MACHINE LEARNING APPLICATIONS IN GRADUATION PREDICTION AT THE ACADEMY OF FINANCE, VIETNAM

PhD. Cu Thu Thuy* - Nguyen Duc Thanh**

Abstract: ***This study investigates the application of supervised machine learning methods for predicting student learning outcomes based on academic performance in previous semesters. The experimental data comprises of 2344 graduated from the CQ56 and CQ57 (2022 and 2023) cohorts across all majors at the Academy of Finance. The result indicates that J48 decision tree algorithm achieved the highest prediction accuracy for graduation classification, both with the original data and the preprocessed data that addresses missing values. This preliminary study demonstrates the potential for an effective application of machine learning to student data mining at the Academy of Finance.***

• Keywords: *machine learning, predicting student learning outcomes, educational data mining.*

Date of receipt: 04th Dec., 2024
Date of receipt revision: 10th Feb., 2025
Date of delivery revision: 30th Dec., 2024
Date of approval: 18th Mar., 2025
DOI: https://doi.org/10.71374/jfar.v25.i2.09

1. Introduction

Nowadays, higher education institutions are facing intense competition due to the changing of the education market, as well as demands from learners and society. To thrive, they must improve on various aspects, with student learning quality being a crucial factor. Minimizing student attrition and improving academic performance are common goals shared across institutions. Beyond academic rules and regulations, analysis of prior academic performance could assist in prediction of students' final learning outcomes. These results could provide students with an effectively guide toward improving their learning strategies and achieving better results in the future.

Educational data mining has gathered significant attention from researchers and developers as it becomes essential for institutions seeking to enhance educational quality based on innovative techs (Sumitha et al., 2016). Predicting student learning outcomes is challenging due to its multi-faceted nature, however this is the central issue to create breakthroughts. Various methods have been employed to generate these predictions. Recently, Machine Learning (ML) has been used to develop models utilizing big data to support student academic performance prediction and decision-making in educational settings (Alalawi et al., 2023).

Currently, the application of machine learning models to predict student learning outcomes at the Academy of Finance remains unexplored. This study aims to address the question of which methods are suitable for this purpose and provide empirical evidence for their reliability. Several machine learning models are applied to predict graduation outcomes at the Academy of Finance based on students' prior academic data.

The remainder of this paper is structured as follows: Section 2 reviews relevant literature; Section 3 presents the experimental results from applying machine learning methods to student data at the Academy of Finance; and the Conclusion summarizes the findings, limitations, and future research directions.

2. Literature review

In recent years, extensive research have been conducted on the application of machine learning to predict student learning outcomes, identify at-risk students, and forecast dropout rates. Key studies include:

Khalid et al. (2023) conducted a systematic literature review on the employment of Machine Learning (ML) to predict student learning outcomes in Educational Data Mining (EDM), covering publications from January 2010 to October 2022. Over 50 distinct ML algorithms were identified across 162 papers. Three primary ML methods for predicting student learning outcomes are classification, regression, and clustering, with classification being the most prevalent, accounting for 80% of applications. Regression and clustering follow, representing 14.6% and 5.2%, respectively. The most employed algorithms include Decision Tree, Random Forest, Naive Bayes (NB), Artificial Neural Networks (ANN), and Support Vector Machine (SVM). Based on assessment of training time complexity, Naive Bayes and Decision Tree are identified as the least computationally demanding algorithms, which may explain their widespread adoption.

The comprehensive research conducted by Hellas et al. (2018) explores the question, "What is the most advanced technology currently used in predicting students' learning outcomes?" Decision Trees, Neural Networks, and SVM were found to be highly prevalent.

Experimental results from specific data sets in recent years are summarized as follows:

Sumitha et al. (2016) predicted student grades (Excellent, Good, Average, Below Average) based on

** Academy of Finance; email: cuthuthuy@hvtc.edu.vn*
*** Newton Grammar School; email: thanh.nguyenduc247@gmail.com*

demographic information, school details, attendance records, CGPA, and results from the previous semester for approximately 350 students from KLN College of Information Technology (affiliated with Anna University). The results showed that J48 algorithm achieved the highest accuracy rate at 97.27%.

Yaacob et al. (2019) analyzed a dataset consists of 631 graduates from 2013 to 2016 from the Statistics Department at Universiti Teknologi MARA Cawangan Kelantan and Universiti Teknologi MARA Cawangan Negeri Sembilan. The Naive Bayes algorithm achieved the highest prediction accuracy (89.26%), compared to Decision Tree, k-Nearest Neighbors (k-NN), and Logistic Regression.

Bujang et al. (2021) employed multiple machine learning models, including J48, Logistic Regression, Random Forest, NB, k-NN, and SVM, to predict student learning outcomes. The dataset comprised of 641 students from the Information Technology and Media Department at Malaysia Polytechnic University. The study compared six models and found that J48 and Random Forest achieved the highest accuracy rate of 98.9%. Applying the Synthetic Minority Over-sampling Technique (SMOTE) further improved the accuracy of all six models to 99.5%.

N. G. Cu et al. (2024) improved performance of machine learning algorithm such as J48, Logistic Regression, Random Forest, MLP, k-NN, and SVM on prediction of at-risk students by a novel data rebalancing technique.

In addition to utilizing data from the institutional information system, some researchers have also conducted surveys to gather direct feedback from students.

In Vietnam, numerous researchers have utilized machine learning models to predict students' academic performance. Uyen and Tam (2019) analyzed data from 555 students across the 54th, 55th, and 56th cohorts of the Information Technology Department at Vinh University. Similarly, Son et al. (2022) recommended employing machine learning models for early predictions of academic performance based on education and enrollment data from Hanoi Metropolitan University. More recently, Thuy (2023) analyzed a dataset of 6,696 students from 2000 to 2010 from the Banking Department at the Banking Academy of Vietnam.

The findings of these studies highlight several key advantages of using machine learning to predict learning outcomes, including increased prediction accuracy, automation and time efficiency, reduced subjectivity in analysis, improved education quality, and the ability to identify potential trends. Given that machine learning models can process vast amounts of data and be applied on a large scale, their expansion is relatively straightforward, making them particularly beneficial for large schools and institutions that educate thousands of students. Educators could also modify machine learning models, allowing them to be applied to a wide range of educational sectors and systems.

Building on these insights, this study utilizes widely adopted machine learning models in education to predict graduation classifications using accessible data from the Academy of Finance.

3. Experimental results

The sample dataset includes 2344 graduated from the CQ56 and CQ57 cohorts across all majors at the Academy. Due to data extraction limitations, a subset of the total student population from the two cohorts were selected.

Input features for prediction are academic results from the first two years of study. Graduation classifications are: Excellent, Very Good, Good, and Average. Table 1 summarizes the dataset.

Table 1: General information of the student dataset

Graduation classification	Number of Observations	Percentage (%)
Average	20	0.85%
Good	1332	56.83%
Very Good	829	35.37%
Excellent	163	6.95%
Total	2344	100%

Source: Authors' calculations

The data includes students from a variety of programs. Because the institution uses a credit-based system, students can enroll in different courses (within certain limits). Building a prediction model on this data poses a significant challenge due to the need to handle missing data: courses which not all students are required to take, or only a portion of students could participate.

When using raw data, only some machine learning algorithms can be directly applied, such as prediction models using association rules or decision tree algorithms (Do et al., 2023). In this study, the decision tree model was selected to experiment with missing data based on the research of other authors.

The prediction results of the J48 decision tree model on the raw data are quite promising, as detailed in Table 2 and Table 3 below.

Table 2: Performance of the J48 model with raw data

Algorithm	Accuracy	Precision	Recall	F1-Score
J48	97.82%	98.79%	98.71%	98.75%

Source: Authors' calculations

Below are the detailed results for each class:

Table 3: Performance of the J48 model for each class with raw data

Class	Precision	Recall	F1-Score
Very Good	97.32%	96.50%	96.91%
Good	97.83%	98.35%	98.09%
Average	100.00%	100.00%	100.00%
Excellent	100.00%	100.00%	100.00%

Source: Authors' calculations

To accommodate to the data format requirements of other machine learning models, missing data was replaced with the mean of variable. Specifically, numerical missing values were replaced with the mean of the observed values for the respective feature.

Categorically missing values were replaced with the mode (most frequent category) of the observed data. The performance of different machine learning models for predicting student learning outcomes on this processed data is presented in Table 4.

Table 4: Performance of machine learning models with processed data

Algorithm	Accuracy	Precision	Recall	F1-Score
J48	**98.04%**	**98.91%**	**98.84%**	**98.87%**
Nnge	78.86%	78.10%	65.42%	71.20%
Logistic Regression	85.79%	61.43%	59.76%	60.58%
MLP	84.64%	61.66%	58.15%	59.85%
SMO	78.33%	77.32%	40.47%	53.13%
Adaboost	91.25%	90.13%	49.59%	63.98%

Source: Authors' calculations

Considering all the metrics, including accuracy, precision, recall, and F1-score, the J48 decision tree model provides the best prediction results. With the J48 model, these values are 98.04%, 98.91%, 98.84%, and 98.87%, respectively; this result suggests that applying this method in practice at the Academy is entirely feasible. The results of the other models are more limited compared to the decision tree-based model. Among the remaining models, Adaboost performed the best, followed by logistic regression, the multilayer perceptron (MLP) neural network, the Nnge algorithm (a variant of the K-nearest neighbor strategy), and SMO (a support vector machine optimization algorithm). These algorithms gave more limited prediction results compared to the J48 decision tree algorithm. One possible explanation is supplementing missing data with the average values of variables is not the best approach. In the future, the research will be expanded further to address this issue.

Regarding the best-performing algorithm, the J48 decision tree, the classification results for each class in the prediction task are presented in Table 5.

Table 5: Performance of the J48 model for each class with processed data

Class	Precision	Recall	F1-Score
Very Good	97.57%	96.86%	97.22%
Good	98.06%	98.50%	98.28%
Average	100.00%	100.00%	100.00%
Excellent	100.00%	100.00%	100.00%

Source: Authors' calculations

The results demonstrate that the J48 algorithm performed well across all classes. Additionally, minority classes such as "Average", which accounts for only 0.85% of observations, and the "Excellent" class, which represents 6.95% of observations, also achieved strong predictive performance. The prediction model performed well (balanced) across all classes in the student academic performance classification task. This ensures that the predictions do not suffer from bias, particularly in the case of minority classes. Bias in classification could lead to incorrect treatment of minority-class instances when applying the predictions in real-world scenarios.

4. Conclusion

This study contributes to the application of modern machine learning methods in educational data mining at the Academy of Finance. It provides a framework for prediction of final academic outcomes based on students' performance in earlier stages of their studies. The experimental results, using data from graduates in 2022 and 2023 (cohorts CQ56 and CQ57), indicate that the J48 decision tree model offers the best predictive performance. These findings align with the results observed in similar studies by Sumitha et al. (2016) and Bujang et al. (2021).

These predictions can serve as a warning for students currently enrolled at the Academy, prompting them to adjust learning strategies and enhance their commitment to academic pursuits. Furthermore, these insights can assist faculty and administrators in providing targeted support to struggling students and improving overall learning outcomes, ultimately increasing students' employment prospects.

One limitation of the study is that it relies solely on academic performance data from the first two years of study to predict graduation classification. In reality, several other factors could influence graduation outcomes, including extracurricular activities (e.g., sports, performing arts), participation in supplementary courses, involvement in professional clubs, engagement in scientific research, and family-related factors such as parental education level, income, financial support, and family dynamics. Additionally, the issue of handling missing data in machine learning models requires further research when applied to student datasets. Another limitation of this study is the employment of only a few standalone machine learning models. Future research should explore the application of ensemble learning methods, deep learning models, and improvements to existing models to enhance prediction accuracy. Addressing these limitations will be the focus of future research.

References:

Alalawi, K., Athauda, R., Chiong, R. (2023), "Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review", Volume 5, Issue 12, Special Section title: Additive Manufacturing for Advanced Materials and Structures, https://doi.org/10.1002/eng2.12699.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2021), "Multiclass prediction model for student grade prediction using machine learning", IEEE Access (Volume 9), pp. 95608- 95621. DOI: 10.1109/ ACCESS.2021.3093563.

Do, V., T., Nguyen, G. C., Thi, H. D., & Ngoc, L. P. (2023), "Classification and variable selection using the mining of positive and negative association rules", Information Sciences, 631, pp.218-240.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao., S. N. (2018), "Predicting academic performance: A systematic literature review", In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '18 Companion), Larnaca, Cyprus. ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3293881.3295783.

N. T. Uyên, N. M. Tâm (2019), "Áp dụng thuật toán khai phá dữ liệu trong dự báo kết quả học tập của sinh viên", Tạp chí khoa học, Tập 48 - Số 3A, pp. 68-73.

N. T. K. Son, N. V. Bien, N. H. Quynh, C. C. Tho (2022), "Machine learning based admission data processing for early forecasting students' learning outcomes", International Journal of Data Warehousing and Mining, Volume 18, Issue 1. DOI: 10.4018/IJDWM.313585.

N. V. Thủy (2023), "Sử dụng các mô hình Machine learning dự đoán tình trạng sinh viên tốt nghiệp đúng hạn", Tạp chí Khoa học & Đào tạo Ngân hàng, Số 255- Tháng 8. 2023, tr 52-64.

N. G. Cu, T. L. Nghiem, T. H. Ngo, M. T. L. Nguyen, H. Q. Phung (2024), "Increment of academic performance prediction of at-risk student by dealing with data imbalance problem", Applied Computational Intelligence and Soft Computing 2024 (1), 4795606, DOI: https://doi.org/10.1155/2024/4795606.

Sumitha, R., Vinothkumar, E. S., and Scholar, P. (2016), "Prediction of students outcome using data mining techniques", International Journal of Scientific Engineering and Applied Science (IJSEAS), Volume 2, Issue 6, pp.8.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., &, Sobri, N. M. (2019), "Supervised data mining approach for predicting student performance", Indonesian Journal of Electrical Engineering and Computer Science. Vol. 16, No. 3, December 2019, pp. 1584~1592.