English Language Teaching; Vol. 16, No. 4; 2023
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Prediction of Students' Performance in English Using Machine Learning Algorithms

Nipa Jun-on[1], Pimpaka Intaros[1] & Sarawut Suwannaut[1]

[1] Department of Mathematics, Faculty of Science, Lampang Rajabhat University, Lampang 52100, Thailand

Correspondence: Sarawut Suwannaut, Department of Mathematics, Faculty of Science, Lampang Rajabhat University, Lampang 52100, Thailand. E-mail: sarawut-suwan@hotmail.co.th

Received: January 1, 2023 Accepted: March 18, 2023 Online Published: March 20, 2023

doi: 10.5539/elt.v16n4p24 URL: https://doi.org/10.5539/elt.v16n4p24

Abstract

In this work, a new machine learning-based model is proposed to predict undergraduate students' reading scores using their listening scores as the primary data. The performance of several machine learning techniques, including neural networks, gaussian process regression, and random forests, was calculated and compared in order to predict the reading test results of the students. The dataset included the listening and reading test results of 1145 students who took the English proficiency exam at Lampang Rajabhat University's language center in Lampang, Thailand. According to the results, the suggested model has a classification accuracy range of 64–75%. Only three different types of parameters—listening scores, departmental data, and faculty data—were used to make the predictions.

Keywords: machine learning algorithms, educational data mining, prediction, English

1. Introduction

Education data mining (EDM) is the use of data mining (DM) techniques on data acquired from various educational systems with the goal of improving education (Baker & Yacef, 2009). Recent EDM research mainly concentrates on assessing students' learning and behavior, analyzing educational tactics and interventions, predicting students' performance and dropout, and giving students tailored suggestions (Romero & Ventura, 2020).

One of the top considerations for educators in assessing educational achievement at all educational levels is students' academic performance (Tan et al., 2019). Educators can develop and execute early interventions and support to improve students' performance when their performance is predicted, especially for students who are at risk of failure (Wakelam et al., 2020). Even while DM is becoming more and more common in educational settings, EDM studies are still quite few and are not well covered, especially in developing nations like Thailand. The limited number of prediction-related EDM research conducted in the Thai education area (Pattanaphanchai et al., 2019; Iam-On & Boongoen, 2017). As a result, relatively little is known about the factors that predict academic achievement for students in this area.

The most common activity in educational data mining to create a prediction model of student performance is classification. To forecast students' performance, a variety of classification approach algorithms were used, including Decision Tree, Neural Network, Naive Bayes, and Support Vector Machine (SVM) (Shahiri et al., 2015). A decision tree is used to find useful information in a small or large data collection. Decision tree algorithm is a popular prediction technique because it is easy to understand. In a given data collection, Naïve Bayes computes a set of probabilities in a given data set (Patil & Sherekar, 2013). A supervised machine learning approach called Support Vector Machine locates the hyperplane in an N-dimensional space that optimally separates the data into two categories (Patil & Sherekar, 2013).

EDM provides new knowledge to educators by identifying hidden patterns in educational data. Many components of the educational system may be evaluated and changed using this method to guarantee the quality of education.

2. Literature Review

Numerous studies have been conducted on forecasting students' performance. The methodology for predicting a student's grade point average at graduation was presented by Tekin (2014). The prediction model was created using three algorithms: neural networks (NN), support vector machines (SVM), and extreme learning machines (ELM). The results showed that the SVM approach consistently makes correct predictions at a rate of 97.98%. Additionally, utilizing decision tree and Naive Bayes approaches, Abu-Oda and El-Halees (2015) suggested a prediction model of students dropping out of university. The information gathered included the students' academic performance, grade point average, and high school GPA. The findings demonstrated that decision tree and Naïve Bayes develop the model with the accuracy of prognosis as 98.14% and 96.86% respectively.

Cruz-Jesus et al. (2020) predicted student academic success using 16 demographic variables, including age, gender, class attendance, internet access, computer ownership, and the number of courses taken. Support vector machines, logistic regression, random forest, and k-nearest neighbors are machine learning algorithms that cloud be used to predict student academic achievement with an accuracy range of 50 to 81%.

Ahmad and Shahzadi (2018) developed a machine learning-based approach to identify students who would struggle academically. Based on the students' study habits, academic interaction behaviors, and learning styles, they created a prediction with an accuracy of 85%. The researchers came to the conclusion that the method they recommended may be used to identify academically unsuccessful students.

At Mae Fah Luang University in Thailand, Iam-On and Boongoen (2017) presented the findings of an investigation on the effectiveness of education data mining approach for identifying first-year students at risk of failure or dropping out as well as variables impacting student success. Additionally, utilizing data from the representative of the major public institution in the southern region of Thailand, Pattanaphanchai, Leelertpanyakul, and Theppalak (2019) developed a model for forecasting students' dropout rates. Data from the Prince of Songkla University's Faculty of Science was gathered for this study during the academic years of 2013 and 2017. The experiment's findings demonstrate that JRip rule induction is the best method for creating prediction models, with a 77.30% accuracy rate.

The study aims at utilizing a small number of scores to predict students' academic performance in English without accounting for socioeconomic or demographic characteristics. The goal of this study was to create a new machine learning algorithm-based model to forecast undergraduate students' reading scores based on their listening grades, faculty, and department.

Machine learning classification algorithms were used to identify the classification algorithms that performed the best in predicting students' academic performance in English. The reading test results of the students were forecast using this methodology. These models will enable the development of new educational strategies and regulations that will enhance students' academic achievement in English. This will help to decrease the number of students who might not succeed based on the assessments made after taking the listening exam.

3. Method

The dataset, pre-processing methods, and machine learning algorithms employed in this work are all described in depth in this section.

3.1 Dataset

The Language Center at Lampang Rajabhat University, Lampang, Thailand, keeps all student records, and a total of 1,145 students participated in this study. Listening (LN) and reading (RD) scores, faculty, and departments of 1,145 students who took the English Exit-Examination in the 2020 academic year were chosen as the dataset in these records. The distribution of each academic unit's total student is shown in Table 1.

Table 1. The dataset

Academic unit	Number of Students
Faculty of Education	217
Faculty of Science	215
Faculty of Industrial Technology	71
Faculty of Science and Agricultural Technology	23
Faculty of Management Sciences	265
Faculty of Humanities and Social Sciences	354

The test was administered in computer labs at Language centre of Lampang Rajabhat University. The labs were equipped with desktop computers and heaphones. Test takers were provided with account details to log in. Before the test, a tutorial video was provided to test takers on how to take the test. First section is listening part

(45 minutes) followed by reading part (75 minutes). In this system, listening and reading score are ranging from 0 to 100.

3.2 Experiments and Results

In order to forecast students' English proficiency, Random Forest (RF), Gaussian Process Regression (GPR), and Neural Networks (NN) were all considered. Utilizing cross validation, the prediction accuracy was assessed. Confusion matrix indicators were used to evaluate the model's performance. According to the previous studies, no one classifier offers the best prediction outcomes. As a result, it's crucial to examine which classifiers have received the most research for the data being analyzed (Asif, 2018, 117-194).

The results of the full set of tests were obtained using MATLAB R2021a. A strong, complete, and user-friendly platform for scientific computation is Matlab. It offers both experienced data scientists and beginners to the field an interactive solution that combines computations and visuals. One of Matlab's most appealing features, referred to as ToolBoxes, are the specific tools it provides for particular areas. The Matlab environment is extended by these ToolBoxes, which are sets of functions that make it possible to address particular sorts of issues.

The dataset included LN score, RD score, Faculty and Department of 1,145 students who taken English test in 2022 academic year. Table 2 shows some part of the dataset.

Table 2. Part of the dataset with 1,145 rows

stdID	Listening	Reading	Faculty	*Department*
std1	2	2	Faculty of Education	Thai Language
std2	3	5	Faculty of Education	Thai Language
std3	5	4	Faculty of Education	Thai Language
…	…	…	…	…
std577	26	28	Faculty of Management Sciences	General Management
std578	35	35	Faculty of Management Sciences	General Management
std579	40	31	Faculty of Management Sciences	General Management
…	…	…	…	…
std1143	15	24	Faculty of Humanities and Social Sciences	Politics and Government
std1144	22	13	Faculty of Humanities and Social Sciences	Politics and Government
std1145	14	26	Faculty of Humanities and Social Sciences	Politics and Government

Based on the CEFR level and student's score range, the following categorical critiria is defined in Table 3. Table 3 shows the criteria used to convert LN and RD scores into the categorical format.

Table 3. Categorical criteria

Category	Criteria
1	score $\leq$ 10
2	11 $\leq$ score $\leq$ 30
3	31 $\leq$ score $\leq$ 40
4	score $\geq$ 41

3.3 Confusion Matrix

The confusion matrix displays the dataset's current situation as well as the number of accurate and inaccurate model predictions. The confusion matrix is displayed in Table 4. The ratio of successfully classified examples to wrongly classified instances serves as a measure of the model's performance. The columns show the model's estimation, while the rows display the actual sample counts in the test set.

The number of cases that were successfully identified is shown in Table 4 under true positive (TP) and true negative (TN). False positive (FP) displays occurrences that are anticipated to be 1 (positive) when they should be 0 (negative). False negative (FN) indicates the proportion of cases that are projected to be 0 (negative) but are really in class 1. (positive).

Table 4. The Confusion matrix

		Predicted	
		Positive (1)	Negative (0)
Actual	Positive (1)	TP	FP
	Negative (0)	FN	TN

The confusion matrix for the RF method is displayed in Table 5. The main diagonal of the confusion matrix, which has four dimensions in all, displays the proportion of instances that were correctly predicted, while the matrix's other sections display the percentage of incorrect predictions.

Table 5 shows that 74.2% of those with the actual RD less than or equal to 10, 77.1% of those with range 11–30, 70.9% of those with range 31–40, and 65.5% of those with greater than or equal to 41 were predicted correctly.

Table 5. Confusion matrix of the RF algorithm

		Predicted				
		≤ 10	11-30	31-40	≥ 41	sum
Actual	≤ 10	**74.2%**	16.4%	15.2%	20.6%	622
	11-30	18.7%	**77.1%**	9.6%	12.7%	390
	31-40	6.8%	5.2%	**70.9%**	1.2%	114
	≥ 41	0.3%	1.3%	4.3%	**65.5%**	19
	sum	710	289	131	15	1145

Table 6 shows the confusion matrix for the GPR algorithm. Table 6 shows that 70.8% of those with the actual RD less than or equal to 10, 79.7% of those with range 11–30, 55.4% of those with range 31–40, and 42.6% of those with greater than or equal to 41 were predicted correctly.

Table 6. Confusion matrix of the GPR algorithm

		Predicted				
		≤ 10	11-30	31-40	≥ 41	sum
Actual	≤ 10	**70.8%**	15.3%	37.3%	29.5%	622
	11-30	12.6%	**79.7%**	6.8%	14.6%	390
	31-40	9.4%	4.2%	**55.4%**	13.3%	114
	≥ 41	7.2%	0.8%	0.5%	**42.6%**	19
	sum	682	325	120	18	1145

Table 7 shows the confusion matrix for the NN algorithm. Table 7 shows that 65.7% of those with the actual RD less than or equal to 10, 68.2% of those with range 11–30, 46.2% of those with range 31–40, and 41.8% of those with greater than or equal to 41 were predicted correctly.

3.4 Classification Accuracy

Classification accuracy: CA is the ratio of the correct predictions (TP+TN) to the total number of instances (TP+TN+FP+FN).

$$\text{Accuracy} = \frac{TN + TP}{FN + TN + TP + FP}$$

Table 7. Confusion matrix of the NN algorithm

		Predicted				
		≤ 10	11-30	31-40	≥ 41	sum
Actual	≤ 10	**65.7%**	24.9%	32.4%	49.2%	622
	11-30	18.7%	**68.2%**	15.9	6.1%	390
	31-40	9.4%	4.7%	**46.2%**	2.9%	114
	≥ 41	6.2%	2.2%	5.5%	**41.8%**	19
	sum	649	282	179	35	1145

The classification accuracy of the RF, GPR and NN algorithms were also 0.745, 0.716 and 0.640 respectively. For instance, the RF algorithm was able to attain 74.5% accuracy in accordance with these results. In other words, the correlation between the predicted data and real data was quite strong. Hence, 74.5% of the samples were properly classified.

Table 8. Classification accuracy of the models

Models	**Classification accuracy (CA)**
Random Forest (RF)	74.5%
Gaussian Process Regression (GPR)	71.6%
Neural Networks (NN)	64.0%

4. Discussion and Conclusion

This study suggests a new machine learning-based model to forecast undergraduate students' reading scores using their listening scores as the primary data. To forecast the students' reading scores, the performances of the machine learning algorithms Random Forests, Gaussian Process Regression, and Neural Networks were computed and compared. Two criteria were the main focus of this investigation. Prediction of English performance based on previous achievement scores was the first parameter. The second one was the comparison of performance indicators of machine learning algorithms.

According to the findings, the suggested model has a classification accuracy range of 64–75%. This finding implies that students' listening test results are a significant predictor to be used in predicting their reading test results. Algorithms with a very high accuracy rate that may be used to forecast students' reading scores include RF, GPR, and NN. Additionally, just three types of parameters—listening scores, department data, and faculty data—were used to make the predictions.

Another result of this study showed that RF and GPR algorithms had the highest classification accuracy, whereas NN algorithms have the lowest. This finding suggests that RF and GPR algorithms perform more accurately when forecasting students' reading scores using machine learning techniques.

The suggested model has a 74.5% accuracy rate in predicting students' reading test results. This finding implies that this model coule be used to forecast students' English proficiency in the future. Students could be given the opportunity to examine their working procedures and enhance their performance if listening and reading tests are separated for a month, for instance.

The outcomes show that it is possible to predict students' English performance using machine learning techniques. More significantly, the prediction was just based on the listening test results, faculty, and department. Other factors can be used as input variables and other machine learning algorithms added to the modeling process to perform future research.

Acknowledgments

This research paper is supported by Lampang Rajabhat University.

References

Abu-Oda, G. & El-Halees, A. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process, 5*(1), 15-27. https://doi.org/10.5121/ijdkp.2015.5102

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

Asif, R., Merceron, A., Sli, S.A. & Haider, N.G. (2018). Analyzing undergraduate students' performance using educational data mining. *Computer and Education, 113*, 117-194. https://doi.org/10.1016/j.compedu.2017.05.007

Baker, R. S. J. D. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

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

Iam-On, N. & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics, 8*(2), 497-508. https://doi.org/10.1007/s13042-015-0341-x

Patil, T. & Sherekar, S. (2013). Performance Analysis of Naive Bayes and J48, *International Journal of Computer Science and Applications, 6*(2), 256-261.

Pattanaphanchai, J., Leelertpanyakul, K. & Theppalak, N. (2019). The investigation of student dropout prediction model in Thai higher education using educational data mining: A case study of faculty of science, Prince of Songkla University. *Journal of University of Babylon for Pure and Applied Sciences, 27*(1), 356–367. https://doi.org/10.29196/jubpas.v27i1.2191

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. https://doi.org/10.1002/widm.1355

Shahiri, A. M., Husain, W. & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *In Procedia Computer Science, 72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Tan, R. Z., Wang, P. C., Lim, W. H., Ong, S. H. C. & Avnit, K. (2019). Early prediction of students' mathematics performance. *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, 651–656. https://doi.org/10.1109/TALE.2018.8615289

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research, 54*, 207-226. https://doi.org/10.14689/ejer.2014.54.12

Wakelam, E., Jefferies, A., Davey, N. & Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology, 51*(2), 347–370. https://doi.org/10.1111/bjet.12836

Copyrights