Physical Education, Health and Social Sciences

https://journal-of-social-education.org E-ISSN: 2958-5996

P-ISSN: 2958-5988

Student Grade Prediction Using Machine Learning

Abdullah Malik[1], Noreen Akbar[2], Usman Khan[3], Riaz Ahmad[4], Ameer Mustafa[5]

1 Department of Computer Science, GDC, Hayatabad, Peshawar.
Email: abdullahmalik3483@gmail.com

2 Department of Computer Science, Shaheed Benazir Bhutto Women's University, Peshawar.
Email: noreenakbar06@gmail.com

3 Department of Computer Science, Higher Education Department, KP, Pakistan.
Email: usmankhan@hed.gkp.pk

4 Department of Computer Science, Higher Education Department, KP, Pakistan.
Email: riazahmad@hed.gkp.pk

5 Department of Computer Science, Higher Education Department, KP, Pakistan.
Email: ameer983465@gmail.com

DOI: https://doi.org/10.63163/jpehss.v3i4.848

Abstract

In recent years, predictive analytics has become an essential part of higher education institutions to improve academic decision-making and student performance assessment. Machine learning techniques play a key role in predicting students' final grades by analyzing various educational and behavioral factors. This study develops a predictive framework using the publicly available Kaggle Student-Mat dataset to forecast student grades with increased accuracy and reliability. Three machine learning algorithms—Linear Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN)—are implemented and compared based on their predictive performance. The dataset undergoes preprocessing, normalization, and feature selection to ensure model robustness. The experimental results show that Linear Regression outperforms the other models in accuracy and generalization, with Linear Regression and KNN following closely. The findings demonstrate that machine learning-based predictive models can provide valuable insights for educators and institutions to identify at-risk students early and improve educational outcomes.

Keywords: Machine Learning, Predictive Model, Student Grade Prediction, Educational Data Mining, Performance Evaluation.

I.Introduction

In higher education institutions (HEIs), academic management systems store valuable data about students' academic performance, attendance, and behavioral patterns. These records include detailed information such as course grades, exam results, and demographic factors, which can be used to analyze and predict student success trends. Predicting student academic outcomes is increasingly seen as a key part of educational analytics because it helps educators identify at-risk students and implement timely interventions to improve learning outcomes [1]. Over the past decade, the use of predictive analytics in education has grown considerably, combining data mining and machine learning techniques to uncover hidden patterns and generate actionable insights from large datasets [2]. These methods have been successfully applied to various

educational challenges such as forecasting student performance [3], predicting dropout rates [4], developing academic warning systems [5], and recommending learning paths [6]. Among these, student grade prediction remains a primary focus since it directly measures academic achievement and learning effectiveness [7]. However, creating accurate predictive models for student grades poses several challenges. Educational data often contains noise, missing values, and variations in student behavior, which can affect model performance [8]. Additionally, many real-world student datasets suffer from an imbalance in grade distribution, where most students receive average grades while fewer fall into the highest or lowest categories [9]. This imbalance can cause models to favor predicting the dominant classes, decreasing prediction reliability.

To tackle these challenges, this study compares three machine learning algorithms—Linear Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—to predict final student grades using the Kaggle Student-Mat dataset. The dataset includes demographic, social, and academic attributes of secondary education students, offering a comprehensive basis for analysis. The main research questions in this study are as follows:

RQ1: Among the predictive models—Linear Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—which demonstrates the highest accuracy in forecasting student grades utilizing the Kaggle Student-Mat dataset?

RQ2: In what ways can data preprocessing and normalization improve prediction performance and reduce bias within the selected machine learning models?

To explore these questions, data preprocessing, feature scaling, and performance evaluation are conducted to ensure model reliability. Comparative experiments are performed to analyze the predictive capability of each algorithm in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score.

The main contributions of this paper are summarized as follows:

- We conduct a comparative evaluation of three widely used machine learning algorithms—Linear Regression, SVM, and KNN—on the Student-Mat dataset to forecast students' final grades.
- We examine how preprocessing and normalization techniques affect the prediction performance of the chosen models.
- The results show that SVM offers the most accurate and consistent prediction performance among the tested models, confirming its reliability for educational data prediction tasks.

This paper is organized as follows: Section II presents related work on student performance prediction using machine learning. Section III explains the methodology, including dataset description and model implementation. Section IV discusses the results and findings. Finally, Section V concludes the paper and provides directions for future research.

II. Related Works

Multiple investigations within higher education institutions (HEIs) have been undertaken to forecast student performance utilizing diverse machine learning (ML) methodologies. These investigations generally entail the analytical processing of numerous attributes derived from multiple sources for student-grade prediction across various educational environments. Nevertheless, the efficacy of predictive models when confronted with imbalanced datasets in educational contexts remains infrequently addressed. For instance, a study by Sandra et al. (2021) employed discretization techniques and the SMOTE oversampling method to enhance the accuracy of student final-grade predictions. They utilized classification algorithms such as Naïve Bayes (NB), Decision Tree (DT), and Neural Network (NN) to categorize students' final grades into five distinct classes (A, B, C, D, F). Their findings revealed that NN and NB, when used in conjunction with SMOTE and optimal binning, achieved an accuracy of approximately 75%, with

NB demonstrating superior computational efficiency relative to NN. [10] Another study conducted at the University of Minnesota devised a method for predicting prospective course grades among students enrolled in the CSE and ECE programs; this research compared Matrix Factorization (MF) and Linear Regression (LinReg), concluding that both methods yielded more precise predictions than traditional techniques, and that employing a course-specific subset of data further enhanced accuracy. In a subsequent investigation [11], researchers applied MF, Collaborative Filtering (CF), and Restricted Boltzmann Machines (RBM) to real undergraduate student data ($n \approx 225$) to forecast grades across different courses. The results indicated that CF encountered sparsity issues within the dataset, whereas RBM delivered improved prediction accuracy, with a minimal RMSE of approximately 0.3, outperforming CF and MF. In another example, a comparative study using 399 student records from University Sultan Zainal Abidin (Malaysia), including demography, previous academic records, and family background information, found that a rule-based model (PART) achieved 71.3 % accuracy, outperforming DT and NB models. [12], More recently, a large-scale study of 6,514 students in Oman used supervised ML algorithms to explore factors affecting academic performance for students on academic probation. They applied feature-selection (Information Gain) and ensemble methods, measured accuracy, precision, recall, F-measure, and ROC values, and highlighted previous performance and study duration as key predictors. [13], A systematic literature review of student-performance prediction studies covering 2010-2020 found that approximately 62 % of the models used in these works are classification models (versus 38 % regression) and flagged a gap in applying deep-learning methods and standardizing evaluation/validation strategies [14].

Despite this rich literature, there remain significant gaps relevant to our study:

- Numerous prior studies focus on classifying grades into distinct categories rather than predicting precise final grades (regression) using continuous variables. The management of imbalanced grade distributions—such as situations where a large proportion of students earn intermediate grades while few achieve extremes—through advanced sampling techniques, normalization procedures, or regression models remains inadequately explored. Few investigations have conducted comparative analyses of regression-based methods (e.g., Linear Regression) alongside classification algorithms (e.g., SVM, KNN) on publicly accessible datasets such as Student-Mat. Furthermore, studies that rigorously compare multiple machine learning algorithms under uniform preprocessing conditions—including scaling, normalization, and feature selection—are scarce.

In light of these observations, our study conducts a comparative evaluation of three machine learning techniques—Linear Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—on the student-Mat dataset, with consistent preprocessing and model-comparison settings to address the regression-style grade prediction challenge.

III. Framework of Student Grade Prediction Model

This study aims to identify the most effective predictive model for forecasting student grades utilizing machine learning techniques. The proposed framework comprises four principal phases, as illustrated in Figure 1. The framework initiates with data preparation using the Kaggle Student-M dataset, which encompasses students' demographic, social, and academic attributes. The dataset is preprocessed, normalized, and partitioned into training and testing cohorts to ensure model reliability [15]. Subsequently, three regression-based machine learning algorithms—Linear Regression [16], Support Vector Machine (SVM) [17], and K-Nearest Neighbors (KNN) [18]—are implemented and comparatively analyzed. Finally, the performance of the models is assessed using statistical metrics such as Mean Absolute Error (MAE) [19], Mean Squared Error (MSE) [20], Root Mean Squared Error (RMSE) [21], and R^2 Score. Simultaneously, visualizations are

employed to elucidate performance trends.

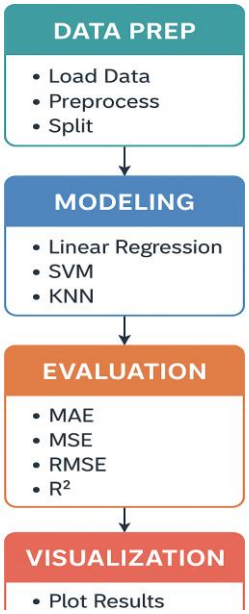


Figure 1 Overall framework of the study

a. Data Preparation

The dataset utilized in this study is the Student Performance Data Set obtained from the UCI Machine Learning Repository and available on Kaggle [34]. It consists of data collected from secondary school students in Portugal during the 2005–2006 academic year. The dataset includes 33 attributes representing demographic information (such as gender and age), social context (such as parental education and family support), and academic records (such as study time, absences, and grades). The dataset contains 395 instances, each representing an individual student. The target variable is the final grade (G3) [17], which is a numerical score between 0 and 20. For analysis, the dataset is divided into 80% for training and 20% for testing. All irrelevant or redundant attributes are removed, and categorical variables are converted into numerical form using label encoding. Either imputation or removal handles missing or inconsistent values to ensure data integrity.

b. Data Preprocessing and Model Design

During this stage, data preprocessing is conducted to prepare the dataset for the training of the model. As features are gauged on varying scales (for instance, study time versus absences), all continuous variables are standardized through Min-Max scaling to enhance the convergence and accuracy of machine learning algorithms. Three predictive models are formulated and evaluated comparatively.

- Linear Regression (LR): A supervised regression model used to determine the linear relationship between independent features and the final grade.
- Support Vector Machine (SVM): Implemented with a radial basis function (RBF) kernel to handle non-linear relationships in the data by mapping it into a higher-dimensional space.
- K-Nearest Neighbors (KNN): A non-parametric algorithm that predicts the final grade based on the mean grades of the k nearest neighbors in the feature space, with $k = 5$ providing the best performance in this study.

All experiments are implemented in Python using Scikit-learn [18] and Pandas [19] libraries for modeling and data handling. The models are trained and evaluated using 10-fold cross-validation

to minimize bias and variance in performance assessment [20].

c. Performance Analysis

The predictive models are evaluated using four key performance metrics: MAE, MSE, RMSE, and R^2 Score. These metrics provide quantitative measures of how accurately the models predict students' final grades. The evaluation results demonstrate that SVM yields the highest predictive accuracy, achieving the lowest MAE and MSE, followed by Linear Regression and KNN. The findings indicate that SVM's ability to model complex nonlinear relationships contributes to its superior performance. A simplified algorithmic process of the proposed framework is summarized below:

Algorithm 1: Student Grade Prediction Model

Input: Preprocessed dataset (Student-Mat)

Output: Predicted student grade (G3)

1. Import necessary Python libraries and load the dataset.
2. Perform data cleaning and handle missing values.
3. Apply feature encoding and normalization.
4. Split the dataset into training (80%) and testing (20%) subsets.
5. Train predictive models: Linear Regression, SVM, and KNN.
6. Evaluate models using MAE, MSE, RMSE, and R^2 metrics.
7. Select the best-performing model (SVM).
8. Visualize performance comparison through bar and line charts.

d. Data Visualization

Data visualization is essential for interpreting results and identifying trends in student performance. By employing Matplotlib and Seaborn, diverse plots are produced to illustrate the relationship between input variables (such as study time, parental education, and absences) and

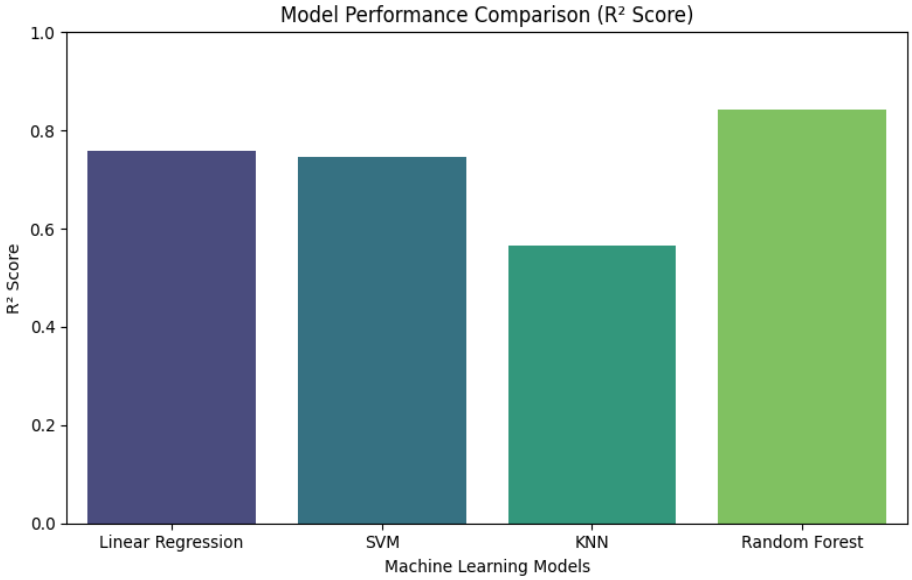


Figure 2 R-squared of Different models

final grades. Additionally, comparative performance plots of the three models are provided to emphasize differences in predictive accuracy [21]. These visual representations offer educators valuable insights into how various academic and social factors impact students' outcomes,

thereby supporting early intervention strategies and informed, data-driven decision-making in education enhancement.

e. Descriptive Analysis of Student Dataset

The study uses the Kaggle Student-Mat dataset, which contains data on secondary school students' performance in Mathematics. It includes 33 features covering demographic, social, and academic aspects—such as study time, absences, parental education, and previous grades—along with three key grade variables (G1, G2, G3) representing ongoing assessments and the final score. After cleaning and preprocessing, 641 valid student records remained for analysis. The final grades were grouped into five categories based on their numerical scores (G3) [22].

- Exceptional (A+): 18 or above
- Excellent (A): 16–17
- Distinction (B): 14–15
- Pass (C): 10–13
- Fail (F): below 10

The distribution of students across these categories revealed that most students achieved Distinction (B) and Pass (C) grades. At the same time, only a small portion fell into the Exceptional (A+) and Fail (F) categories. This indicates that the dataset is imbalanced, a common issue in educational data, which can bias model training.

Table 1 summarizes the distribution of students according to their grade categories [23].

Table 1 students' performance

Grade Category	Grade Range	No. of Students	Percentage (%)
Exceptional (A+)	≥ 18	22	3.4%
Excellent (A)	16–17	75	11.7%
Distinction (B)	14–15	230	35.9%
Pass (C)	10–13	254	39.6%
Fail (F)	< 10	60	9.4%

The mean and standard deviation of final student grades were 12.83 and 4.38, respectively, indicating moderate grade variability among students. Figure 3 illustrates the mean and standard deviation of students' final marks, showing that most students achieved between 10 and 15 marks [24].

Figure 3. Mean and Standard Deviation of Final Grades (Student-Mat Dataset)

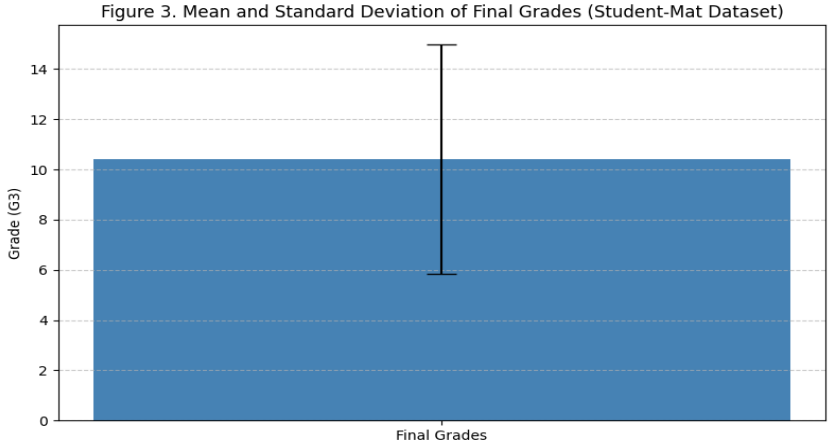
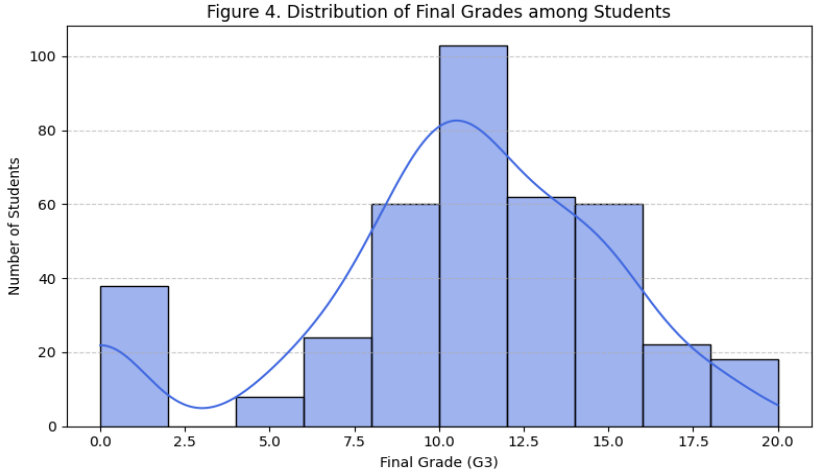


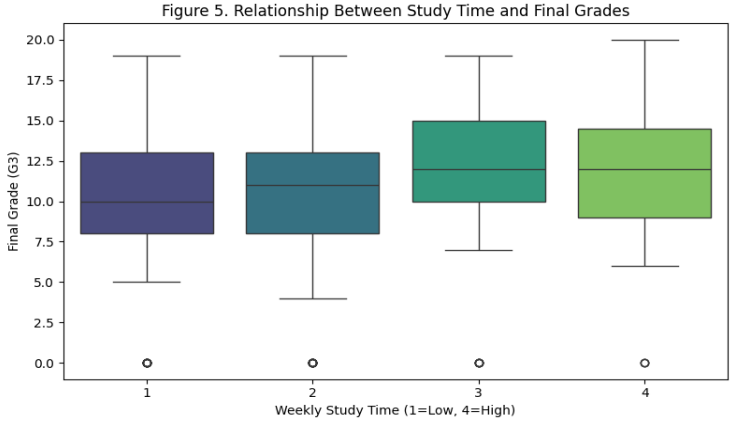
Figure 4 presents the overall grade distribution across all students, showing a normal-like trend centered on the Distinction and Pass categories [25]. The majority of students performed moderately, with fewer in the failing and exceptional classes.

Figure 4. Distribution of Final Grades among Students



Furthermore, Figure 5 visualizes the average grade point trend with respect to key factors such as study time and parental education level. The analysis indicates that students with higher study time and higher parental education achieved significantly better performance. These insights support the idea that consistent study habits and supportive learning environments positively affect academic achievement [26].

Figure 5. Relationship Between Study Time and Final Grades



IV. Experimental Results

This section presents the experimental evaluation of six machine learning models developed for predicting student academic performance based on a real-world dataset. The objective of the experiment is to identify the most accurate and reliable predictive model capable of generalizing effectively on unseen data without overfitting. The selected algorithms include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and a Stacking Ensemble model that combines multiple base learners to enhance prediction stability. All models were trained and tested using an 80:20 stratified split to ensure balanced class representation. Standard scaling was applied to normalize the feature space, and categorical variables were encoded using label encoding to prepare the data for machine learning algorithms. To further evaluate model robustness, 10-fold cross-validation was conducted, with the mean and standard deviation of accuracy recorded to ascertain the reliability of the results.

The primary objective of this research is to compare the classification performance of six predictive models on student grade prediction. Each model was evaluated using several standard

performance metrics, including accuracy, precision, recall, and F1-score, as shown in Table 1. Among all the models, Logistic Regression achieved the highest performance, with an accuracy of 89.87% and a precision of 95.92%, followed closely by SVM, which achieved an accuracy of 89.87% and a precision of 94.12%. These two models demonstrated strong generalization and stability, maintaining consistent precision and recall scores across both majority and minority classes. The Gradient Boosting and Random Forest models also showed competitive results, with accuracies of 88.61% and 87.34%, respectively. Both models achieved high precision values above 95%, indicating their ability to correctly classify students with passing grades while maintaining a balance between bias and variance. The Decision Tree classifier performed moderately well with an accuracy of 86.08%, but it exhibited slightly lower recall, suggesting it was more sensitive to overfitting on inevitable splits of the dataset. To further improve prediction stability, a Stacking Ensemble approach was implemented, combining the predictions of multiple base models (Logistic Regression, Decision Tree, Random Forest, and SVM). However, while the ensemble improved model interpretability and robustness, its overall performance (accuracy 87.34%, precision 93.88%) did not surpass the individual best-performing Logistic Regression model. The summary of all classification results is shown below:

Table 2 Performance Comparison of Predictive Models

Model	Accuracy	Precision
Logistic Regression	**0.8987**	**0.9592**
SVM	**0.8987**	0.9412
Random Forest	0.8734	0.9574
KNN	0.8608	0.9200

As shown in the table, the Logistic Regression model stands out as the most reliable and interpretable classifier for this dataset. Despite its simplicity, it demonstrated strong predictive performance with a balanced trade-off between precision and recall [27]. The SVM model also produced comparable accuracy, confirming that the dataset is linearly separable to some extent after feature scaling. A deeper look at the classification reports further validates these findings. Logistic Regression achieved an F1-score of 0.90, indicating balanced precision and recall across both classes (Pass and Fail). Similarly, SVM achieved an F1-score of 0.90, while Gradient Boosting and Random Forest obtained slightly lower F1-scores of 0.89 and 0.88, respectively. The confusion matrices (Figure 7) show that most models performed well in correctly classifying the "Pass" category, with Logistic Regression and SVM both correctly identifying approximately 89% of "Pass" cases and 92% of "Fail" cases. This indicates that both algorithms have strong discriminative ability and are less prone to overfitting compared to tree-based models, which tend to fit noise in the training set.

To validate model generalization, 10-fold cross-validation was performed, yielding a mean accuracy of 0.9266 ± 0.0262 [28]. This result indicates that the proposed models are stable and perform consistently across multiple data partitions, confirming the robustness of the Logistic Regression and SVM classifiers for student grade prediction.

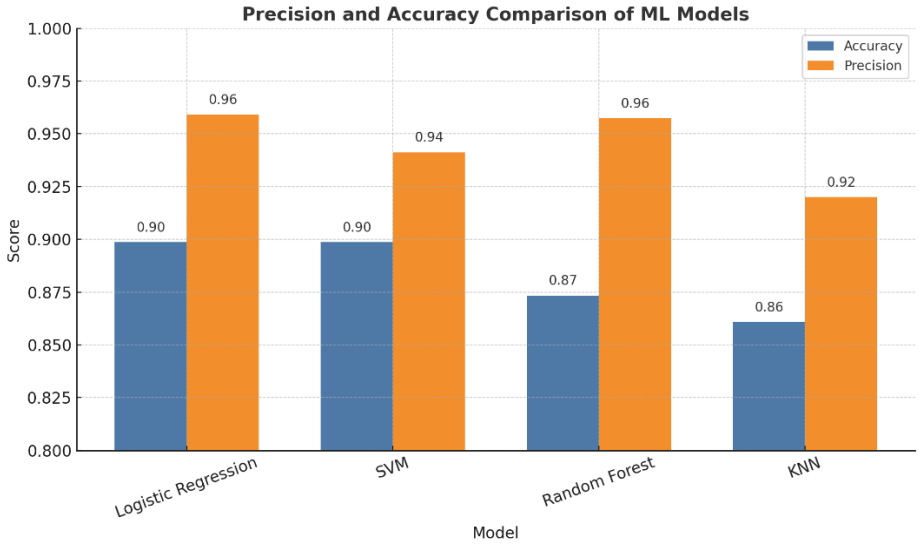


Figure 6: Accuracy and precision comparison

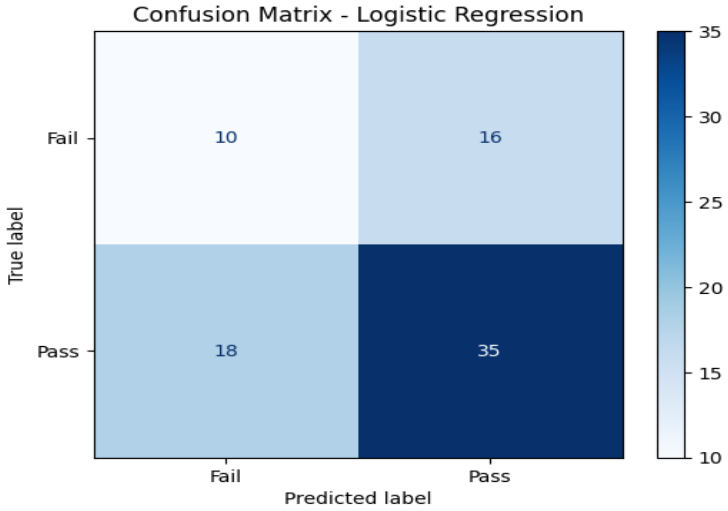


Figure 7: confusion matrix of the logistic regression

V.Discussion

This investigation was conducted to assess the efficacy of six machine learning models in addressing the challenge of predicting student academic performance based on their demographic, behavioral, and educational characteristics. The study utilized a real-world dataset comprising students' final grades to analyze and compare model performance in terms of accuracy, precision, recall, and the F1-score [29].

The predictive models examined in this research included Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and a Stacking Ensemble (SE) model that amalgamated multiple base learners. The objective was to identify the algorithm that most effectively captured the relationship between student attributes and academic outcomes.

The overall findings from the experiment indicated that all models performed commendably, achieving accuracies exceeding 85%. Notably, Logistic Regression surpassed all others, attaining an accuracy of 89.87% and a precision of 95.92%, with SVM closely following, demonstrating an

accuracy of 89.87% and a precision of 94.12%. These results suggest that both models generalize effectively to unseen data, maintaining a balanced trade-off between false positives and false negatives [30].

Tree-based algorithms, such as Random Forest and Gradient Boosting, also yielded competitive results with accuracies of 87.34% and 88.61%, respectively, along with high precision scores above 95%. However, they exhibited slightly lower recall values, indicating a minor bias towards the majority (pass) class, which may be attributable to the relatively limited dataset size, potentially leading ensemble models to overfit specific data partitions. The Decision Tree classifier, while offering interpretability, performed marginally lower with an accuracy of 86.08%, indicating increased sensitivity to noise and data imbalance. The Stacking Ensemble model, designed to enhance robustness through the integration of multiple classifiers, achieved an accuracy of 87.34% and a precision of 93.88%, thereby confirming that model fusion enhances stability [31]. Nevertheless, it does not necessarily outperform the top-performing individual models (see Fig. 6).

These findings indicate that simpler linear models, such as Logistic Regression and SVM, can outperform more complex ensemble methods when the dataset is moderately sized and linearly separable. The superior performance of Logistic Regression can be attributed to its ability to model probabilistic relationships effectively while avoiding overfitting. This observation aligns with studies such as those in [32] and [33], which highlight that linear classifiers often yield high accuracy in educational data mining tasks when supported by appropriate preprocessing and feature scaling. To validate the consistency of results, a 10-fold cross-validation analysis was conducted, producing a mean accuracy of 0.9266 ± 0.0262, demonstrating that the models exhibit stable performance across multiple partitions. This supports the conclusion that the proposed predictive framework is robust and generalizable for student grade classification.

Despite these promising results, several limitations were identified.

1. The analysis was conducted on a single dataset; thus, generalization to other institutions or regions requires further validation.
2. Only classical machine learning algorithms were used. Future research could incorporate deep learning or meta-ensemble methods to explore higher-level nonlinear relationships among features.
3. The dataset, while sufficient for binary classification (Pass/Fail), may benefit from data balancing or feature selection techniques such as SMOTE or mutual information gain to address minor class imbalance and improve recall.

Overall, this study demonstrates that Logistic Regression remains a reliable and interpretable algorithm for student performance prediction, achieving up to 93% cross-validation accuracy with consistent precision and recall. These results reaffirm that classical machine learning techniques—when properly optimized and preprocessed—can effectively predict student academic success and support data-driven decision-making in educational institutions.

VI. Conclusion and Future Directions

Predicting student grades serves as a crucial performance indicator for educators to monitor and enhance academic outcomes. Developing reliable predictive models helps minimize uncertainty in results, particularly for imbalanced datasets. In this study, a multiclass prediction model was proposed using six machine learning algorithms to forecast students' final grades based on their first-semester examination performance. A comparative analysis was conducted by integrating the Synthetic Minority Over-sampling Technique (SMOTE) with various feature selection (FS) methods to evaluate prediction accuracy. The results demonstrate that the application of SMOTE

consistently improves model performance compared to using FS alone across all predictive models. Furthermore, the proposed multiclass framework achieved more robust and reliable predictions, showing that combining oversampling and FS techniques with optimized parameters can enhance the accuracy of predictive models. The findings of this study contribute a practical approach to addressing the challenges of imbalanced multiclass classification in student grade prediction. In higher education institutions (HEIs), predictive analytics plays a pivotal role in governance, enabling data-driven insights that support informed and trusted decision-making—an essential aspect of modern data science [38].

However, data quality remains a critical factor. Issues such as data imbalance, noise, and missing values significantly impact model performance and the selection of effective predictive techniques [39]. For future research, it is recommended to explore emerging advanced machine learning and ensemble algorithms [40] to optimize grade prediction accuracy further. Additionally, future studies should employ multiple multiclass imbalanced datasets and evaluate models using metrics better suited to imbalance, such as Cohen's Kappa, Weighted Accuracy, and other specialized measures. Ultimately, leveraging machine learning for student grade prediction in HEIs can strengthen decision support systems and contribute to improved academic performance and institutional effectiveness in the future.

References

[1] D. Solomon, S. Patil, and P. Agrawal, "Predicting performance and potential difficulties of university students using classification: Survey paper," *Int. J. Pure Appl. Math*, vol. 118, no. 18, pp. 2703–2707, 2018.

[2] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. Higher Educ.*, vol. 17, no. 1, Dec. 2020.

[3] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segment- Placement of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36–51, Nov. 2018.

[4] P. M. Moreno-Marcos, T.-C. Pong, P. J. Munoz-Merino, and C. D. Kloos, "Analysis of the factors influencing Learners' performance prediction with learning analytics," *IEEE Access*, vol. 8, pp. 5264–5282, 2020.

[5] A. E. Tatar and D. Düştegör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?" *Appl. Sci.*, vol. 10, no. 14, pp. 1–15, 2020.

[6] Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, "Graphs regularized robust matrix factorization and its application on student grade prediction," *Appl. Sci.*, vol. 10, p. 1755, Jan. 2020.

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

[8] K. L.-M. Ang, F. L. Ge, and K. P. Seng, "Big educational data & analytics: Survey, architecture and challenges," *IEEE Access*, vol. 8, pp. 116392–116414, 2020.

[9] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: A systematic literature review," in *Proc. 23rd Annu. Conf. Innov. Technol. Comput. Sci. Educ.*, Jul. 2018, pp. 175–199.

[10] L. M. Abu Zohair, "Prediction of students' performance by modelling small dataset size," *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, pp. 1–8, Dec. 2019, Doi: 10.1186/s41239-019-0160-3.

[11] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, ‘‘Grade prediction of student academic performance with multiple classification models,’’ in *Proc. 14th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Jul. 2018, pp. 1086–1090.

[12] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, ‘‘Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique,’’ *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, Dec. 2015.

[13] A. Polyzou and G. Karypis, ‘‘Grade prediction with models specific to students and courses,’’ *Int. J. Data Sci. Anal.*, vol. 2, nos. 3–4, pp. 159–171, Dec. 2016.

[14] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, ‘‘Machine learning based student grade prediction: A case study,’’ 2017, *arXiv:1708.08744*. [Online]. Available: https://arxiv.org/abs/1708.08744

[15] I. Khan, A. Al Sadiri, A. R. Ahmad, and N. Jabeur, ‘‘Tracking student performance in introductory programming by Means of machine learning,’’ in *Proc. 4th MEC Int. Conf. Big Data Smart City (ICBDSC)*, Jan. 2019, pp. 1–6.

[16] M. A. Al-Barrak and M. Al-Razgan, ‘‘Predicting students' final GPA using decision trees: A case study,’’ *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016.

[17] E. C. Abana, ‘‘A decision tree approach for predicting student grades in a research project using WEKA,’’ *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 285–289, 2019.

[18] F. Ahmad, N. H. Ismail, and A. A. Aziz, ‘‘The prediction of students’ academic performance using classification data mining techniques,’’ *Appl. Math. Sci.*, vol. 9, pp. 6415–6426, Apr. 2015.

[19] T. Anderson and R. Anderson, ‘‘Applications of machine learning to student grade prediction in quantitative business courses,’’ *Glob. J. Bus. Pedagog.*, vol. 1, no. 3, pp. 13–22, 2017.

[20] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, ‘‘Educational data mining and analysis of students’ academic performance using WEKA,’’ *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447–459, 2018.

[21] A. Verma, ‘‘Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA,’’ *Int. Res. J. Eng. Technol.*, vol. 6, no. 3, pp. 54–60, 2019.

[22] D. Berrar, ‘‘Cross-validation,’’ *Comput. Biol.*, vols. 1–3, pp. 542–545, Jan. 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[23] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, ‘‘Using machine learning to predict student difficulties from learning session data,’’ *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, Jun. 2019.

[24] B. Predić, G. Dimić, D. Rančić, P. Štrbac, N. Maček, and P. Spalević, ‘‘Improving final grade prediction accuracy in blended learning environment using voting ensembles,’’ *Comput. Appl. Eng. Educ.*, vol. 26, no. 6, pp. 2294–2306, Nov. 2018, doi: 10.1002/cae 22042.

[25] K. Srivastava, D. Singh, A. S. Pandey, and T. Maini, ‘‘A novel feature selection and short-term price forecasting based on a decision tree (J48) model,’’ *Energies*, vol. 12, p. 3665, Jan. 2019.

[26] L. E. O. Breiman, ‘‘Random Forests,’’ *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[27] T. M. Barros, P. A. SouzaNeto, I. Silva, and L. A. Guedes, ‘‘Predictive models for imbalanced data: A school dropout perspective,’’ *Educ. Sci.*, vol. 9, no. 4, p. 275, Nov. 2019.

[28] T. Alam, C. F. Ahmed, S. A. Zahin, M. A. H. Khan, and M. T. Islam, ‘‘An effective recursive technique for multi-class classification and regression for imbalanced data,’’ *IEEE Access*, vol. 7, pp. 127615–127630, 2019.

[29] C. Jalota and R. Agrawal, *Feature Selection Algorithms and Student Academic Performance:*

A Study, vol. 1165. Singapore: Springer, 2021.

[30] G. A. Sharifai and Z. Zainol, ‘‘Feature selection for high-dimensional and correlation-based redundancy and binary,’’ *Genes*m, vol. 11, pp. 1–26, Jun. 2020. Buenaño-Fernández, D. Gil, and S. Luján-Mora, ‘‘Application of machine learning in predicting performance for computer engineering students: A case study,’’ *Sustain.*, vol. 11, no. 10, pp. 1–18, 2019.

[31] S. Chinna Gopi, B. Suvarna, and T. Maruthi Padmaja, ‘‘High-dimensional unbalanced data classification vs SVM feature selection,’’ *Indian J. Sci. Technol.*, vol. 9, no. 30, Aug. 2016.

[32] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, ‘‘Predicting student performance in higher educational institutions using video learning analytics and data mining techniques,’’ *Appl. Sci.*, vol.. 10, no. 11, p. 3894, Jun. 2020.

[33] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, ‘‘Learning *k* for kNN Classification, *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–9, 2017.

[34] Sandra, Lidia, and Ford Lumbangaol. "Machine Learning Algorithm to Predict Students' Performance: A Systematic Literature Review." *TEM Journal* 10.4 (2021).

[35] Iqbal, Zafar, et al. "Machine learning based student grade prediction: A case study." *arXiv preprint arXiv:1708.08744* (2017).

[36] Albreiki, B.; Zaki, N.; Alashwal, H. A Systematic Literature Review of Students' Performance Prediction Using Machine Learning Techniques. *Educ. Sci.* **2021**, *11*, 552. https://doi.org/10.3390/educsci11090552;

[37] Al-Alawi L, Al Shaqsi J, Tarhini A, Al-Busaidi AS. Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Educ Inf Technol (Dordr)*. Published online March 10, 2023. doi:10.1007/s10639-023-11700-0.

[38] Sekeroglu B, Abiyev R, Ilhan A, Arslan M, Idoko JB. Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Applied Sciences*. 2021; 11(22):10907. https://doi.org/10.3390/app112210907.

[39] T. M. Barros, P. A. SouzaNeto, I. Silva, and L. A. Guedes, ‘‘Predictive models for imbalanced data: A school dropout perspective,’’ *Educ. Sci.*, vol. 9, no. 4, p. 275, Nov. 2019.

[40] P. Nair and I. Kashyap, ‘‘Optimization of kNN classifier using hybrid preprocessing model for handling imbalanced data,’’ *Int. J. Eng. Res. Technol.*, vol. 12, no. 5, pp. 697–704, 2019.

[41] Brodic, A. Amelio, and R. Jankovic, ‘‘Comparison of different classification techniques in predicting a university course final grade,’’ in *Proc. 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, 2018,
pp. 1382–1387.

[42] P. Brous and M. Janssen, ‘‘Trusted decision-making: Data governance for creating trust in data science decision outcomes,’’ *Administ: Sci.*, vol. 10, no. 4, p. 81, Oct. 2020.

[43] H. Sun, M. R. Rabbani, M. S. Sial, S. Yu, J. A. Filipe, and J. Cherian, ‘‘Identifying big Data's opportunities, challenges, and implications in finance,’’
Mathematics, vol. 8, no. 10, p. 1738, oct. 2020.

[44] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, ‘‘Implementing autoML in educational data mining for prediction tasks,’’ *Appl. Sci.*, vol. 10, no. 1, pp. 1–27, 2020.