Machine Learning Approaches for Student Success Prediction: Evidence from a Higher Education Institution

Abdullah M. Alammari
Faculty of Education, Curriculums and Teaching Department,
Umm Al-Qura University, Makkah, Saudi Arabia
amammari@uqu.edu.sa

نماذج تعلم الآلة للتنبؤ بنجاح الطلاب الأكاديمي: دليل من مؤسسة للتعليم العالي

مستخلص البحث باللغة العربية

يستكشف هذا البحث تطبيق تعلم الآلة في التنبؤ بالنتائج الأكاديمية للطلاب، وذلك من خلال التطبيق على حزمة بيانات 236 طالبًا مسجلين في إحدى الجامعات السعودية خلال الفترة من 2021 إلى 2024. شملت حزمة البيانات المتكاملة العديد من المتغيرات، بما في ذلك الجنس، والسجل الأكاديمي (المعدل التراكمي السابق)، وأداءات التقييمات (الاختبارات القصيرة، والامتحانات النصفية، والواجبات، والامتحانات النهائية)، ومؤشرات التفاعل (ساعات الدراسة)، وأنماط الحضور. قام البحث بتقييم عدة نماذج من تقنيات تعلم الآلة، ومقارنة مدى فاعلية كل من الغابة العشوائية (Random Forest - RF) ، وآلات المتجهات الداعمة (Support Vector Machines – SVM)، والانحدار اللوجستي، وأقرب الجيران- k-Nearest Neighbors (kNN). وقد أظهر نموذج التعزيز التدريجي (Gradient Boosting) أداءً متفوقًا من خلال التحقق المتقاطع المكون من عشر طيات، حيث حقق درجة AUC بلغت 0.987. بينما قدمت نماذج الغابة العشوائية (RF) وأقرب الجيران (kNN) نتائج قوية أيضًا، كما سجل كل من الانحدار اللوجستي وآلات المتجهات الداعمة دقة تنبؤية أقل نسبيًا. كشفت تحليلات أهمية الميزات أن أداء الطلاب في الواجبات ودرجات الامتحانات النهائية كانا أهم العوامل المؤثرة في النجاح الأكاديمي. تقدم هذه النتائج رؤى عملية للمؤسسات التعليمية، مما يمكنها من تطوير تدخلات مستهدفة وتحسين بيئات التعلم بتقنيات تعلم الآلة لتعزيز التنبؤ بأداء الطلاب الأكاديمي ورفع معدلات نجاحهم.

الكلمات المفتاحية: نماذج تعلم الآلة, التنبؤ بنجاح الطلاب, التعليم الآلي.

Abstract

This research explores the application of machine learning in predicting student academic outcomes, analyzing data from 236 students enrolled at a Saudi university from 2021-2024. The comprehensive dataset encompassed multiple variables, including gender, academic history (previous GPA), assessment performance (quizzes, midterms, assignments, and final exams), engagement metrics (course hours), and attendance patterns. The study evaluated multiple machine learning approaches, comparing the effectiveness of Random Forest (RF), Support Vector Machines (SVM), Logistic Regression, and k-nearest Neighbors (kNN). Gradient Boosting demonstrated superior performance through tenfold cross-validation, achieving an AUC score of 0.987. While Random Forest and kNN also yielded strong results, Logistic Regression and SVM showed comparatively lower predictive accuracy. Feature importance analysis revealed that assignment performance and final examination scores were the most significant predictors of academic success. These findings provide educators with actionable insights to develop targeted interventions and optimize learning environments for enhanced student achievement according to the evidence shown in this study serving higher education institutions.

Keywords: Machine Learning Approaches, Student Success Prediction, Higher Education.

1. Introduction

There has been a notable surge in interest in applying data mining (DM) techniques within education in recent years. DM, fundamentally centered on the exploration of data, endeavors to uncover novel and potentially valuable insights or meaningful outcomes from extensive datasets (Witten et al., 2011). Its primary objective is to identify emerging trends and patterns from vast datasets using a variety of classification algorithms (Baker & Inventado, 2014). Educational data mining (EDM) involves adapting conventional data mining methodologies to address educational challenges (Fernandes et al., 2019). It involves applying DM techniques to educational datasets encompassing student particulars,

academic records, examination outcomes, class participation metrics, and the frequency of student inquiries. In recent times, EDM has emerged as a powerful tool for uncovering hidden patterns within educational data, predicting academic performance, and transforming the educational environment. The integration of EDM has endowed learning analytics with a newfound dimension (Waheed et al., 2020; Dolmark et al., 2022). Learning analytics involves a comprehensive approach to analyzing student data to gain a deep understanding of their learning environment. This analysis helps optimize learning and instructional effectiveness (Long & Siemens, 2011). It encompasses aggregating, quantifying, and disseminating data about students and their contextual milieu to comprehend and enhance learning experiences and the associated environments. Additionally, it involves the formulation of novel institutional strategies.

Learning analytics encompasses predicting student academic performance, unraveling patterns in system interactions and navigational behaviors, and identifying students at risk of academic underperformance (Waheed et al., 2020). Learning management systems (LMS), student information systems (SIS), intelligent teaching systems (ITS), Massive Open Online Courses (MOOCs), and other web-based educational platforms generate digital footprints ripe for analysis to discern behavioral trends among successful students and those at academic risk. By employing EDM techniques, these data can be leveraged to scrutinize the behaviors of successful students and those in jeopardy of failure, devise corrective measures informed by academic performance, and thereby aid educators in refining pedagogical methodologies (Kang et al. 2023; Casquero et al., 2016; Fidalgo-Blanco et al., 2015). The collation of educational process data offers novel avenues for enhancing the learning journey and optimizing user engagement with technological interfaces (Shorfuzzaman et al., 2019). The processing of educational data engenders enhancements across various domains, such as predictive modeling of student behavior, analytical learning methodologies, and formulation of novel educational policies (Capuano & Toti, 2019; Viberg et al., 2018). This holistic data assimilation furnishes education authorities with empirical foundations for policymaking and lays the groundwork for developing AI-infused learning platforms (Qahl & Sohaib, 2023).

EDM empowers educators to predict student attrition or declining course engagement scenarios, analyze internal factors affecting academic performance, and utilize statistical methods to forecast students' academic achievements. Diverse data mining techniques are employed to predict student performance, identify struggling learners, and anticipate dropout cases (Hardman et al., 2013; Kaur et al., 2015). Early prediction, a nascent phenomenon, encompasses evaluative methodologies to support students by proffering tailored corrective strategies and policies within this domain (Waheed et al., 2020). Especially amidst the backdrop of the pandemic, the rapid deployment of learning management systems has rendered them indispensable within higher education. As students engage with these systems, the generated log records have become increasingly accessible (Binsawad et al., 2022; Macfadyen & Dawson, 2010; Kotsiantis et al., 2013; Saqr et al., 2017). Universities are thus tasked with enhancing their capacity to harness these data reservoirs to foster student progression (Bernacki et al., 2020).

This study examines the use of machine learning (ML) models to predict student academic performance, a well-explored area in educational data mining (EDM) and learning analytics. It contributes by offering a detailed comparison of various ML models, including Random Forest (RF), Support Vector Machines (SVM), Logistic Regression, and k-nearest Neighbors (kNN) algorithms, and their effectiveness in predicting student grades within a specific institutional context. The various input features, such as gender, students' previous GPA, quiz scores, midterm grades, assignments, final exam results, the amount of time they spend on the course on the blackboard, their attendance records, and their final marks. By synthesizing these heterogeneous variables, the research provides a nuanced comprehension of the intricate determinants impinging upon students' scholastic achievements within the Saudi University context. To enhance the practical value and clarity of the research, the study is reframed around three distinct objectives, each addressing a critical aspect of the prediction process.

The study aims to provide educators and administrators with practical insights to design tailored interventions, allocate resources

wisely, and create an educational environment that fosters academic excellence and student success within the Saudi University ecosystem (Alammari et al., 2022).

1.1 Research Objectives and Research Questions:

Objective 1: Identify significant predictors of student academic performance through data-driven analysis.

Related Question 1: What are the key factors that influence student success in higher education?

This objective aims to uncover the most impactful features (e.g., assignment scores, attendance, and final exams) that contribute to student performance. Identifying these factors equips educators with insights to focus on areas that directly affect student outcomes, thereby enabling targeted interventions.

Objective 2: Evaluate and compare the performance of machine learning models to determine the most effective one for grade prediction.

Related Question 2: Which machine learning model demonstrates the highest accuracy in predicting student grades?

This objective focuses on assessing various algorithms, such as Gradient Boosting, Random Forest, and SVM, to identify the most reliable approach. Highlighting the effectiveness of the ML models to ensure that educators and institutions can adopt the best-performing tools for accurate predictions.

Objective 3: Demonstrate the practical application of the selected model, emphasizing ease of use and its integration into educational systems.

Related Question 3: How can the selected model be practically applied to enhance educational outcomes?

This objective addresses the application of the ML models in real-world educational settings. By integrating the model into platforms like learning management systems, educators can input student data to generate actionable insights, enabling early identification of at-risk students and fostering a data-driven approach to improving academic success.

2. Literature Review and Theoretical Perspective

Students' academic performance is paramount in higher education, prompting researchers to utilize Educational Data Mining (EDM) applications for predictive analysis and decision-making processes. The most common indicators for predicting and evaluating students' academic achievement at the university level include Cumulative Grade Point Average (CGPA) and Grade Point Average (GPA). Additionally, attributes such as quiz grades, midterm marks, assessments, attendance, and lab work have been employed. Some researchers have utilized students' academic achievements in previous courses to predict their performance in upcoming courses. Traditional ML techniques have been utilized to predict students' grades in forthcoming courses and identify at-risk students early in the semester.

Research on predicting students' academic performance has employed various machine-learning techniques to address diverse contexts. Al Mayahi et al. (2020) utilized Support Vector Classifiers and Naïve Bayes algorithms to predict students' grades in a mathematics course, demonstrating the effectiveness of these methods in educational settings. Similarly, Badr et al. (2016) explored grade prediction in programming courses, employing classification techniques based on association rule mining, highlighting the relevance of rule-based approaches for targeted educational domains. Extending the scope to deep learning, Nabil et al. (2021) proposed a framework using Deep Neural Networks to predict academic performance based on students' course grades, showcasing the potential of advanced neural architectures for capturing complex relationships in educational data.

In a series of studies, Akour et al. (2020), Sultana et al. (2019), Pujianto et al. (2020), Gajwani and Chakraborty (2021); Kumar et al. (2021) predicted student performance. Various ML algorithms were applied, including Convolutional Neural Network (CNN), Artificial Neural Networks (ANN), Decision Trees, Random Forest, and Naïve Bayes. Different ensemble methods and oversampling techniques were also explored to enhance model performance (Alharbi and Sohaib, 2021). CNN emerged as the most effective approach, outperforming other methods in accuracy without requiring manual feature selection. Malini (2021) utilized an online dataset from the machine learning repository to predict students' academic performance in high school. Features such as academic background, personal attributes, and economic background were analyzed, and ML techniques were employed, including bagging, ANNs, and boosting. The study underscored the significant impact of economic background on student performance. In studies (Kostopoulos et al. (2020), Olalekan et al. (2020), Aggarwal et al. (2021)), efforts were focused on identifying students at risk of failure early in the semester. ML techniques such as deep, dense neural networks, decision trees, K-nearest neighbor, random forest, and naive Bayes were employed. It was observed that non-academic parameters significantly influenced student performance, with results improved when considering both academic and non-academic parameters.

Agaoglu (2016) employed four classification techniques to forecast instructors' performance based on students' evaluations of courses. Qiu et al. (2018) introduced an integrated framework named Feature Selection Prediction to anticipate dropout rates in Massive Open Online Courses (MOOCs), encompassing feature generation, feature selection, and dropout prediction. Akram et al. (2019) presented the Students' Academic Performance Enhancement through Homework Late/Non-submission Detection (SAPE) algorithm for predicting students' academic performance. Considering MOOC learning characteristics, Wen et al. (2020) proposed a simplified feature matrix to retain information regarding the local correlation of learning behavior and introduced a new Convolutional Neural Network (CNN) model to forecast dropout rates. Lin et al. (2020) devised a method for continuous facial emotion pattern recognition using deep learning, amalgamating Convolutional Neural

Networks (CNNs) and Long Short-Term Memory (LSTM) networks to analyze students' continuous facial expressions and predict academic emotions. Farissi et al. (2020) suggested integrating genetic algorithm feature selection with classification methods to anticipate student academic performance. Turabieh et al. (2021) proposed a modified version of the Harris Hawks Optimization (HHO) algorithm to enhance feature selection for predicting student performance by regulating population diversity. Ma et al. (2021) introduced a novel approach called progressive imitation learning to train a lightweight CNN model by mimicking the learning trajectory of a teacher model for constructing a prediction model. Lastly, Gao et al. (2022) proposed a deep cognitive diagnosis framework to assess students' mastery of skills and problem-solving abilities, enhancing traditional cognitive diagnosis methods with deep learning techniques.

Asif et al. (2017) delved into the performance of undergraduate students using Data Mining (DM) methods. Their study aimed to predict academic achievements at the culmination of a four-year program while examining students' developmental trajectories. Notably, they emphasized the importance of identifying specific courses indicative of exceptional or poor performance to provide timely support to struggling students and foster opportunities for high achievers. Cruz-Jesus et al. (2020) undertook the prediction of student academic performance by considering a range of demographic variables. Employing machine learning techniques such as Random Forest, Logistic Regression, k-nearest Neighbors, and Support Vector Machines, they achieved prediction accuracies varying from 50% to 81%. Fernandes et al. (2019) developed a model integrating demographic characteristics and achievement grades to forecast academic success. Their findings highlighted the significance of previous achievement scores and attendance in estimating academic performance.

Hoffait and Schyns (2017) utilized DM methods to identify students at risk of failure, leveraging registration data and environmental factors. Their approach facilitated precise classification and ranking of students based on their risk levels. Rebai et al. (2020) proposed a machine-learning model to elucidate key factors influencing school performance, emphasizing school size, competition, and parental pressure. Ahmad and

Shahzadi (2018) employed machine learning techniques to identify academically at-risk students based on learning skills and study habits, achieving an accuracy of 85%. Musso et al. (2020) predicted academic performance and dropouts using learning strategies and socio-demographic factors, underlining the influence of background information. Waheed et al. (2020) designed a model employing artificial neural networks to analyze students' navigation through Learning Management Systems (LMS), identifying significant impacts of demographics and clickstream activities on student performance.

Similarly, Xu et al. (2019) explored the relationship between internet usage behaviors and academic performance, achieving high prediction accuracy. Bernacki et al. (2020) investigated the predictive power of log records in LMS, successfully identifying students needing course repeats and potential future failures. Burgos et al. (2018) predicted subsequent semester achievement grades and developed interventions for at-risk students, resulting in decreased dropout rates.

Collectively, these studies underscore the growing reliance on machine learning and deep learning models for enhancing academic performance prediction and personalized learning interventions. These studies underscore the multifaceted nature of factors influencing student academic performance. Researchers aim to develop robust predictive models that inform effective educational interventions and policies by employing machine learning algorithms and considering diverse variables. Overall, the literature underscores the significance of EDM in predicting academic performance, guiding targeted interventions, and informing educational policies.

To strengthen the educational focus of this study, it is imperative to delve deeper into the applications of learning analytics (LA) within educational contexts. Learning analytics represents the intersection of educational research, data analysis, and technology, providing a powerful means to enhance teaching and learning processes. This addition shifts the study’s perspective from a predominantly computational lens to one that resonates more with the educational community.

2.1 Learning Analytics in Education

Learning analytics (LA) refers to the measurement, collection, analysis, and reporting of data concerning learners and their contexts to understand and enhance learning as well as the environments in which it takes place (Siemens & Long, 2011). It includes a broad range of applications, such as predicting student performance, personalizing learning pathways, identifying at-risk students, and improving instructional design. LA can predict academic success or failure by analyzing historical data and real-time student behaviors. Tools like dashboards enable educators to monitor student progress and intervene early for those at risk (Waheed et al., 2020; Viberg et al., 2018). For example, early warning systems driven by LA can alert instructors to students with declining engagement or poor academic performance, prompting timely interventions (Macfadyen & Dawson, 2010).

In addition, By analyzing student interaction patterns, LA enables the customization of learning materials to suit individual needs. This approach has been shown to improve learner engagement and outcomes (Shorfuzzaman et al., 2019). Adaptive learning platforms leverage LA to recommend tailored activities and resources, accommodating diverse learning styles and paces (Fidalgo-Blanco et al., 2015).

Moreover, Data-driven insights derived from LA can inform curriculum revisions and teaching methodologies. For instance, identifying topics where students struggle most allows educators to adjust course content or provide additional resources (Capuano & Toti, 2019). LA also facilitates the alignment of instructional strategies with student needs, creating more inclusive and effective learning environments (Fernandes et al., 2019).

Also, Learning analytics systems analyze participation metrics, such as forum interactions or resource usage, to gauge engagement levels. Research has shown that higher engagement correlates strongly with academic success (Viberg et al., 2018). Platforms like LMS (Learning Management Systems) use these insights to recommend peer collaborations or suggest interactive activities to re-engage students.

Learning analytics can enhance formative assessment by tracking student progress and providing detailed feedback. Visualization tools, such as heatmaps and progression graphs, help both students and instructors understand learning trajectories (Kang et al., 2023). Automated

feedback systems powered by LA ensure timely, specific, and actionable insights for learners.

While this study leverages machine learning methods—an essential aspect of EDM—to predict student outcomes, integrating learning analytics applications enriches the theoretical foundation. Unlike EDM, which often focuses on algorithms and techniques, LA emphasizes actionable insights for educational practice (Baker & Yacef, 2009). The synergy between these domains enhances the capacity of educators and administrators to foster academic success through evidence-based strategies.

3. Method

This study collected data from the students' records at Umm Al-Qura University in Saudi Arabia. The dataset comprises information from 236 students who successfully completed the course between 2021 and 2024. These records encompass a wide range of details, including student demographics, academic achievements, and course-related data such as gender, previous GPA, quiz scores, midterm exam grades, assignment scores, final exam grades, hours spent in the course recorded on the blackboard, attendance records, final marks, and the target variable, grade. Table 1 shows the data information.

Table 1: Variables used in the study

Features/Target	Type	Descriptions/Values
Student ID	Numeric	Unique identifier for each student. Meta attribute.
Gender	Catogorical	Gender of the student.
Previous GPA	Numeric	Grade Point Average from the previous semester term.
Quiz	Numeric	Score on quizzes during the course.
Midterm	Numeric	Score on the midterm exam.
Assignment	Numeric	Score on assignment.
Final Exam	Numeric	The score obtained in the final exam.
Hours Spent in Course	Numeric	Number of hours spent in the course recorded on the blackboard.
Attendance	Numeric	Attendance percentage in the course.
Final Marks	Numeric	Final total marks out of 100
Grade (Target)	Catogorial	The corresponding grade symbol based on the GPA.

Features/Target	Type	Descriptions/Values
		● Excellent: 3.50 – 4.00 GPA ● Very Good: 2.75 to less than 3.50 GPA ● Good: 1.75 to less than 2.75 GPA ● Pass: 1.00 to less than 1.75 GPA

3.1 Machine Learning Models

To predict students' academic performance, Random Forest (RF), Support Vector Machines (SVM), Logistic Regression and k-nearest Neighbors (kNN) algorithms were deployed. Prediction accuracy was evaluated using tenfold cross-validation. The Data Mining (DM) process serves two purposes: predictive and descriptive modeling. Predictive models leverage known data to predict outcomes for unknown datasets, while descriptive models identify patterns to inform decision-making. In prediction analysis, machine learning techniques such as support vector machines, kNN, decision trees, and random forests offer greater efficiency and accuracy for forecasting purposes (Huang & Fang, 2013; Delen, 2010; Hani et al., 2024; Nassar and Sohaib, 2024). Statistical techniques aim to construct predictive models based on available input data, whereas machine learning methods automatically generate models matching input data with expected target values. Model performance was assessed using confusion matrix metrics.

3.2 Experiment

The experimental phase was conducted using the Orange machine learning software, renowned for its user-friendly interface and powerful capabilities in data mining. Orange provides a component-based programming environment suitable for both expert data scientists and beginners in data science. Its workflow-based approach allows users to stack widgets for various data analysis tasks, including retrieval, preprocessing, visualization, modeling, and evaluation.

A workflow in Orange represents a sequence of actions to achieve a specific task, facilitating comprehensive data analysis by combining different components. Figure 1 illustrates the workflow diagram designed for this study. Table 2 describes the parameters used in this study. Model performance was assessed using various metrics, including confusion

matrix, classification accuracy (CA), precision, recall, F1-score, and area under the ROC curve (AUC). Table 3 describes the metrics.

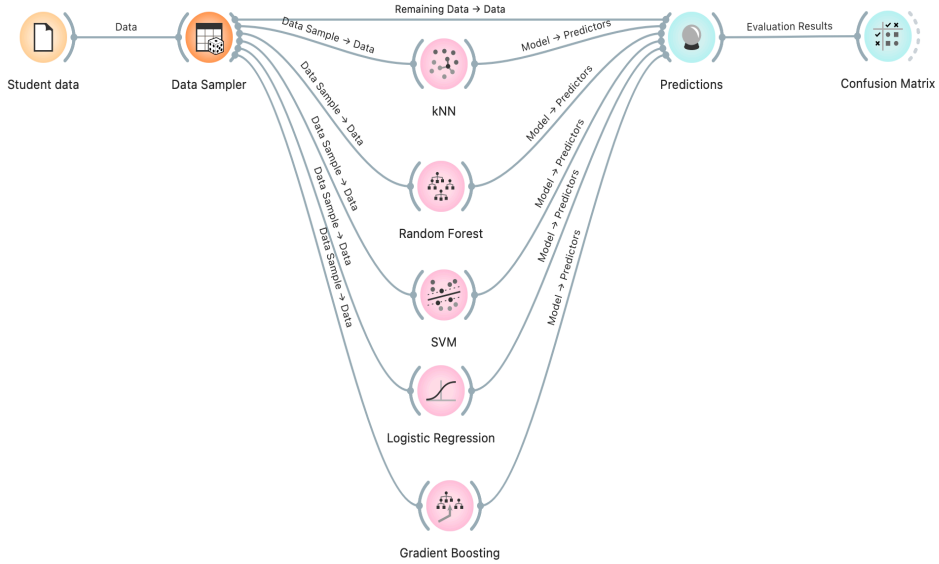

Figure 1: ML workflow

Table 2: ML parameters

ML Method	Parameters
Random Forest (RF)	- Number of trees <10> - Criterion (entropy) - Minimum samples split/leaf <5>
Support Vector Machines (SVM)	- Kernel type (radial basis function)- Regularization parameter (C)= 1.8
Logistic Regression	- Penalty (L2) - Regularization parameter (C)= 1
Gradient Boosting	- Number of trees <100> - Learning rate <0.1> - Maximum depth <5>
k-Nearest Neighbors (kNN)	- Number of neighbors <5> - Distance metric (Euclidean)

Table 3: Metrics descriptions

Metric	Description

AUC	Area Under the ROC Curve: A measure of the model's ability to distinguish between classes. A higher AUC indicates better performance, with a maximum value of 1 indicating perfect classification.
CA	Classification Accuracy: The proportion of correctly classified instances out of the total instances. It represents the overall correctness of the model's predictions.
F1 Score	The harmonic means of precision and recall. It considers both false positives and false negatives and is useful when the classes are imbalanced.
Precision	The proportion of true positive predictions out of all positive predictions. It measures the accuracy of positive predictions.
Recall (Sensitivity)	The proportion of true positive predictions out of all actual positive instances. It measures the ability of the model to correctly identify positive instances.
MCC	Matthews Correlation Coefficient: A correlation coefficient between the observed and predicted binary classifications. It ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement.

4. Results

The evaluation results depict the performance of various machine learning models in predicting grades based on features such as gender, previous GPA, quiz scores, midterm scores, assignment scores, final exam scores, hours spent in the course on the blackboard, and attendance. Table 4 shows the results. Among the models assessed, Gradient Boosting emerges as the most effective, achieving the highest AUC score of 0.987 and consistently strong performance across all metrics. Following closely, Random Forest demonstrates robust performance, albeit with slightly lower scores than Gradient Boosting. kNN also shows competitive performance, although it slightly trails behind the top-performing models. Conversely, Logistic Regression delivers moderate performance, while SVM exhibits the lowest overall performance among the models evaluated. These results suggest that Gradient Boosting is the most reliable model for accurately predicting grades based on the given features, with the potential for further optimization to enhance the performance of other models like Logistic Regression and SVM.

Table 4: Target class (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Gradient Boosting	0.987	0.985	0.985	0.986	0.985	0.980
Random Forest	0.986	0.956	0.955	0.956	0.956	0.939
kNN	0.986	0.941	0.940	0.942	0.941	0.919
Logistic Regression	0.982	0.933	0.933	0.938	0.933	0.910
SVM	0.979	0.896	0.888	0.903	0.896	0.859

Table 5 confusion matrices, which show the proportion of predicted grades compared to the actual grades, help evaluate the performance of the models in accurately classifying students into their respective grade categories.

Table 5: Confusion matrix (showing the proportion of predicted)

Model	Excellent	Fail	Good	Pass	Very Good	Total
Random Forest	96.1%	0.0%	0.0%	0.0%	0.0%	49
	0.0%	100.0%	0.0%	0.0%	0.0%	5
	0.0%	0.0%	92.3%	0.0%	5.1%	26
	0.0%	0.0%	7.7%	100.0%	0.0%	16
	3.9%	0.0%	0.0%	0.0%	94.9%	39
SVM	89.1%	0.0%	0.0%	0.0%	0.0%	49
	0.0%	100.0%	0.0%	22.2%	0.0%	5
	1.8%	0.0%	92.0%	0.0%	5.6%	26
	0.0%	0.0%	8.0%	77.8%	0.0%	16
	9.1%	0.0%	0.0%	0.0%	94.4%	39
kNN	94.2%	0.0%	0.0%	0.0%	0.0%	49
	0.0%	100.0%	0.0%	0.0%	0.0%	5
	0.0%	0.0%	88.9%	0.0%	5.3%	26
	0.0%	0.0%	11.1%	100.0%	0.0%	16
	5.8%	0.0%	0.0%	0.0%	94.7%	39
Logistic Regression	96.1%	0.0%	0.0%	0.0%	0.0%	49
	0.0%	71.4%	0.0%	0.0%	0.0%	5
	0.0%	0.0%	88.9%	0.0%	5.4%	26
	0.0%	28.6%	3.7%	100.0%	0.0%	16
	3.9%	0.0%	7.4%	0.0%	94.6%	39

Gradient Boosting	96.1%	0.0%	0.0%	0.0%	0.0%	49
	0.0%	100.0%	0.0%	0.0%	0.0%	5
	0.0%	0.0%	100.0%	0.0%	0.0%	26
	0.0%	0.0%	0.0%	100.0%	0.0%	16
	3.9%	0.0%	0.0%	0.0%	100.0%	39

The confusion matrices comprehensively summarize each model's performance in predicting grade categories.

- Random Forest: This model demonstrated exceptional performance across all grade categories, achieving high accuracy rates for each grade. Notably, it attained perfect classification accuracy for "Fail," "Pass," and "Very Good" grades.
- SVM (Support Vector Machine): SVM also performed well, particularly excelling in accurately classifying "Excellent" and "Very Good" grades. However, it faced challenges in accurately predicting "Fail" and "Good" grades, resulting in lower accuracy percentages for those categories.
- kNN (k-Nearest Neighbors): Similar to Random Forest, kNN achieved high accuracy across all grade categories. It notably performed excellently in classifying "Excellent" and "Very Good" grades without misclassifications.
- Logistic Regression: While Logistic Regression demonstrated strong performance in classifying "Excellent" and "Very Good" grades, it encountered difficulties in accurately predicting "Fail" and "Pass" grades. This led to comparatively lower accuracy percentages in those categories.
- Gradient Boosting: Like Random Forest and kNN, Gradient Boosting achieved high accuracy across all grade categories. It particularly excelled in accurately classifying "Excellent" and "Very Good" grades.

Overall, Random Forest and Gradient Boosting showcased the highest classification accuracy across all grade categories, while Logistic Regression and SVM exhibited slightly lower accuracy in specific categories. Notably, the "Fail" category generally had lower accuracy

across all models, suggesting potential areas for improvement in predicting this grade category.

Table 6 analysis also reveals which factors are most important for predicting final grades in this study. We used three methods to rank these factors: Information Gain, Gain Ratio, and Gini Index.

Table 6: Feature Ranking

	Features	Info. gain	Gain ratio	Gini
1	Assignment	1.137	0.569	0.332
2	Final Marks	1.133	0.568	0.331
3	Previous GPA	1.101	0.551	0.313
4	Final Exam	1.096	0.551	0.339
5	Midterm	1.051	0.529	0.318
6	Quiz	1.044	0.532	0.306
7	Attendance	0.727	0.375	0.222
8	Hours Spent in Course on Blackboard	0.647	0.327	0.180

- **Information Gain:** This method examines how much each factor reduces the overall uncertainty about final grades. Here, "Assignment" and "Final Marks" showed the most significant reduction in uncertainty, suggesting they're the most informative for predicting final grades.
- **Gain Ratio:** This method considers the information gain while accounting for the number of categories within a factor. Similar to Information Gain, "Assignment" and "Final Marks" came out on top, solidifying their importance.
- **Gini Index:** This method focuses on how well each factor separates students into different final grade categories. "Attendance" and "Hours Spent in Course" had the lowest values, indicating they were most effective in separating students based on final grades.

Overall, "Assignment" and "Final Marks" consistently ranked high across all three methods, highlighting their significant role in predicting final grades. "Previous GPA" and "Final Exam" also scored well, suggesting their importance. Interestingly, "Attendance" and "Hours Spent in Course" ranked lower, implying they might not be as influential in predicting final grades as other factors.

5. Discussions

The evaluation results demonstrate the effectiveness of various machine learning models in predicting final grades based on several

features, including gender, previous GPA, quiz scores, midterm scores, assignment scores, final exam scores, hours spent in the course on the blackboard, and attendance. These results indicate that Gradient Boosting is the most reliable model for accurately predicting grades based on the given features, with the potential for further optimization to enhance the performance of other models like Logistic Regression and SVM.

The confusion matrices provide a detailed breakdown of each model's performance in accurately classifying students into their respective grade categories. Random Forest and Gradient Boosting consistently demonstrate high accuracy rates across all grade categories, with Random Forest achieving perfect classification accuracy for "Fail," "Pass," and "Very Good" grades. SVM also performs well, particularly excelling in accurately classifying "Excellent" and "Very Good" grades, although it faces challenges with "Fail" and "Good" grades. kNN showcases strong performance with no misclassifications in "Excellent" and "Very Good" grades. Logistic Regression, while showing robust classification for "Excellent" and "Very Good" grades, struggles with "Fail" and "Pass" grades, resulting in lower accuracy percentages in those categories.

The feature ranking analysis using the Information Gain, Gain Ratio, and Gini Index highlights the importance of specific features in predicting final grades. "Assignment" and "Final Marks" consistently rank high across all three methods, indicating their significant contribution to predicting final grades. "Previous GPA" and "Final Exam" also emerge as essential features, while "Attendance" and "Hours Spent in Course" rank lower, suggesting they may not be as influential in predicting final grades compared to other factors.

The results suggest that Gradient Boosting is the most effective model for predicting final grades based on the given features, with "Assignment" and "Final Marks" being crucial predictors. Further optimization of other models, like Logistic Regression and SVM, could enhance their performance. Additionally, while certain features like "Attendance" and "Hours Spent in Course on the Blackboard" may not play a significant role in predicting final grades, further investigation into

their impact may provide insights for improvement. However, achieving accurate predictions of student academic performance hinges upon thoroughly comprehending the factors and variables influencing students' outcomes and accomplishments (Dolmark et al., 2021; Alshanqiti & Namoun, 2020).

5.1 Conclusion and Future Work

The study shows that machine learning can be a powerful tool for predicting student grades, with Gradient Boosting leading the pack in this study. However, other models like Random Forest and KNN were also very effective. The best choice depends on the specific needs, such as prioritizing identifying failing students. It is important to remember that these models rely on the quality of their training data. Biases or missing information in the data can lead to biased predictions. These models are for informational purposes and should not solely determine student grades. However, they can be precious for educators by flagging students who might be struggling or could benefit from extra help. Future research can focus on improving these models by including more data sources or exploring more advanced machine-learning techniques.

References

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. IEEE Access, 4, 2379–2387.

Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting Student performance using ensemble learning techniques. International Journal of Systems Dynamics and Applications, 10(3), 38–49.

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. Bulletin of Education and Research, 40(3), 157–164.

Alharbi A. and Sohaib, O. (2021). Technology Readiness and Cryptocurrency Adoption: PLS-SEM and Deep Learning Neural

Network Analysis," in IEEE Access, vol. 9, pp. 21388-21394, doi: 10.1109/ACCESS.2021.3055785

Alammari, A., Sohaib, O. and Younes, S., (2022). Developing and evaluating cybersecurity competencies for students in computing programs. PeerJ Computer Science, 8, p.e827.

Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE Access, 8, 203827–203844. https://doi.org/10.1109/access.2020.3036572

Akour, M., Sghaier, H. A., & Qasem, O. A. (2020). The effectiveness of using deep learning algorithms in predicting students' achievements. Indonesian Journal of Electrical Engineering and Computer Science, 19, 388–394.

Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., & Jiang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. IEEE Access, 7, 102487–102498.

Al Mayahi, K., & Al-Bahri, M. (2020). Machine learning based predicting student academic success. In Proceedings of the 12th International Congress on Ultra Modern Telecommunications and Control Systems Workshops (ICUMT) (pp. 264–268).

Algobail, A., Badr, G., Almutairi, H., & Almutery, M. (2016). Predicting students' performance in university courses: A case study and tool in KSU mathematics department. Procedia Computer Science, 82, 80–89.

Ali, J., Anzer, A., & Tabaza, H. A. (2018). Predicting academic performance of students in UAE using data mining techniques. In Proceedings of the International Conference on Advanced Computing and Communication Engineering (ICACCE) (pp. 179–183).

Anzer, A., Tabaza, H. A., & Ali, J. (2018). Predicting academic performance of students in UAE using data mining techniques. In Proceedings of the International Conference on Advanced Computing and Communication Engineering (ICACCE) (pp. 179–183).

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers and Education, 113, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. Learning analytics (pp. 61–75). Springer.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educa- tional Data Mining, 1(1), 3–17.

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158(August), 103999. https://doi.org/10.1016/j.compedu.2020.103999

Binsawad, M., Abbasi, G. A., & Sohaib, O. (2022). People's expectations and experiences of big data collection in the Saudi context. PeerJ Computer Science, 8, e926.

Burgos, C., Campanario, M. L., De, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers and Electrical Engineering, 66(2018), 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Capuano, N., & Toti, D. (2019). Experimentation of a smart learning system for law based on knowledge discovery and cognitive

computing. Computers in Human Behavior, 92, 459–467. https://doi.org/10.1016/j.chb.2018.03.034

Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: A case study in higher education using learning analytics approach. Interactive Learning Environments, 24(1), 49–67. https://doi.org/10.1080/10494820.2013.817441

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intel- ligence methods to assess academic achievement in public high schools of a European Union country. Heliyon. https://doi.org/10.1016/j.heliyon.2020.e04081

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Dolmark, T., Sohaib, O., Beydoun, G., Wu, K., & Taghikhah, F. (2022). The effect of technology readiness on individual absorptive capacity toward learning behavior in Australian universities. Journal of Global Information Management (JGIM), 30(1), 1-21.

Dolmark, T., Sohaib, O., Beydoun, G. and Wu, K., (2021). The effect of individual's technological belief and usage on their absorptive capacity towards their learning behaviour in learning environment. Sustainability, 13(2), p.718.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining : Predic- tive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94(February 2018), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Farissi, A., Dahlan, H. M., & Samsuryadi. (2020). Genetic algorithm based feature selection with ensemble methods for student academic performance prediction. Journal of Physics: Conference Series, 1500, Article no. 012110.

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using Learning Analytics to improve teamwork assessment. Computers in Human Behavior, 47, 149–156. https://doi.org/10.1016/j.chb.2014.11.050

Gajwani, J., & Chakraborty, P. (2021). Students' performance prediction using feature selection and supervised machine learning algorithms. In Proceedings of the International Conference on Innovative Computing and Communications (pp. 347–354).

Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. Future Generation Computer Systems, 126, 252–262.

Hani, U., Sohaib, O., Khan, K., Aleidi, A., & Islam, N. (2024). Psychological profiling of hackers via machine learning toward sustainable cybersecurity. Frontiers in Computer Science, 6, 1381351.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. Systems Research and Behavioral Science, 30(2), 194–203. https://doi.org/10.1002/sres.2130

Hoffait, A., & Schyns, M. (2017). Early detection of university students with potential difficulties. Decision Support Systems, 101(2017), 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A compari- son of four types of predictive mathematical models. Computers and Education, 61(1), 133–145. https://doi.org/10. 1016/j.compedu.2012.08.015

Kang, K., Li, L., & Sohaib, O. (2023). Graduates' intention to develop live commerce: The educational background perspective using multi-group analysis. Entrepreneurial Business and Economics Review, 11(1), 113-126.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learn- ers in education sector. Procedia Computer Science, 57, 500–508. https://doi.org/10.1016/j.procs.2015.07.372

Kostopoulos, G., Tsiakmaki, M., Kotsiantis, S., & Ragos, O. (2020). Deep dense neural network for early prediction of failure-prone students. In G. A. Tsihrintzis & L. C. Jain (Eds.), Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications (pp. 291–306). Springer.

Kostopoulos, G., Tsiakmaki, M., Koutsonikos, G., Pierrakeas, C., Kotsiantis, S., & Ragos, O. (2018). Predicting university students' grades based on previous academic achievements. In Proceedings of the 9th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1–6).

Kotsiantis, S., Tselios, N., Filippidi, A., & Komis, V. (2013). Using learning analytics to identify successful learners in a blended learning course. International Journal of Technology Enhanced Learning, 5(2), 133–150. https://doi.org/10.1504/IJTEL. 2013.059088

Kumar, M., Mehta, G., Nayar, N., & Sharma, A. (2021). EMT: Ensemble meta-based tree model for predicting student performance in academics. IOP Conference Series: Materials Science and Engineering, 1022(1), Article no. 012062.

Lin, S. Y., Wu, C. M., Chen, S. L., & Lin, T. L. (2020). Continuous facial emotion recognition method based on deep learning of academic emotions. Sensors and Materials, 32(10), 3243–3259.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. Educause Review, 46(5), 31–40. Macfadyen, Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. Higher Education, 80(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Malini, J. (2021). Analysis of factors affecting student performance evaluation using education data mining technique. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12, 2413–2424.

Ma, H. B., Yang, S. Y., Feng, D. Z., & Jiao, L. C. (2021). Progressive mimic learning: A new perspective to train lightweight CNN models. Neurocomputing, 456, 220–231.

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9, 140731–140746.

Nassar, R.U.D. and Sohaib, O., 2024, April. Prediction of the Compressive Strength of Sustainable Concrete Produced with Powder Glass Using Standalone and Stack Machine Learning Methods. In Asian Conference on Intelligent Information and Database Systems (pp. 147-158). Singapore: Springer Nature Singapore.

Olalekan, A. M., Egwuche, O. S., & Olatunji, S. O. (2020). Performance evaluation of machine learning techniques for prediction of graduating students in tertiary institution. In Proceedings of the International Conference on Mathematics, Computer Engineering and Computer Science (ICMCECS) (pp. 1–7).

Pujianto, U., Prasetyo, W. A., & Taufani, A. R. (2020). Students academic performance prediction with k-nearest neighbor and C4.5 on SMOTE-balanced data. In Proceedings of the 3rd International

Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 348–353).

Qahl, M., & Sohaib, O. (2023). Key factors for a creative environment in Saudi Arabian higher education institutions. Journal of Information Technology Education: Innovations in Practice, 22, 001-048.

Qiu, L., Liu, Y., & Liu, Y. (2018). An integrated framework with feature selection for dropout prediction in massive open online courses. IEEE Access, 6, 71474–71484

Rebai, S., Ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70(August 2018), 100724. https://doi.org/10.1016/j.seps. 2019.06.009

Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. Medical Teacher, 39(7), 757–767. https://doi.org/10.1080/0142159X.2017.1309376

Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., & Alamri, A. (2019). Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. Computers in Human Behavior, 92(Febru- ary 2017), 578–588. https://doi.org/10.1016/j.chb.2018.07.002

Sultana, J., Usha, M., & Farquad, R. (2019). An efficient deep learning method to predict student's performance. Higher Education Quality Assurance Enhancement, Technical Report.

Turabieh, H., Azwari, S. A., Rokaya, M., Alosaimi, W., Alharbi, A., Alhakami, W., & Aln ai, M. (2021). Enhanced Harris hawks optimization as a feature selection for the prediction of student performance. Computing, 7, 1417–1438.

Tsiakmaki, M., Kostopoulos, G., Koutsonikos, G., Pierrakeas, C., Kotsiantis, S., & Ragos, O. (2018). Predicting university students' grades based on previous academic achievements. In Proceedings of the 9th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1–6).

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. Computers in Human Behavior, 89(July), 98–110. https://doi.org/10.1016/j.chb.2018.07.027

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104(October 2019), 106189. https://doi.org/10.1016/j.chb.2019.106189

Wen, Y., Tian, Y., Wen, B., Zhou, Q., Cai, G., & Liu, S. (2020). Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. Tsinghua Science and Technology, 25(3), 336–347.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98(January), 166–173. https://doi.org/10.1016/j. chb.2019.04.015