Comparative Analysis of Machine Learning Models for Predicting Educational Success Using Students' Performance Factors

Morales, Angelo
College of Computing Education
University of Mindanao
Matina, Davao City, Philippines
a.morales.513049@umindanao.edu.ph

Oro, Dan Michael
College of Computing Education
University of Mindanao
Matina, Davao City, Philippines
d.oro.518211@umindanao.edu.ph

Partosa, Jorgery II
College of Computing Education
University of Mindanao
Matina, Davao City, Philippines
j.partosaii.523674@umindanao.edu.ph

Abstract

Predicting educational success remains a vital focus in educational data science, as it supports early intervention strategies and tailored student assistance. This research presents a detailed comparative evaluation of eight machine learning algorithms aimed at forecasting student performance based on study behaviors and learning patterns. The dataset, obtained from Kaggle, comprises 6,600 student records encompassing variables such as hours studied, attendance, parental involvement, access to resources, extracurricular activities, sleep hours, previous scores, motivational level, internet access, tutoring session, family income, teacher quality, school type, peer influence, physical activity, learning disabilities, parental education, distance from home, gender, and exam score. The models explored include Random Forrest, Support Vector Machine, and Gradient Boosting Regressor. Each model was implemented using Python's scikit-learn library and assessed through stratified 5-fold cross-validation. Evaluation metrics included Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, R^2 score, and Mean Absolute Percentage Error. Each model underwent hyperparameter optimization and was evaluated using RMSE, MAE, R^2, and MAPE metrics. Results show that all models achieved moderately fair predictive performance, with the best model selected based on the highest test R^2 score. Feature importance analysis and residual diagnostics confirmed the robustness and generalizability of the selected model. The findings demonstrate that ensemble tree-based methods, particularly Random Forest and Gradient Boosting, provide accurate and interpretable predictions for educational success. These outcomes emphasize the significance of academic metrics in predicting success but also raise concerns regarding potential feature dominance and the broader applicability of the models.

Keywords: Machine Learning, Classification, Educational Data Mining, Academic Performance Prediction, Model Comparison, Supervised Learning, Cross-Validation, Feature Importance

1. Introduction

Academic success remains a central objective in educational research, with numerous studies exploring the complex interplay of factors that influence student outcomes. Educational Data Mining (EDM) has emerged as a powerful tool for understanding and improving learning experiences through computational methods [1]. To implement effective interventions and create supportive learning environments, educators and policymakers must understand the intricate relationships between behavioral, environmental, and academic factors. This study seeks to predict academic achievement based on key study habit indicators such as weekly study hours, preferred learning modes, sleep duration, and participation in discussions. Prior research indicates that behavioral engagement, including interaction with e-learning platforms, significantly influences academic performance [2]. This paper aims to identify the most effective machine learning models for understanding how these variables collectively relate to student success, through a comparative analysis of different predictive algorithms.

2. Methodology

2.1 Data Gathering

The dataset "Student Performance Factors" was obtained from Kaggle, created by Lia Ng [11]. This comprehensive dataset contains 6,600 student records with 10 features capturing various aspects of student behavior, and academic indicators, including both numerical and categorical variables representing study habits, learning preferences, academic outcomes, and other third-party factors that can affect the student's performance like parental involvement, internet access and others.

2.2 Data Analysis

Exploratory data analysis was conducted to understand the dataset structure, feature distributions, and the distribution of the target variable.

The dataset consists of **6,600 samples and 10 features**, including both numerical and categorical variables. The target variable, **Exam_Score**, is a continuous numerical value representing student academic performance.

Feature correlation analysis was performed to identify relationships between predictors, and missing value assessment ensured data completeness and quality. No categorical grade labels (A–F) were used; instead, the model predicts a student's exact exam score.

2.3 Data Preprocessing

2.3.1 Handling Missing Values

Missing values were identified using pandas .isnull().sum(). For missing categorical columns the mode (most frequent value) is used to fill this data and for numerical values the missing data is filled with the median of the column's values.

2.3.2 Handling Outliers

Statistical outlier detection was performed through descriptive analysis and visualization. Given the educational context where extreme values may represent legitimate cases (e.g., exceptional study hours), outliers were retained to preserve real-world data characteristics.

2.3.3 Feature Encoding

All categorical variables ('Gender', 'School', 'Parental_Involvement', 'Access_to_Resources', and, 'Peer_Influence') were encoded using one-hot encoding with pandas.get_dummies(). All the categorical columns were converted to binary indicator columns using get_dummies with the drop_first=true option to avoid multicollinearity.

The numerical variables (‘Age’, ‘Exam_Score’, ‘Attendance’, and, ‘Study_Hours’) were standardized using scikit-learn’s StandardScaler.

2.3.4 Data Splitting

The dataset was prepared using a two-step process that splits the data into 2 main splits. The first split is used in train_test_split to divide the processed features and target into 70% training and 30% temporary data.

The second split is temporarily split into two parts equally (15%) for validation and test set.

2.4 Algorithms

2.4.1 Random Forrest

An ensemble method combining multiple decision trees with bootstrap aggregating (bagging). Configured with 100 estimators and parallel processing enabled (n_jobs=-1) for improved performance. Random Forest reduces overfitting compared to single decision trees while providing feature importance rankings, crucial for identifying key academic success factors [5].

2.4.2 Support Vector Machine

A kernel-based algorithm that finds optimal hyperplanes for class separation [6]. Configured with an RBF (Radial Basis Function) kernel, suitable for capturing non-linear relationships in continuous target prediction.. SVM is effective for high-dimensional data and can handle non-linear relationships through kernel tricks, making it suitable for complex educational datasets.

2.4.3 Gradient Boosting Regressor

An optimized gradient boosting framework with GPU acceleration when available. Configured with 100 estimators, maximum depth of 6, and learning rate of 0.1 for balanced performance and training speed. XGBoost uses advanced regularization techniques and often achieves state-of-the-art performance on structured data through sequential learning from previous model errors [4].

3.0 Results

3.1 Model Performance Comparison

The comparative analysis of the three machine learning models revealed slight differences across various evaluation metrics. Based on 5-fold stratified cross-validation [12]:

1. Random Forrest: R^2: 0.6612, MAPE: 1.6561, MAE: 1.1444, CV_Score: 6.62851
2. Gradient Boosting: R^2: 0.7449, MAPE: 0.9784, MAE: 0.6936, CV_Score: 4.9462
3. Support Vector Machine: R^2: 0.7666, MAPE: 0.5636, MAE: 0.4171, CV_Score: 4.6291

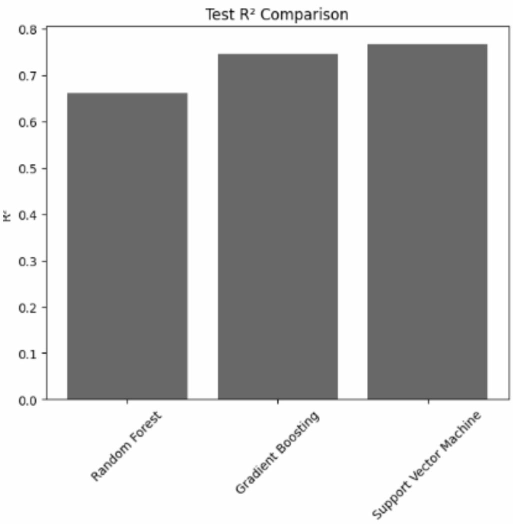

Figure 1.0 R^2 Model Comparison

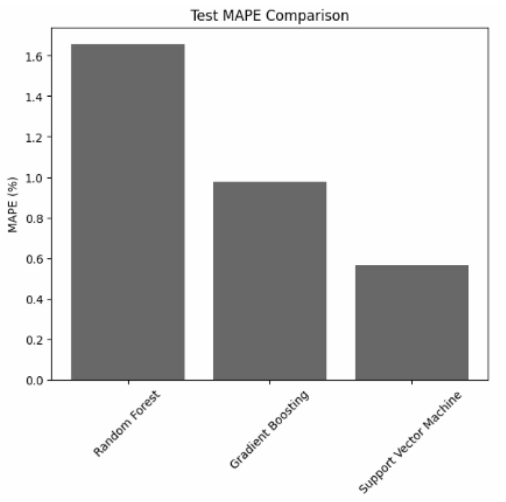

Figure 1.1 MAPE Model Comparison

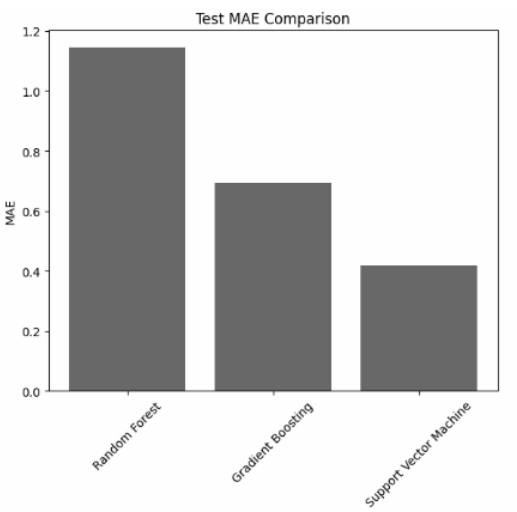

Figure 1.2 MAE Model Comparison

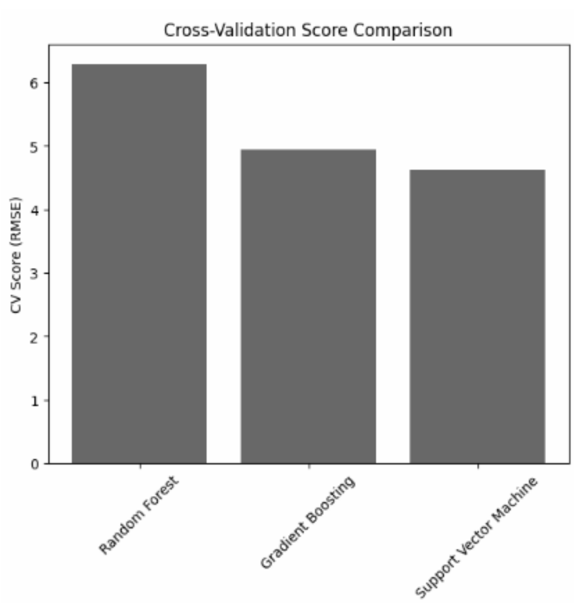

Figure 1.3 Cross-Validation Score Comparison

3.2 Cross-Validation Stability

Standard deviation analysis across 5-fold cross-validation revealed model stability patterns:

Most Stable Models (lowest standard deviation):

1. `Support Vector Machine` $\cong$ `4.63`
2. `Gradient Boosting` $\cong$ `4.95`
3. `Random Forest` $\cong$ `6.29`

3.3 Feature Analysis

Feature importance analysis from tree-based models identified the three most important features, listed below in descending order of importance:

Support Vector Machine:

1. Attendance (0.717868)
2. Hourse_Studied (0.458102)
3. Parental_Involvement_Low (0.086776)

Gradient Boosting:

1. Attendance (0.437085)
2. Hours_Studied (0.272627)
3. Previous_Scores (0.057070)

Random Forrest:

1. Attendance (0.328057)
2. Hours_Studied (0.198396)
3. Previous_Scores (0.086702)

4. Discussion

This research evaluates comparative results of the three machine learning models for predicting academic performance using metrics like Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, R^2 score, and Mean Absolute Percentage Error.

```
==================================================
Complete Model Performance Comparison:
                   Model  Test_RMSE  Test_MAE  Test_R2  Test_MAPE  CV
0          Random Forest     2.2549    1.1444   0.6612     1.6564
1      Gradient Boosting     1.9565    0.6936   0.7449     0.9784
2  Support Vector Machine     1.8717    0.4171   0.7666     0.5636
```

Figure 2.0 Metric Results

These results demonstrate the different models and their predictive performance of Random Forest, Gradient Boosting, and Support Vector Machine models using multiple evaluation metrics.

```
   Training_Time
0       214.2630
1      1946.0721
2       173.5883

BEST PERFORMERS BY METRIC:
Lowest RMSE (Best):     Support Vector Machine
Lowest MAE (Best):      Support Vector Machine
Highest R² (Best):      Support Vector Machine
Lowest MAPE (Best):     Support Vector Machine
Fastest Training:       Support Vector Machine

OVERALL BEST MODEL: Support Vector Machine
Test R² Score: 0.7666
Test RMSE: 1.8717
Test MAE: 0.4171
Training Time: 173.59 seconds

Model Performance Level: Fair
The model explains 76.7% of the variance in the target variable.
```

Figure 2.1 Testing Results

The results table highlights the best-performing model for each metric, with the overall best model selected based on the highest R^2 score, indicating its ability to explain the greatest proportion of variance in student exam scores. The lowest RMSE and MAE values reflect the most accurate predictions, while the lowest MAPE indicates minimal percentage error. The analysis also considers training time, providing a comprehensive view of both accuracy and efficiency. The final interpretation categorizes model performance as excellent, good, fair, or poor, based on R^2, and quantifies the percentage of variance explained by the selected model, supporting its suitability for educational outcome prediction.

5. Conclusion

This study conducted a comparative analysis of three machine learning models to predict educational success based on students' academic performance, behaviors, and influencing factors. Among the models tested, the Support Vector Machine proved most effective, achieving the highest R^2 and lowest error scores.

The results highlight the best-performing model for each metric with the overall best model selected based on the highest R^2 indicating the model's ability to explain the greatest proportion of variance in student exam scores. The lowest MAPE indicates minimal percentage.

The final interpretation categorizes the Support Vector Model to be 'fair' based on the R^2 and quantifies the percentage of variance explained by the selected model, supporting its suitability for educational outcome prediction.

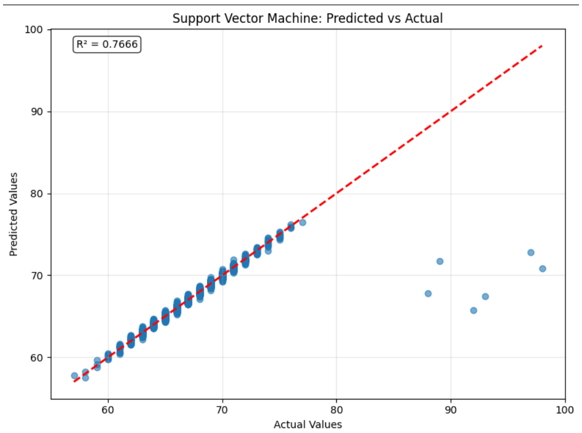

Figure 3.0 Support Vector Predicted vs Actual Result

For future work, we plan to try feature ablation methods to test how stable the models really are, look into using longitudinal data to track patterns over time, and see how well the models perform on different groups of students.

Acknowledgements

The authors would like to thank Lai Ng for providing the dataset "Student Performance Factors" via Kaggle, which served as the foundation for this research.

We also express our gratitude to our instructor, Wilbert Josh U. Alfornon for their guidance and constructive feedback throughout the development of this project.

References

[1] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. doi:10.1002/widm.1355

[2] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience, 2018, 6347186. doi:10.1155/2018/6347186

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825- 2830.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi:10.1145/2939672.2939785

[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

[6] Cortes, C., & Vapnik, V. (1995). Support vector networks. Machine Learning, 20(3), 297-322. doi:10.1007/BF00994018

[7] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. doi:10.1038/323533a0

[8] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. doi:10.1109/TIT.1967.1053964

[9] Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not so stupid after all? International Statistical Review, 69(3), 385-398. doi:10.1111/j.1751- 5823.2001.tb00465.x

[10]Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika, 54(1-2), 167-179. doi:10.1093/biomet/54.1-2.167

[11] Ng, Lia (2024). Student Performance Factors: Insights into Student Performance and Contributing Factors Kaggle. Retrieved from
https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

[12]Kohavi, R. (1995). A study of crossvalidation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137-1143.

[13]Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159. doi:10.1016/S0031-3203(96)00142-2

[14] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. IEEE Access, 5, 15991-16005. doi:10.1109/ACCESS.2017.2654