PREDICTING STUDENT ACADEMIC PERFORMANCE WITH MACHINE LEARNING IN UDNR, MYANMAR

Myat Nyein Moe[1], Soe Moe Lwin[2], Zin Mar Oo[3]

[1,2] Department of Information Technology, Defence Services Technological Academy, Pyin Oo Lwin, Myanmar

[3] Graduate School of Economics, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577 Japan

[1]myatnyein1982@gmail.com, [2]smoelwin@iuj.ac.jp, [3]zinmaroo@iuj.ac.jp

ABSTRACT

This study assesses the predictive power of five prominent machine learning algorithms—Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Regressor (GBR), and Extreme Gradient Boosting (XGBoost)—in forecasting student academic performance using survey data from the University for the Development of the National Races of the Union (UDNR) in Myanmar. While SVR emerges as the most accurate model, the other algorithms demonstrate close performance. Utilizing Permutation Feature Importance (PFI) analysis, the study identifies lagged GPA, total marks obtained in grade 11, gender, weekly study hours, and ethnicity as significant predictors. These findings underscore the potential of machine learning for precise student performance predictions and personalized education strategies, with multiple viable algorithmic choices.

Keywords: academic performance, machine learning, feature importance, Myanmar

I. INTRODUCTION

Academic achievements in higher education significantly impact employment, income, and social status (Riegle-Crumb, 2006). Given higher education's importance for personal and socio-economic growth, much research focuses on factors affecting academic performance. Grade point average (GPA) is a crucial measure used to assess academic performance. It plays a pivotal role in various aspects, including college admission, scholarships, and career opportunities (Volwerk and Tindal, 2012). While other measures and outcomes may be considered, GPA is often preferred for its simplicity, numerical nature, and comparability. Factors influencing academic performance encompass internal factors, such as a student's innate ability and self-motivation, and external factors, including residential area, ethnicity, and gender (Kudari, 2016).

This study utilizes machine learning to predict student academic performance, focusing on data collected from UDNR students. The study evaluates five algorithms and employs rigorous methodologies such as data sampling, hyperparameter tuning, and model evaluation. The goals of this study are to showcase the predictive capabilities of machine learning in student performance forecasting and to uncover factors shaping student outcomes. Preliminary findings demonstrate machine learning's potential in accurately forecasting academic performance, emphasizing the

influence of past academic performance, study hours, gender, and ethnicity, with further details in subsequent sections.

II. LITERATURE REVIEW

In recent years, machine learning and data mining techniques have gained traction in predicting student academic performance worldwide. Albreiki et al. (2021) conducted a comprehensive review spanning 2009 to 2021, confirming the transformative potential of machine learning in education, emphasizing its role in helping educators understand student progress and intervene early. Similarly, Yakubu and Abubakar (2022) applied machine learning to predict student performance at a Nigerian university, highlighting the influence of factors like gender and high school examination scores. Chen and Ding (2023) found neural networks demonstrating high accuracy and potential in shaping educational policies. Many existing studies focus on specific institutions or regions, highlighting the need for broader analyses encompassing diverse settings like Myanmar.

III. DATA

The dataset, collected between July 12 and July 25, 2018, includes information from 735 students in their third, fourth, and fifth academic years at UDNR. Their GPAs, used as the outcome, were sourced from administrative records. This research aimed to identify factors affecting academic performance, adopting a 'lag' approach using prior GPAs as a predictor for the subsequent year, resulting in a dataset with 1333 observations.

Features include lagged GPA, grade 11 total marks, weekly study hours, ethnicity, gender, religion, residential background, and Basic Education High School location, each providing distinct insights. The feature data is preprocessed, applying z-score normalization to numerical attributes, and employing various encoding strategies for categorical features.

The dataset is complete, with no missing values, enabling credible analyses without data imputation (refer to Table 1 for dataset statistics).

IV. METHODOLOGY

A. Examination Procedures

The procedure comprises four stages to ensure robust, accurate, and reliable models.

1. **Data Sampling**: The dataset is segmented into five subsets using the K-fold cross-validation technique (K=5). This technique ensures that every data point is used for validation exactly once while the remaining data points form the training set.
2. **Hyperparameter Tuning**: Grid Search with K-fold cross-validation identifies optimal hyperparameters for each algorithm. This method considers all possible combinations of

the hyperparameters to find the combination that minimizes the error and improves the prediction performance of the model (see Table 2 for details).

Table 1: Summary Statistics

Variable	Obs.	Mean	Std	Min	Max
GPA	1,333	4.3	0.6	3.0	5.0
GPA_lag	1,333	-0.1	1	-2.0	1.2
TotalmarksGrade11	1,333	-0.1	1	-2.5	2.5
Studyhrperweek	1,333	1.8	1.7	0.0	6.0
ownethpct	1,333	0.0	1	-1.7	2.8
Gender_					
male	1,333	0.3	0.4	0	1
Religion_					
Buddhism	1,333	0.8	0.4	0	1
Christianity	1,333	0.2	0.4	0	1
Other	1,333	0.0	0.1	0	1
Ethnicity_					
Bamar	1,333	0.2	0.4	0	1
Chin	1,333	0.1	0.3	0	1
Kachin	1,333	0.1	0.2	0	1
Kayah	1,333	0.04	0.2	0	1
Kayin	1,333	0.1	0.3	0	1
Mon	1,333	0.02	0.2	0	1
Rakhine	1,333	0.1	0.3	0	1
Shan	1,333	0.3	0.5	0	1
Residential]_					
rural	1,333	0.6	0.5	0	1
suburban	1,333	0.3	0.5	0	1
urban	1,333	0.1	0.3	0	1
LocationofBEHS_					
centeroftown	1,333	0.1	0.3	0	1
isolatedarea	1,333	0.1	0.3	0	1
outskirtoftown	1,333	0.4	0.5	0	1
rural	1,333	0.3	0.5	0	1

Table 2: Hyper-parameters

	Grid search	Choice model
ANN		
- No. of nodes in 1st hidden layers	9, 18, 27	9
- No. of nodes in 2nd hidden layers	4, 9, 18	4,9
SVR		
- C value	1, 5, 10	1
- Gamma value	scale, auto	auto
RF		
- Maximum depth	None, 5, 10	5
- No. of estimators	100, 200, 300	100, 200, 300
GBR		
- Learning rate	0.01, 0.05, 0.1	0.01
- No. of estimators	300, 500, 1000	300, 500
XGBoost		
- Maximum depth	6, 8, 10	6
- No. of estimators	500, 1000, 1500	500, 1000

Note: For the grid search, two parameters for each algorithm were selected for tuning, while suitable values were assigned to other parameters not included in this table.

3. **Model Training**: With optimal hyperparameters, each model undergoes training using cross-validation. Specifically, for each of the five subsets created in the first stage, a model is trained using the four remaining subsets as the training data and the validation data. This process is repeated five times, resulting in five models for each algorithm.

4. **Model Testing**: The saved models are tested using the testing data, which is the one-fold left out in each iteration. Performance evaluation employs three metrics: Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). These metrics enable comprehensive assessment and comparison of the algorithms' performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (3)$$

where y_i represents the observed data, $\hat{y}_i$ represents the predicted data, $\bar{y}$ is the mean of the observed data, n is the total number of observations.

B. Machine Learning Algorithms and Feature Importance

Five algorithms are selected for this study: ANNs, SVR (Cortes and Vapnik, 1995), RF (Breiman, 2001a), GBR (Friedman, 2001), and XGBoost (Chen and Guestrin, 2016). To assess feature significance, Permutation Feature Importance (PFI) (Breiman, 2001b) is employed, evaluating variable importance by shuffling feature values, and measuring their impact on model performance. These variables are then ranked by importance, offering practical insights for enhancing complex models.

V. RESULTS

A. Prediction Performance

Table 3 presents the prediction performance for five algorithms. In the training data, XGBoost outperformed others with MAE of 0.16, RMSE of 0.22, and R2 of 0.88, indicating a superior fit. For the testing data, models showed similar results. SVR stood out with MAE 0.22, RMSE 0.30, and R2 0.77, making it the best-performing model. Figure 1 illustrates actual versus predicted GPA outcomes for all five algorithms.

The promising results of these machine learning models, particularly in terms of predictive accuracy, highlight their potential as tools for early assessment of students at risk of academic underperformance, enabling timely interventions and support strategies to aid in their successful completion of the academic year."

Table 3: Prediction Performance

	MAE	RMSE	R^2
Testing			
ANN	0.223	0.302	0.768
GBR	0.235	0.308	0.759
RF	0.220	0.303	0.767
SVR	0.219	0.301	0.770
XGBoost	0.234	0.315	0.749
Training			
ANN	0.211	0.282	0.799
GBR	0.208	0.267	0.819
RF	0.190	0.258	0.831
SVR	0.199	0.279	0.802
XGBoost	0.163	0.221	0.877

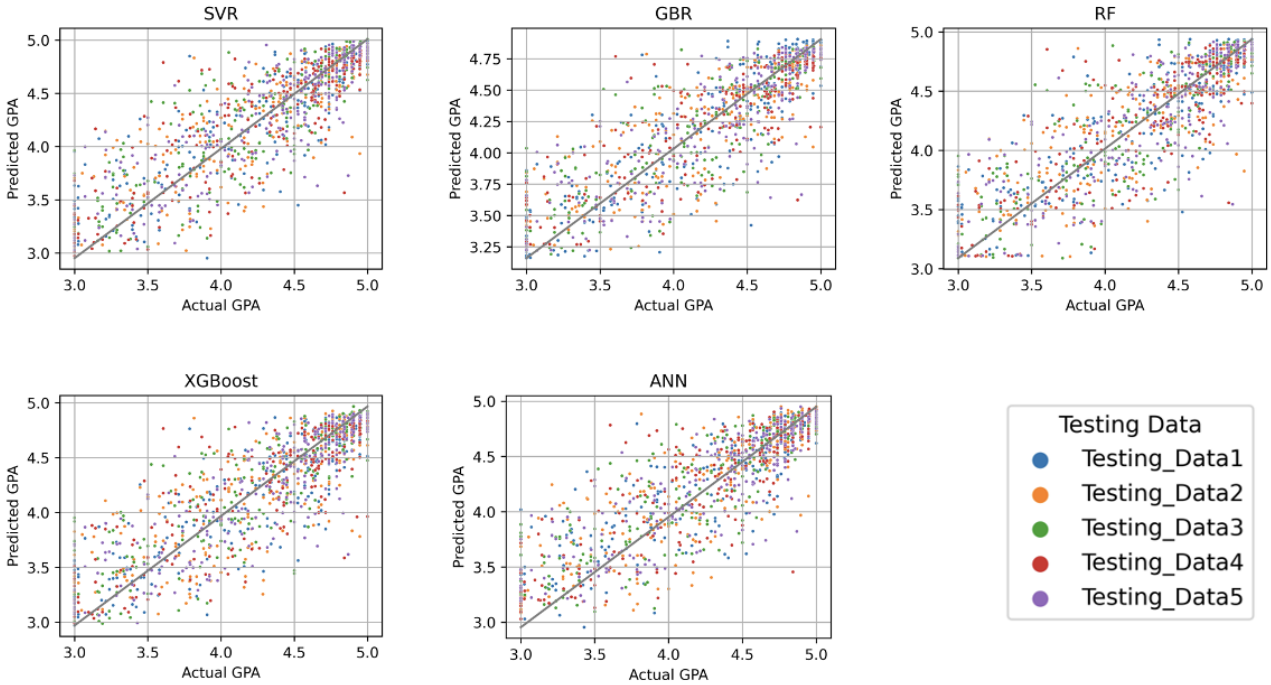

Figure 1: Performance comparason of 5 algorithms on testing data

B. Feature Importance

Feature importance quantifies variables' contributions to the model's predictions, defining their relative utility. Figure 2 displays two PFI plots based on the best-performing model (SVR), outlining crucial elements in the academic performance analysis.

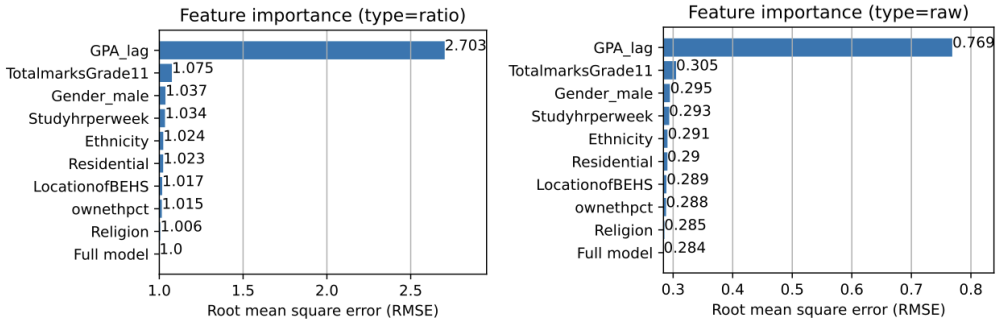

Figure 2: Feature importance for GPA prediction (SVR model)

In the left panel, the RMSE loss after accounting for the first important feature, *GPA_lag*, stands at 2.703, while it is 0.769 in the right panel. This indicates that the RMSE of the model escalated from 0.284 to 0.769, experiencing a 2.7-fold increase following the permutation of the *GPA_lag* variable. Following the *GPA_lag*, the features *TotalmarksGrade*11, *Gender*, *Studyhrperweek*, and *Ethnicity* rank as the second to fifth most important factors in predicting GPA.

It is important to note that analysis of other algorithms provides a similar pattern — the top features identified here were consistently important in other models too. Although rankings varied slightly, these features consistently emerged as key factors in GPA prediction. This alignment underscores their significance in understanding academic performance.

VI. CONCLUSIONS

The research yields valuable insights for predicting student academic performance at UDNR, contributing to data-driven decision-making, and improving educational outcomes. SVR proved proficient in predicting student performance, with the other four algorithms also demonstrating strong predictive capabilities, closely matching SVR's results. This suggests the robust efficacy of machine learning in educational settings, enhancing accurate predictions for informed educational strategies.

The findings reinforce the literature's emphasis on machine learning as a transformative tool in education, particularly in predicting academic performance factors like gender, study hours, and high school examination scores. In line with the study's initial goals, these results contribute to a broader understanding of educational outcomes, aligning with global research trends and underscoring the need for data-driven educational strategies in diverse settings, including Myanmar.

For a more detailed understanding of influencing factors, future research can explore advanced interpretable machine learning methods like Accumulated Local Effects (Apley and Zhu, 2020) and SHapley Additive Explanations (Lundberg and Lee, 2017).

REFERENCES

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82*(4), 1059–1086.

Breiman, L. (2001a). Random forests. *Machine Learning, 45*, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.

Chen, S., & Ding, Y. (2023). A machine learning approach to predicting academic performance in Pennsylvania's schools. *Social Sciences, 12*(3), 118.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Kudari, J. M. (2016). Survey on the factors influencing students' academic performance. *International Journal of Emerging Research in Management & Technology, 5*(6), 30–36.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

Riegle-Crumb, C. (2006). The path through math: Course sequences and academic performance at the intersection of race-ethnicity and gender. *American Journal of Education, 113*(1), 101–122.

Volwerk, J. J., & Tindal, G. (2012). Documenting student performance: An alternative to the traditional calculation of grade point averages. *Journal of College Admission, 216*, 16–23.

Yakubu, M. N., & Abubakar, A. M. (2022). Applying a machine learning approach to predict students' performance in higher educational institutions. *Kybernetes, 51*(2), 916–934.