engineering proceedings

MDPI

Proceeding Paper

Predicting the Learning Performance of Minority Students in a Vietnamese High School Using Artificial Intelligence Algorithms †

Hai-Duy Le [1], Thao-Trang Huynh-Cam [2,*], Long-Sheng Chen [3], Vo Phan Thu Ngan [4] and Tzu-Chuen Lu [2]

[1] Thuan Hoa High School, Hue City 952410, Vietnam; lehaiduy.c3th@soctrang.edu.vn
[2] Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan; tclu@cyut.edu.tw
[3] Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei 106344, Taiwan; lschen@ntut.edu.tw
[4] Foreign Languages Faculty, Dong Thap University, Cao Lanh City 81118, Vietnam; vptngan@dthu.edu.vn
* Correspondence: hcttrang@dthu.edu.vn
† Presented at the 2024 4th International Conference on Social Sciences and Intelligence Management (SSIM 2024), Taichung, Taiwan, 20–22 December 2024.

Abstract

This study aims to predict and discover important factors for the learning performance of students belonging to two ethnic groups—Khmer and Chinese (Hoa) students—in Soc Trang with the use of random forest (RF) and Gaussian Naïve Bayes (GNB) classifiers based on students' demographics and grade point average (GPA) scores. The study involved 174 Khmer and Chinese (Hoa) students in Grade 10 in a high school in Soc Trang Province, Vietnam. The results showed that, for Khmer students, GNB was better than RF, with an F1 score of 100%. Mathematics was the most important subject leading Khmer students to very good or poor performance. For Chinese (Hoa) students, both classifiers showed the same accuracy performance. Scores in Literature and English in Semester 1 impacted Chinese (Hoa) students' performance. The results of this study provide a reference for formulating a policy to improve the learning performance of minority students to prevent dropouts.

Keywords: learning performance; learning performance predictions; high-school minority students; Khmer students; Chinese (Hoa) students; artificial intelligence algorithms; Soc Trang Province of Vietnam

check for updates

Academic Editors: Teen-Hang Meen, Liza Lee and Cheng-Fu Yang

Published: 27 June 2025

Citation: Le, H.-D.; Huynh-Cam, T.-T.; Chen, L.-S.; Ngan, V.P.T.; Lu, T.-C. Predicting the Learning Performance of Minority Students in a Vietnamese High School Using Artificial Intelligence Algorithms. *Eng. Proc.* **2025**, *98*, 22. https://doi.org/10.3390/engproc2025098022

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soc Trang Province in Vietnam currently has 27 ethnic minorities living within it, with a total population of 424,914, accounting for more than 35% of the province's population. Among them, the Khmer ethnic community is the largest group, accounting for 30.19% of the population; Chinese (Hoa) people account for 5.22% of the population, and other minorities account for 0.04% [1]. Caring for the lives of ethnic minorities, in particular, the Khmer and Chinese (Hoa) ethnic minorities, is a significant policy of the Soc Trang Provincial Government. Accordingly, the education sector, with a large Khmer population, has received special attention from state and local authorities. Overall, 39.3% of the provincial budget (over VND 3500 billion) was allocated toward education in 2023 and 2024 [2].

However, in the 2023–2024 academic year, the rate of students classified as "Poor (underqualified)" in Soc Trang was 1.44% (with 2.05% in Grade 10 and 0.77% in Grade

11). Notably, the failure rate in Grade 10 was nearly three times higher than that in Grade 11. Also, the drop-out rate was 0.77% as a result of poor learning performance and other factors. Consequently, it is important to predict and discover important factors for learning performance so that teachers and school administrators can make early interventions, safeguard students' learning progress, and provide them with necessary support. However, prediction of the learning performance of Khmer and Chinese (Hoa) students in Soc Trang Province has not been performed. Important factors associated with the performance of these ethnic minority students also remain unknown.

Predicting student performance in academic institutions using an artificial intelligence (AI) algorithm is a practice that has been researched for the past decade. For instance, Zafari et al. [3] developed a machine learning (ML)-based prediction algorithm using random forest (RF), support vector machines (SVMs), logistic regression (LR), and artificial neural networks (ANNs). A dataset encompassing demographic, behavioral, and academic data collected through surveys and teacher input was used to determine influential factors affecting Iranian high school student performance and to identify implications. Zeineddine et al. [4] proposed the use of automated ML (AutoML) to predict student success in higher education. They explored various AI models and developed a model that achieved 75.9% accuracy in predicting student performance based on pre-admission data. The results proved that AutoML was a valuable tool for enhancing early interventions to support student success. Correspondingly, six common AI algorithms, namely decision tree (DT), SVM, Naïve Bayes, K-nearest neighbor (KNN), LR, and RF algorithms, were used to predict Malaysian student grades, of which RF was the most effective algorithm, achieving a high F1 score of 99.5% [5].

In terms of the optimal method for predicting undergraduate student academic performance at graduation, various ML algorithms were used in different prediction timeframes, and two approaches using individual course grades and grade point averages (GPAs) were compared [6]. Individual course grades were found to be more accurate data points for early predictions, while GPAs became more useful in the advanced academic program.

In Vietnam, Duong et al. [7] used various AI algorithms including a SVM and a light gradient boosting machine (LightGBM) to develop a two-stage academic performance warning system for Vietnamese universities. The results achieved high accuracy (an F1 score of over 74% at the semester's start and over 92% before final exams). The algorithms enabled early warnings and timely interventions for at-risk students. Recently, Quynh et al. [8] used models based on RF, XGBoost, SVM, and Voting algorithms to predict how well students did in a pre-English course at Vietnam National University's International School. They compared the performance of different models and discovered that an ensemble model, specifically the Voting algorithm, was the most accurate in predicting student performance, achieving 75.9% accuracy based on information gathered before admission. Specifically, in the Mekong Delta in 2023, RF, XGBoost, and LightGBM models were employed on a dataset of over 21,000 student records from Ca Mau province [9]. Key predictors, including GPAs, age, class, and parents' careers, influenced academic success, and enabled interventions to reduce high school dropout rates. In this study, RF achieved the highest accuracy (81.69%).

Considering previous research results, we used an RF classifier and GNB to predict and discover important factors affecting the performance of students belonging to the two most prevalent ethnic groups—Khmer and Chinese (Hoa) students—in Soc Trang Province, Vietnam, where researchers have shown interest in their research.

2. Materials and Methods

Implementation Process

The implementation process is shown in Figure 1.

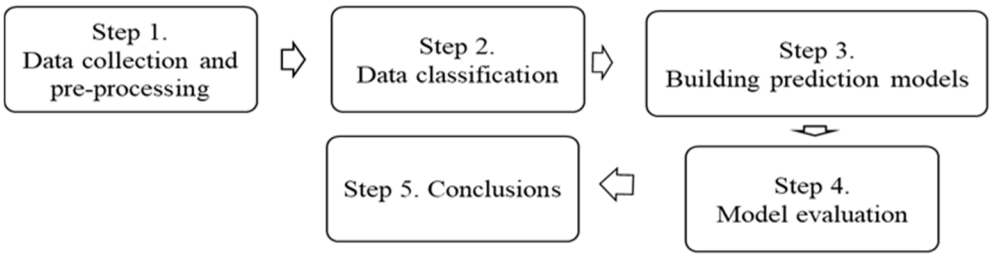
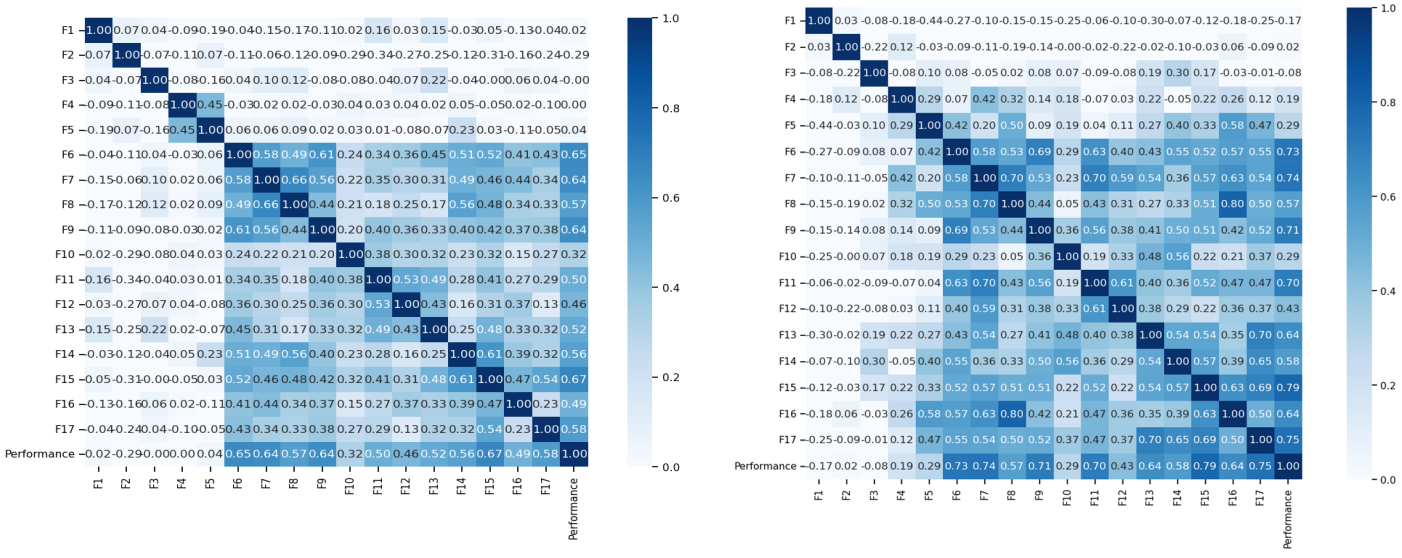


Figure 1. Process used for implementation of algorithms in this study.

- Step 1: Data collection and preprocessing

The data were collected from the school database of a high school in Soc Trang Province, Vietnam, at the end of the 2023–2024 school year. Students' identities were anonymized for ethical purposes. After data collection, Microsoft Excel 2016 software was used for preprocessing. Since we explored the learning performance of students belonging to the two most prevalent ethnic groups—Khmer and Chinese (Hoa) students—the remaining students' data were removed. During this study, no students dropped out or were suspended. The final data used for prediction included 174 Grade 10 students. These students were grouped into two groups. Group 1 included 147 (84.5%) Khmer students, while Group 2 had 27 (15.5%) Chinese (Hoa) students. Demographic information and Mathematics, Literature, and English grades were used as the input data. The output was learning performance. Grades were transformed into numeric values for data processing. Table 1 describes the input and output data and their transformation. Figure 2 shows the correlation matrix among them.



(**a**) (**b**)

Figure 2. Correlation matrix of students of two minority groups: (**a**) Khmer students; (**b**) Chinese (Hoa) students.

Table 1. Input and output data.

Factor ID	Factor Description	Values and Transformed Values	
		Khmer Students	**Chinese (Hoa) Students**
F1	Gender	1 = Girls; 2 = Boys	
F2	Age	1 = 16; 2 = 17; 3 = 18	
F3	Home address	1 = Town; 2 = Rural; 3 = Remote rural	
F4	Father's career	0 = Not applicable; 1 = Agricultural worker; 2 = Teacher; 3 = Laborer; 4 = Truck driver; 5 = Officer; 6 = Manager; 7 = Musician; 8 = Small retailer; 9 = Freelancer; 10 = Other	0 = Not applicable; 1 = Agricultural worker; 2 = Laborer; 3 = Truck driver; 4 = Engineer; 5 = Small retailer; 6 = Freelancer
F5	Mother's career	0 = Not applicable; 1 = Agricultural worker; 2 = Teacher; 3 = Laborer; 4 = Officer; 5 = Photographer; 6 = Small retailer; 7 = Housewife; 8 = Freelancer	0 = Not applicable; 1 = Agricultural worker; 2 = Laborer; 3 = Hairdresser; 4 = Small retailer; 5 = Housewife
F6	Math score for mid-term test_Semester 1	1.4~9.8	1.8~8.8
F7	Math score for final exam_Semeter 1	1.4~10	1.8~9.2
F8	Math score for mid-term test_Semester 2	1.8~10	2.8~10
F9	Math score for final exam_Semeter 2	1.3~10	1.8~9.6
F10	Literature score for mid-term test_Semester 1	3.0~9.0	3.0~8.0
F11	Literature score for final exam_Semeter 1	3.0~9.0	4.0~8.5
F12	Literature score for mid-term test_Semester 2	2.0~10	4.5~8.5
F13	Literature score for final exam_Semeter 2	3.0~9.0	3.0~8.5
F14	English score for mid-term test_Semester 1	2.3~9.5	2.7~7.3
F15	English score for final exam_Semester 1	2.2~9.1	3.4~9.0
F16	English score for mid-term test_Semester 2	3.5~9.8	5.0~10
F17	English score for final exam_Semester 2	2.8~9.5	4.8~8.8
Output	Learning performance	1 = Very good (VG); 2 = Good (G); 3 = Average (AVG); 4 = Poor	

- Step 2: Data classification

For better accuracy, we used "learning performance" as a class label to distribute the datasets into three classification cases, adapted from Huynh-Cam et al. [10]. Case 1 used origin datasets, while Case 2 combined "AVG" and "Poor" to create a new class,

"Poor". Case 3 was created by removing the class "G" and included "VG" and "poor" for comparison. Table 2 describes the classification results in detail.

Table 2. Output data classification.

Cases	Transformed Values	Output Factor	Scores	Number of Samples		Classification
				Khmer Students (number = 147)	Chinese (Hoa) Students (number = 27)	
Case 1: Origin	4	Very good (VG)	8.0–10 points	13 (12.1%)	04 (14.8%)	4 classes: VG, G, AVG, and poor
	3	Good (G)	6.5- 7.9 points	68 (63.8%)	13 (48.2%)	
	2	Average (AVG)	5.0–6.4 points	62 (57.9%)	9 (33.3%)	
	1	Poor	3.5–4.9 points	04 (3.7%)	01 (3.7%)	
Case 2: Combine	3	VG	8.0–10 points	13 (12.1%)	04 (14.8%)	3 classes: VG, G, and poor
	2	G	6.5- 7.9 points	68 (63.8%)	13 (48.2%)	
	1	Poor	3.5–6.4 points	68 (63.8%)	10 (37%)	
Case 3: Focus	1	VG	8.0–10 points	13 (12.1%)	04 (14.8%)	2 classes: VG and poor
	0	Poor	3.5–6.4 points	68 (63.8%)	10 (37%)	

- Step 3: Building prediction models

We built prediction models on Windows Operating Systems with a 3.80 GHz Intel(R) Xeon(R) E-2174G CPU and 64 GB of RAM. RF and GNB classifiers were used to build a prediction model for three classification cases (Table 2) based on the Jupyter Notebook tool, version 6.5.4, with Scikit-learn packages in Python 3 language. The dataset, in each case, was divided into training and testing data in a ratio of 80:20. Each classifier was used to construct prediction models. The mean values and standard deviation (SD) of five datasets in each classification case were used to compare the prediction performance between the results using two classifiers.

- Step 4: Evaluation

We used accuracy and F1 scores to evaluate model performance. After comparing the classification results of the three cases, we selected the best case to retrieve factors.

- Step 5: Conclusions

In this step, the research results were summarized. Several solutions for enhancing the learning performance of Khmer and Chinese (Hoa) students were proposed based on the results.

3. Results

3.1. Classification

We included data pertaining to very good and poor learning performances for Khmer and Chinese (Hoa) students. Therefore, the data of the two classes of "VG" and "Poor" were analyzed in this study. Table 3 presents the results of the three data classification cases for the two student groups. The accuracy and F1-Score for Case 3 were higher than those in Cases 1 and 2. This means that Case 3 was better than Cases 1 and 2.

In Case 1, although the accuracy of the two classifiers was high, the F1 scores were zero. This low performance was attributed to there being too many initial classes (VG, G, AVG, and Poor). Hence, we combined "AVG" and "Poor" to create a new class, "Poor", in Case 2, as the data were similar.

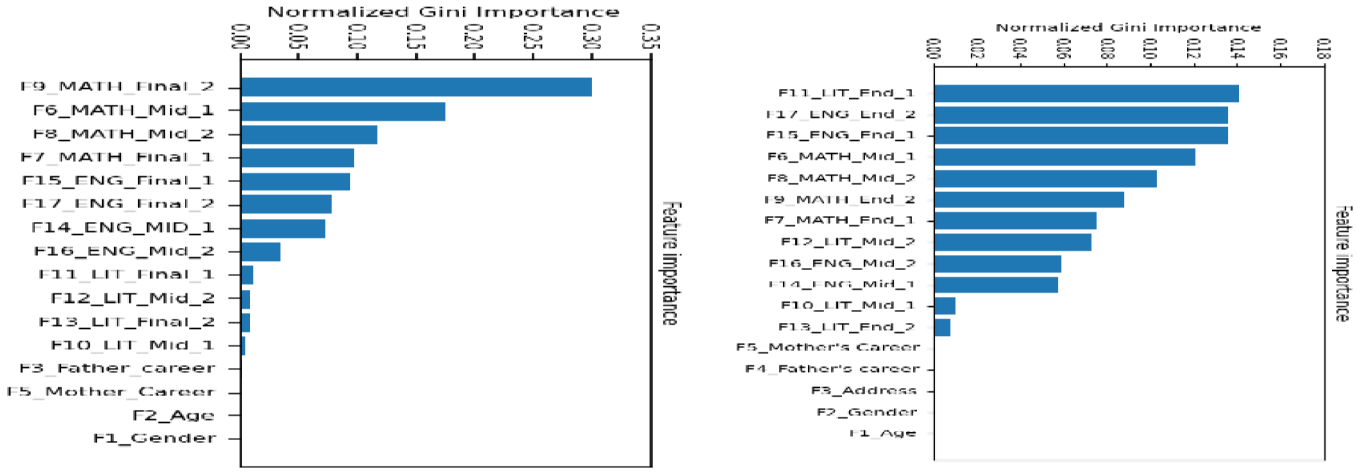
Table 3. Classification results.

Group	Classification Case	Classifier	Accuracy (%)						F1 score (%)					
			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean (SD)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean (SD)
Khmer students	Case 1: Origin	RF	77	77	83	87	80	80.8 (4.27)	00	00	00	00	00	0.0 (0.0)
		GNB	80	80	87	90	80	83.4 (4.77)	00	00	00	00	00	0.0 (0.0)
	Case 2: Combine	RF	83	87	83	77	70	80.0 (6.63)	89	88	89	79	81	85.2 (4.82)
		GNB	90	90	87	73	83	84.6 (7.09)	89	89	85	84	87	86.8 (2.28)
	Case 3: Focus	RF	100	94	100	100	100	98.8 (2.68)	100	96	100	100	100	99.2 (1.79)
		GNB	100	100	100	100	100	100.0 (0.00)	100	100	100	100	100	100.0 (0.00)
Chinese (Hoa) students	Case 1: Origin	RF	67	67	56	78	56	64.8 (9.2)	00	00	00	00	00	0.0 (0.0)
		GNB	83	78	78	67	67	74.6 (7.23)	00	00	00	00	00	0.0 (0.0)
	Case 2: Combine	RF	67	83	83	100	67	80.0 (13.75)	50	67	67	100	00	56.8 (36.56)
		GNB	67	33	67	83	50	60.0 (19.08)	00	00	67	00	00	13.4 (29.96)
	Case 3: Focus	RF	100	100	100	100	100	100.0 (0.00)	100	100	100	100	100	100.0 (0.00)
		GNB	100	100	100	100	100	100.0 (0.00)	100	100	100	100	100	100.0 (0.00)

In Case 2, for Khmer students, the two classifiers performed better, with the F1 score increasing from 0.0 to 85.2% (SD = 4.82) for the RF classifier and to 86.8% (SD = 2.28) for the GNB classifier. For Chinese (Hoa) students, although the accuracy was 80% (SD = 13.75) for the RF classifier and 60% (SD = 19.08) for the GNB classifier, the F1 score of the classifiers was very low. This indicated that the two classifiers could not predict students with very good performance or poor performance for Chinese (Hoa) students. Therefore, in Case 3, we removed the "G" class and only included the "VG" and "poor" classes to obtain higher accuracy. In Case 3, for Khmer students, the GNB classifier predicted better than the RF classifier, with an F1 score of 100%. For Chinese (Hoa) students, the two classifiers showed the same F1 score of 100%. The prediction models built in Case 3 correctly predicted very good and poor performance of minority students. Thus, Case 3 was used for feature selection.

3.2. Feature Selection

Figure 3 displays the rankings of feature importance between two groups of students extracted from Case 3. Mathematics was the most important factor leading Khmer students to very good or poor performance. Literature and English in Semester 1 significantly impacted Chinese (Hoa) students' performance. Age, gender, and family background did not affect minority students' performance.



(**a**) (**b**)

Figure 3. Importance feature ranking: (**a**) Khmer students; (**b**) Chinese (Hoa) students.

4. Conclusions

Table 4 lists the most important factors impacting the performance of minority students in Grade 10. For Khmer students, Mathematics most clearly affected their learning performance, whereas for Chinese (Hoa) students, Literature and English in Semester 1 most clearly affected their learning performance.

Students might not have foundational knowledge in Mathematics and Literature, stemming from middle school, and therefore, when they enter high school, they are overwhelmed by the amount of knowledge they need to learn, which leads to them experiencing difficulty in learning and potentially falling behind. Therefore, support from teachers and tutors to help students review basic concepts is necessary.

Table 4. Ranking of most important factors for minority student performance.

Rank Order	Factors	
	Khmer Students	**Chinese (Hoa) Students**
1	F9. Math score for final exam_Semeter 2	F11. Literature score for final exam_Semeter 1
2	F6. Math score for mid-term test_Semester 1	F17. English score for final exam_Semester 2
3	F8. Math score for mid-term test_Semester 2	F15. English score for final exam_Semester 1
4	F7. Math score for final exam_Semeter 1	F6. Math score for mid-term test_Semester 1
5	F15. English score for final exam_Semester 1	F8. Math score for mid-term test_Semester 2

Regarding English, students in Soc Trang Province, especially those from rural areas, have very limited chances to communicate with foreigners, and they are still shy when speaking in English. Hence, group study must be promoted more in high schools. In the first grade in high schools, teaching methods and learning materials are new. The majority of students are accustomed to traditional learning methods, focusing extensively on vocabulary and grammar or reading and writing skills (for exams and tests) rather than communication skills (listening and speaking). Many students still underestimate the importance of English, as they believe that, in the future, they will not need to use it or will only use it sparingly in their daily work.

School administrators must create more ongoing, practical professional development opportunities to help teachers stay updated on the best practices and innovative teaching methods. Moreover, it is required to allocate sufficient resources, including textbooks, technology, and instructional materials to support teachers and students. They must pay more attention to those who cannot afford tuition fees or learning materials and organize extracurricular activities to help students identify their study styles and improve learning performances.

Minority student learning performance in Vietnamese high schools in the Mekong Delta region has received much attention. In this study, learning performance prediction of Khmer and Chinese (Hoa) students in Grade 10 in a high school in Soc Trang Province using RF and Gaussian NB classifiers showed that, for Khmer students, the GNB classifier predicted better than the RF classifier, with an F1 score of 100%. Mathematics was the most important factor affecting Khmer students' learning performance. For Chinese (Hoa) students, both classifiers showed the same performance. Literature and English in Semester 1 highly impacted Chinese (Hoa) students' performance.

Although this pilot study contributes means to better predict minority students' performance in high schools, it still has limitations. Firstly, we considered demographic information and scores in Mathematics, Literature, and English as the input factors. Other factors need to be included for better predictions and for comparison with the results of this study. Secondly, the research data in this study were obtained from a single high school in one province in the Mekong Delta region, so the results might not be generalizable to other high school students.

Author Contributions: Conceptualization, H.-D.L., T.-T.H.-C. and L.-S.C.; methodology, T.-T.H.-C. and L.-S.C.; software, T.-T.H.-C.; validation, T.-T.H.-C., T.-C.L. and L.-S.C.; formal analysis, H.-D.L. and T.-T.H.-C.; investigation, H.-D.L. and V.P.T.N.; resources, H.-D.L.; data curation, T.-T.H.-C. and H.-D.L.; writing—original draft preparation, H.-D.L. and T.-T.H.-C.; writing—review and editing, L.-S.C. and V.P.T.N.; visualization, V.P.T.N.; supervision, L.-S.C.; project administration, L.-S.C. and T.-C.L.; funding acquisition, V.P.T.N. and T.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The need for ethical review and approval was waived for this study due to it being a low-risk study. Any risk suffered by the research subjects is not higher than those who did not participate in the study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is unavailable due to privacy.

Acknowledgments: We are grateful Thuan Hoa High School in Chau Thanh Town, Soc Trang Province, Vietnam; Dong Thap University, Vietnam; and Chaoyang University of Technology, Taiwan for providing access to their facilities, which allow us to conduct the experiments reported in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Available online: https://baodantoc.vn/dai-hoi-dai-bieu-cac-dtts-tinh-soc-trang-lan-thu-iv-nam-2024-thanh-cong-tot-dep-1723959639108.htm (accessed on 20 October 2024).
2. Regulations on Principles, Criteria, and Norms for Allocation of Central Budget Capital and Ratio of Counterfeit Capital of Local Budget to Implement the National Target Program for Socio-Economic Development in Ethnic Minority and Mountainous Areas for the Period 2021–2030, Phase I: 2021–2025. Available online: https://thuvienphapluat.vn/van-ban/Dau-tu/Quyet-dinh-39-2021-QD-TTg-nguyen-tac-tieu-chi-phan-bo-von-ngan-sach-trung-uong-499362.aspx (accessed on 20 October 2024).
3. Zafari, M.; Sadeghi-Niaraki, A.; Choi, S.M.; Esmaeily, A. A practical model for the evaluation of high school student performance based on machine learning. *App. Sci.* **2021**, *11*, 11534. [CrossRef]
4. Zeineddine, H.; Braendle, U.; Farahm, A. Enhancing prediction of student success: Automated machine learning approach. *Comput. Electr. Eng.* **2020**, *89*, 106903. [CrossRef]
5. Bujang, S.D.A.; Selamat, A.; Ibrahim, R.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H.; Ghani, N.A.M. Multiclass prediction model for student grade prediction using machine learning. *IEEE Access* **2021**, *9*, 95608–95621. [CrossRef]
6. Tatar, A.E.; Düştegör, D. Prediction of academic performance at undergraduate graduation: Course grades or Grade point average? *App. Sci.* **2020**, *10*, 4967. [CrossRef]
7. Duong, H.T.H.; Tran, L.T.M.; To, H.Q.; Van Nguyen, K. Academic performance warning system based on data driven for higher education. *Neural Comput. Appl.* **2022**, *35*, 5819–5837. [CrossRef] [PubMed]
8. Dinh-Thanh, N.; Thi-Ngoc-Diem, P. Predicting academic performance of high school students. In *Nature of Computation and Communication. ICTCC 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*; Phan, C.V., Nguyen, T.D., Eds.; Springer: Cham, Switzerland, 2023; Volume 473. [CrossRef]
9. Quynh, T.D.; Dong, N.D.; Thuan, N.Q. A case study of student performance predictions in English course: The data mining approach. In *International Congress on Information and Communication Technology*; Springer Nature: Singapore, 2024; pp. 419–429.
10. Huynh-Cam, T.T.; Chen, L.S.; Le, H. Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. *Algorithms* **2021**, *14*, 318. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.