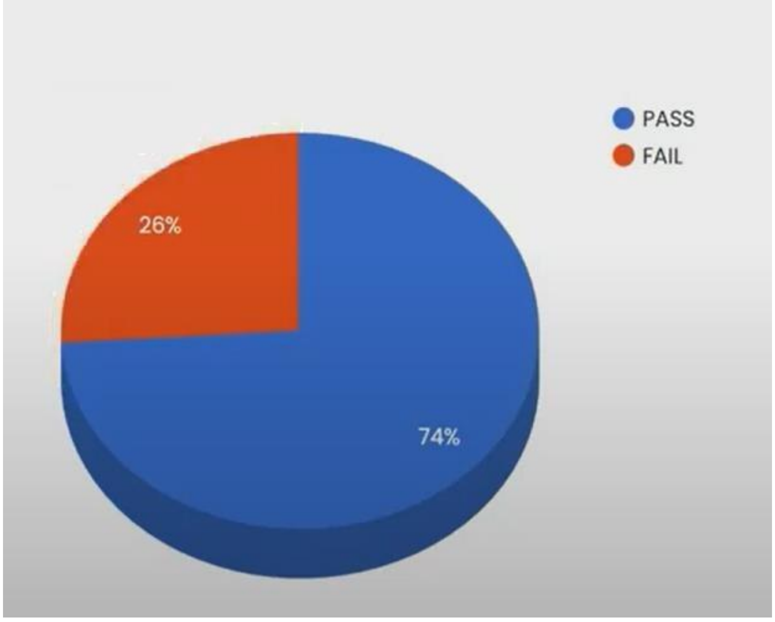


Figure5: Result analysis

6. CONCLUSION

The study confirms that machine learning can effectively predict student academic performance. While no strong correlation was found between high school math scores and university grades, urban students performed better overall. The approach achieved a prediction accuracy of 96.5%, proving the utility of these models for educational institutions. The study demonstrates the potential of ML models in predicting student performance. Decision Trees and SVM offer robust prediction capabilities and can be deployed by educational institutions to identify at-risk students early. Future work could integrate real-time data and expand the feature set for improved prediction.

7. FUTUREWORK

Diversified Data Integration
Incorporating various data sources such as **extracurricular participation**, **personal interests**, and **behavioral metrics** can offer a comprehensive understanding of a student's profile. This holistic approach enables the construction of more nuanced models that go beyond traditional academic indicators.

Real-Time Monitoring and Feedback
By implementing **real-time data tracking** and **automated feedback loops**, educational institutions can intervene early when students exhibit signs of academic struggle. Such timely actions can prevent issues from escalating and provide meaningful support to enhance student outcomes.

Integration with Learning Management Systems (LMS)
Embedding predictive models into existing LMS platforms streamlines **data collection**, **analysis**, and **reporting**, thereby improving the coordination between analytics tools and educational practitioners. This ensures that insights are both accessible and actionable for instructors and administrators.

Personalized Learning Pathways
Leveraging predictive analytics allows for the creation of **individualized learning trajectories** tailored to each student's strengths, weaknesses, and learning preferences. This personalization fosters greater engagement and helps students achieve academic success at their own pace.

Institutional Collaboration
The continuous evolution of predictive analytics in education depends on active **collaboration between universities, researchers, and edtech developers**. Sharing methodologies, tools, and findings can accelerate innovation and contribute to the establishment of best practices in the field.

REFERENCES

[1]Prediction of Students Performance using Machine learning[To cite this article: J. Dhilipan et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1055 012122].

[2]Analysis of machine learning strategies for prediction of passing undergraduate admission test.

https://doi.org/10.1016/j.jjimei.2022.100111

[3]Predicting Academic Success of College Students Using Machine Learning Techniques. Jorge HumbertoGuanin-Fajardo 1, Javier Guan˜a-Moya 2,* and Jorge Casillas

2024, 9, 60. https://doi.org/10.3390/data9040060

[4]Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems.Hanan Abdullah Mengash DOI 10.1109/ACCESS.2020.2981905, IEEE

[5]Development Roadmap for Launching Online Education: A Case Study of an Online Graduate Certificate Course",KevinKam Fung Yuen,AMYOoi Mei Wong

[6]M. Al-Saleem, N. Al-Kathiry, S. Al-Osimi, and G. Badr, "Mining Educational Data to Predict Students' Academic Performance," Springer, Cham, 2015, pp. 403–414