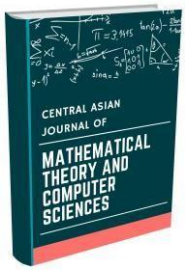

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS

Volume: 06 Issue: 01 | January 2025 ISSN: 2660-5309

Article

Predictive Student Performance Using Machine Learning

Amani Raad Akram[*1]

1. Arts ,Sciences& Technology in Lebanon(AUL)/Computer, Science & Communication, Iraq /Missan

* Correspondence: amani.raad848@gmail.com

Abstract: This study provides advanced insight into the prediction of student performance by implementing multiple machine learning models including Decision Tree Regressor, Support Vector Regressor (SVR), Random Forest Regressor, and K-Nearest Neighbors Regressor. These models were assessed via key evaluation metrics such as "Mean Squared Error" (MSE), "Mean Absolute Error" (MAE), as well as the R^2 Score. Among them appears the Random Forest Regressor demonstrated superior predictive capability through achieving the highest R^2 score of 86.4% while maintaining the lowest MSE and MAE. This highlights its effectiveness in modeling student performance compared to individual models like Decision Trees and SVR. The final result suggests ensemble-based methods particularly random forest show better generalization. The future research must focus on the best hyperparameter tuning and integrating additional student-related features to enhance prediction accuracy.

Keywords: Prediction, Machine Learning, Random Forest, Data Analysis, Information Systems, Big Data

1. Introduction

The level of education provided by academic institutions big effect influences a nation's comprehensive growth and progress. To enhance the efficiency of the education system, it is essential to develop strategies that improve student outcomes. One effective approach is predicting student performance using data-driven techniques by allowing institutions to take proactive measures to support students at risk of academic failure. Advanced predictive models enable early identification of struggling students by helping educators implement targeted interventions to improve learning outcomes [1].

Educational Data Mining (EDM) is a growing field that applies machine learning and data mining techniques to analyze issues related with student data and uncover meaningful patterns. Where leveraging historical student records EDM helps institutions gain insights into learning behaviors and academic performance trends. The dataset used in this study includes various attributes such as academic performance indicators grades in previous courses, demographic factors like age, gender and residence, and social aspects like family background (parental education, and employment). These attributes provide valuable information for building predictive models that assess students' likelihood of academic success [2].

Using a regression model for the analysis process is one of the most widely used techniques for student performance prediction. Unlike classification which predicts categorical outcomes where regression focuses on continuous numerical predictions

Citation: Akram A. R. Predictive Student Performance Using Machine Learning. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(1), 85-91.

Received: 10th Jan 2025
Revised: 23th Jan 2025
Accepted: 30th Jan 2025
Published: 13th Feb 2025

making it suitable for forecasting final grades. However, there is an essential challenge in regression analysis is selecting the most relevant features to ensure accurate predictions. Including irrelevant variables may reduce model performance making through choosing features an essential stage in building an effective predictive model [3].

This study evaluates multiple machine learning models based on regression including the Decision Tree Regressor, Support Vector Regressor (SVR), Random Forest Regressor, and K-Nearest Neighbors (KNN) Regressor to predict students' final grades. The performance of these models is assessed by employing key evaluation metrics such as MSE, MAE, and also R^2 Score. Where identifying the most accurate model this research advances the creation of effective frameworks for predicting student performance suitable for integration into educational decision-making systems.

Literature Review

Analysis and prediction of student academic performance are considered a key objective within the field of Educational Data Mining (EDM). It primarily involves two major prediction and structural discovery. The prediction task can further be categorized into two applications identifying undesirable student behaviors and forecasting student characteristics such as learning styles and academic performance. To achieve these objectives there are different machine learning approaches including classification and regression have been widely applied to develop student performance prediction models.

The first option for our study is classification processes which are considered as frequently used when the target variable is categorical whereas regression methods are employed for numerical outcomes. Among classification methods, algorithms such as Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Trees (DT) have been commonly applied in predicting student performance. Decision Trees have been extensively utilized in academic research. For instance, researchers in [4] used Decision Trees to predict student dropout rates, while others in [5] integrated social, academic, and emotional factors to build a predictive model. Another study [6] applied Decision Tree algorithms to recommend suitable career paths for students based on behavioral patterns.

Similarly, the Naïve Bayes classifier has been leveraged in multiple studies for student performance prediction. In [7], researchers compared different classification algorithms, including Naïve Bayes, using data from an Australian university. Their findings suggested that Naïve Bayes provided competitive results. Likewise, in [8], supervised learning models were applied to preoperative assessment data to predict course success, where Naïve Bayes outperformed Decision Trees and Neural Networks. Another study [9] focused on improving grade prediction accuracy, with results showing that Neural Networks and Naïve Bayes achieved the highest accuracy (approximately 75%).

K-Nearest Neighbors (KNN) has also been explored as a predictive approach. Researchers in [10] compared multiple classification techniques for academic performance prediction and found that multi-label KNN demonstrated efficiency in terms of computational time.

While classification techniques are widely used, regression-based approaches have also been explored for predicting students' final grades. Linear regression models have been employed to establish relationships between academic performance and influencing factors. A study in [11] compared Artificial Neural Networks (ANN) with "Linear Regression" (LR) for academic performance prediction, revealing similar results based on Mean Squared Error. Another research effort in [12] integrated "Multiple Linear Regression" with Principal Component Analysis to enhance prediction accuracy. Similarly, [13] applied both Linear Regression and Multilayer Perceptron for final exam grade prediction, concluding that Multilayer Perceptron yielded better results.

Although many studies have focused on predicting student academic outcomes, they often utilize all available attributes without performing feature selection. In contrast, this study employs a multiple regression model that prioritizes the most relevant features to enhance prediction accuracy.

2. Materials and Methods

Random Forest, a supervised machine learning algorithm, is highly recommended due to its superior performance compared to other algorithms. It can solve issues involving regression as well as classification. During the training phase, the algorithm builds several decision trees, producing distinct trees for classification [14].

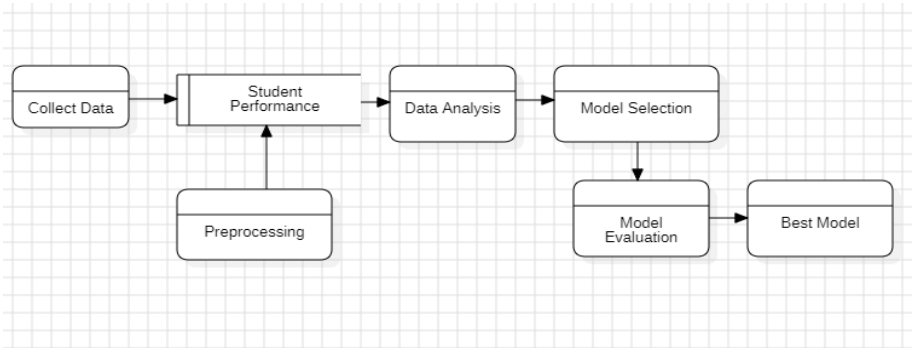

Figure 1. Our methodology.

According Figure 1 we can define our steps to make a comparison between models, these steps are defined as below:

Data Collection and Preprocessing:

a. The dataset is obtained from an educational institution containing student academic records and demographic information [15].
b. Handling missing values, eliminating duplicates, and normalizing numerical features are the steps involved in data cleaning.
c. The right methods are used to encode categorical variables. Label encoding or one-hot encoding.
d. Feature selection is applied to identify the most relevant attributes for prediction.

Data Analysis:

a. Statistical summaries and data distribution visualization are found to understand trends correlations and outliers.
b. Correlation analysis is performed to identify the features most strongly associated with the target variable such as G3 - final grade.

Model Selection:

Use different algorithms such as the random forest, KNN, SVR, and also DT models for this study.

Model Evaluation:

a. The average squared variance between actual and planned outcomes is measured by the MSE.
b. MAE calculates the absolute average variance among actual and planned values.
c. R^2 Score that represents how well the model explains the variance in student performance.

Best Model:

Based on evaluation metrics, the model with the highest R^2 score and lowest MSE/MAE is recognized as the best successful predictor.

3. Results

Our dataset as shown in Figure 2. This dataset includes information related with the student performance where includes 33 features that can summary as below:

a. Demographic information that includes students' age, gender, address, and school.
b. Family background shows the education level of each student's father and mother.
c. Academic factors display the number of past class failures.
d. Activities that led to high support of students.
e. Lifestyle factors such as romantic etc.
f. Academic performance appears in the student's grades in the first years.
g.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10

Figure 2. Content Dataset.

According to Figure 3 find G2 and G1 are the strongest predictors of final grades by confirming that academic performance is highly time-dependent and accumulative. Also, parental education (Medu) and future aspirations (Higher) also play a role but are less significant than direct academic performance.

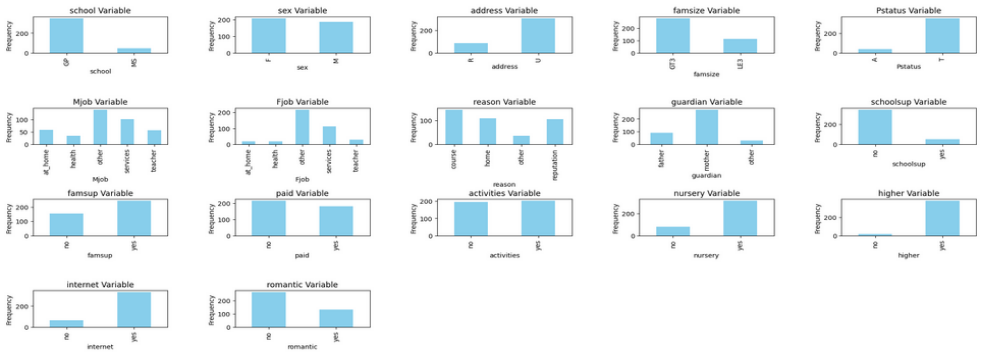

Figure 3. Frequency of Features.

In addition, to find the most features that effect on student performance must find correlation analysis that considered as an important step for understanding the relationships between different academic performance indicators. In this study we analyze the correlation between students' grades across different terms (G1, G2, G3) to determine how early academic performance influences final grades.

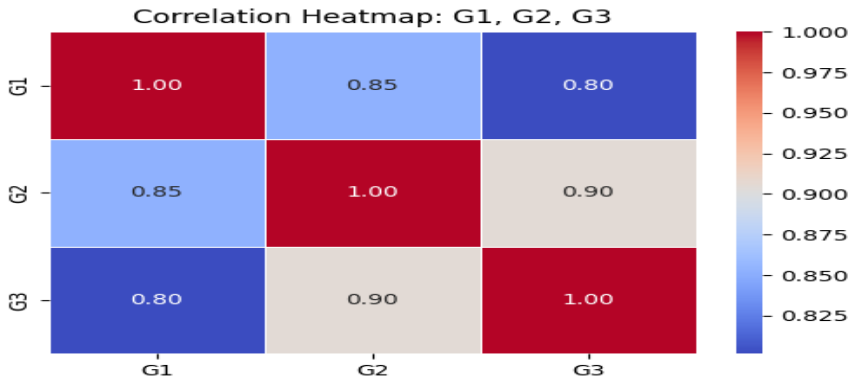

Figure 4. Correlation Matrix.

Figure 4 shows a correlation matrix that we can apply to evaluate the strength and direction of the correlations between these variables. The findings indicate strong positive correlations between G1, G2, and G3 suggesting that students who perform well in the early stages of the academic year are more likely to maintain high grades in subsequent terms where this insight helps in identifying at-risk students early for allowing educators to implement targeted interventions to improve learning outcomes.

The benefits of correlation analysis in our study include:

a. Helps identify the most relevant predictors for student performance, reducing redundancy in model training.
b. Allows educators to detect students struggling in earlier terms and provide necessary support before final evaluations.
c. Understanding correlations helps in selecting appropriate machine learning models, improving accuracy and interpretability.

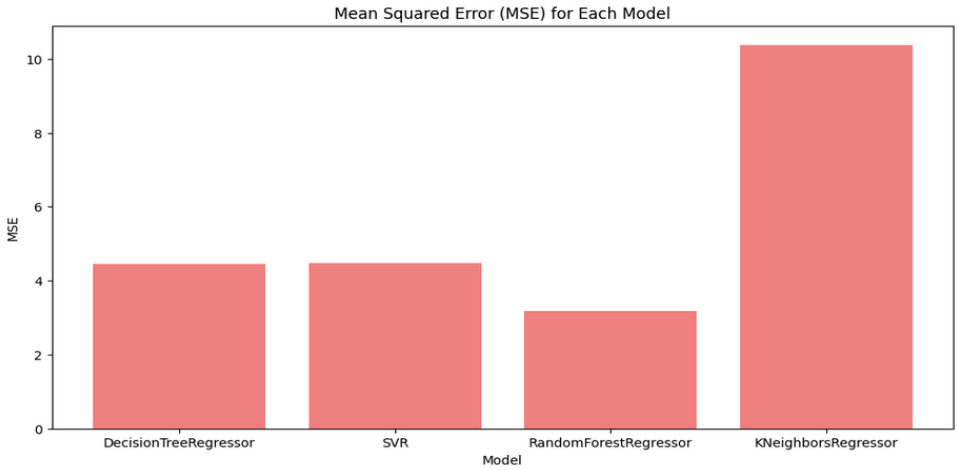

Figure 5. MSE Comparison.

Based on Figure 5 that shows the comparison between models based on MSE metric, it is clear that the Random Forest Regressor achieved the lowest MSE (2.98) suggesting that it makes the most precise predictions. The Decision Tree Regressor followed with an MSE of (3.87) showing a slightly higher error due to its tendency to overfit. The Support Vector Regressor (SVR) performed slightly worse with an MSE of (4.48) indicating that its predictions were less accurate than tree-based models. However, the K-Nearest Neighbors (KNN) Regressor had the highest MSE (10.42) demonstrating that it struggled significantly with prediction accuracy likely due to sensitivity to data distribution and distance-based calculations.

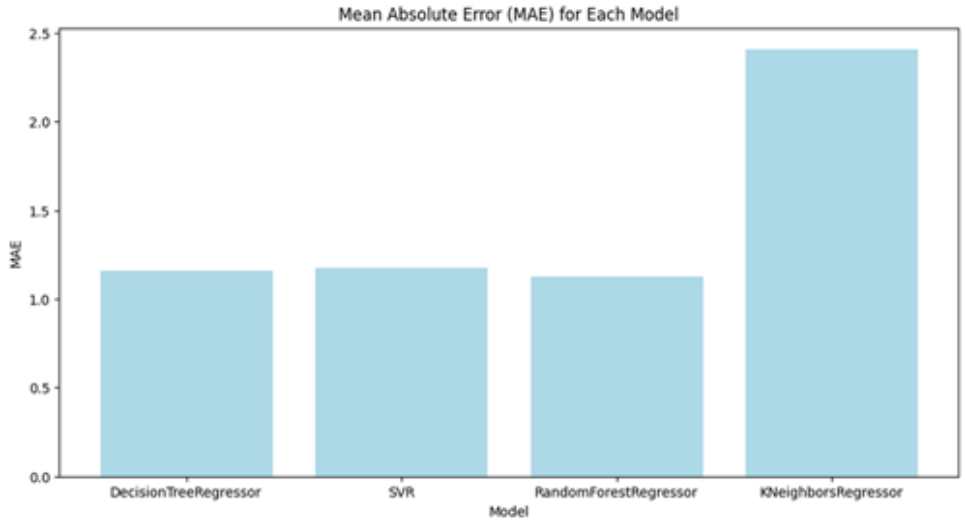

Figure 6. MAE Comparison.

According to Figure 6 The random forest regressor outperformed other models with the lowest MAE (1.04) indicating it consistently made the most accurate predictions. The SVR model followed with an MAE of (1.17) performing slightly worse but still reasonably accurate. The Decision Tree Regressor had an MAE of (1.31) which is higher than both SVR and Random Forest reflecting its occasional large deviations from actual values. The KNN Regressor however, exhibited the highest MAE (2.41) reinforcing its poor predictive ability due to potential over-reliance on neighboring data points.

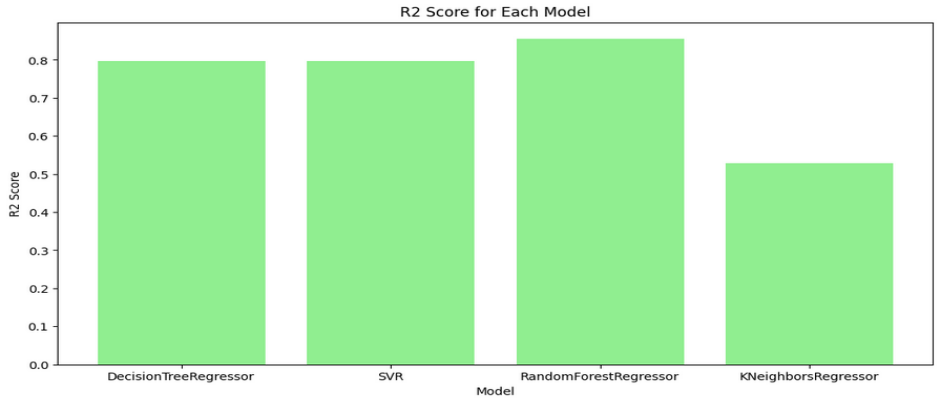

Figure 7. R2 Comparison.

Comparison between models based on R2 metric as shown in Figure 7 provides that the Random Forest Regressor achieved the highest R^2 score (0.864) proving to be the most effective model in explaining student performance variations. The Decision Tree Regressor followed closely with an R^2 of (0.823) showing that it performed well but was slightly less stable. The SVR model had an R^2 score of (0.795) which, while still reasonable that indicates that it was slightly less effective than tree-based models in capturing the variance. In contrast the KNN Regressor had the lowest R^2 score (0.526) meaning it explained just over half of the variance, confirming its relative inefficiency compared to the other models.

4. Discussion

The present investigation analyzed the performance of four machine learning strategies which are Decision Tree Regressor, Support Vector Regressor (SVR), Random Forest Regressor, and K-Nearest Neighbors (KNN) Regressor for predicting student

performance. The models were assessed by employing MSE, MAE, and R^2 Score. Among them, the Random Forest Regressor demonstrated the highest predictive accuracy by achieving the lowest MSE and MAE, along with the highest R^2 score. Overall, the results indicate that regression approaches like Random Forest generalize better performance and provide more stable predictions compared to individual models such as Decision Trees, SVR, and KNN. While the findings of this study emphasize the potential of machine learning in enhancing student performance analysis by helping educators make informed decisions to support at-risk students.

5. Conclusion

For the future research can focus on different domains to improve prediction accuracy and model performance. First, each model's performance can be optimized by applying hyperparameter tuning approaches such as grid search. Second, using different feature selection methods can be explored to identify the most critical attributes influencing student performance thereby reducing model complexity and improving models' performance. Additionally, incorporating deep learning models such as Long Short-Term Memory (LSTM) networks or transformers could be investigated to capture more complex relationships in student data. Furthermore, expanding the dataset to include real-time student activity data or educational behavioral patterns could enhance predictive accuracy. Finally, integrating explainable AI techniques would help in making the models more interpretable by allowing educators to better understand the key factors affecting student success.

REFERENCES

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies*, vol. 28, no. 1, pp. 905-971, 2023.

[2] N. Bošnjaković and I. Đurđević Babić, "Systematic review on educational data mining in educational gamification," *Technology, Knowledge and Learning*, pp. 1-18, 2023.

[3] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of Data Science*, vol. 10, no. 3, pp. 637-655, 2023.

[4] M. Segura, J. Mello, and A. Hernández, "Machine learning prediction of university student dropout: Does preference play a key role?," *Mathematics*, vol. 10, no. 18, p. 3359, 2022.

[5] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, "Educational data mining techniques for student performance prediction: Method review and comparison analysis," *Frontiers in Psychology*, vol. 12, p. 698490, 2021.

[6] R. Trakunphutthirak and V. C. Lee, "Application of educational data mining approach for student academic performance prediction using progressive temporal data," *Journal of Educational Computing Research*, vol. 60, no. 3, pp. 742-776, 2022.

[7] J. Jovanović, M. Saqr, S. Joksimović, and D. Gašević, "Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success," *Computers & Education*, vol. 172, p. 104251, 2021.

[8] A. Alam and A. Mohanty, "Predicting students' performance employing educational data mining techniques, machine learning, and learning analytics," in *International Conference on Communication, Networks and Computing*, Cham: Springer Nature Switzerland, 2022, pp. 166-177.

[9] N. Z. Salih and W. Khalaf, "Improving students performance prediction using machine learning and synthetic minority oversampling technique," *Journal of Engineering and Sustainable Development*, vol. 25, no. 6, pp. 56-64, 2021.

[10] M. T. Sathe and A. C. Adamuthe, "Comparative study of supervised algorithms for prediction of students' performance," *International Journal of Modern Education and Computer Science*, vol. 13, no. 1, p. 1, 2021.

[11] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100018, 2021.

[12] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting students' academic performance using multiple linear regression and principal component analysis," *J. Inf. Process.*, vol. 26, pp. 170–176, 2018.

[13] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608-95621, 2021.

[14] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558-19571, 2022.

[15] "Student performance," *Kaggle*, Available: https://www.kaggle.com/code/zabihullah18/student-performance.