ASIAN JOURNAL OF EDUCATIONAL SCIENCES
ISSN: 0866-7675
eISSN: 2815-567X
DONG THAP UNIVERSITY JOURNAL OF SCIENCE - DONG THAP UNIVERSITY
https://ajesdthu.vn

DOI: https://doi.org/10.52714/dthu.ajes.7.1946

Factors Influencing the Academic Performance of Freshmen: Insights from a Data-driven Analysis

Thao-Trang Huynh-Cam[1*], Long-Sheng Chen[2], and Khai-Vinh Huynh[3]

[1]*Foreign Languages and Informatics Center, Dong Thap University, Cao Lanh 870000, Vietnam*

[2]*Department of Industrial Engineering and Management, National Taipei University of Technology, Taiwan*

[3]*Office of Quality Assurance, Dong Thap University, Cao Lanh 870000, Vietnam*

[*]*Corresponding author, Email: hcttrang@dthu.edu.vn*

Article Info.

Received: March 23, 2026
Revised: April 27, 2026
Accepted: May 15, 2026

Keywords

Academic performance, data-driven analysis, feature selection methods, first-year students, higher education policy, influential factors, predictive models.

Cite

Huynh-Cam, T.-T., Chen, L. S., & Huynh, K. V. (2026). Factors Influencing the Academic Performance of Freshmen: Insights from a Data-driven Analysis. *Asian Journal of Educational Sciences, 1*(1), 31-46. https://doi.org/10.52714/dthu.ajes.7.1946

Abstract

Student academic performance has been continuously examined by theorists and practitioners worldwide. Yet, the academic performance of freshmen as newcomers who often face challenges and need assistance has gained little attention from theorists and practitioners. This pilot study uncovers key factors influencing the academic performance of freshmen at a technical and vocational university in Taiwan. Using predictive analytics, this study explored how data-driven methods can enhance students' performance. The analysis utilized the recursive feature elimination (RFE) method for feature selection using two AI-driven models of a Neural Network and a Decision Tree. The research sample included 1,928 freshmen from a technical and vocational university in Taiwan. The input factor dimensions comprised demographic, socio-economic, and family background variables. The output factor was the grade point average for the first term of the academic year 2020/2021. The results showed that among the tested models, the Decision Tree surpassed the Neural Network with an accuracy, precision, recall, and F1 of 86.0% (D1) and of approximately 90.0% (D2); and an ROC-AUC of 86% (D1) and 87% (D2). The three factors-students' fathers' careers, major, and the average monthly income of students' parents-had the highest positive and significant impact on freshmen's academic performance. This work provided AI-driven models serving as an early warning system and identify the strongest predictors of academic performance among freshmen. It is expected to assist educational policymakers in developing proactive measures to increase the number of excellent students and to reduce the number of underperforming/at-risk students at early stages.

Copyright © 2026 The author(s). This work is licensed under a CC BY-NC 4.0 License

1. INTRODUCTION

Students' academic performance plays a vital role in the quality of education (Alhazmi & Sheneamer, 2023). A high proportion of both outstanding and underperforming students has a crucial impact on universities' reputations, public image, and financial implications for governments. It also significantly affects students' motivation and shapes the design and delivery of university courses (Bai et al., 2021). Even though universities have continuously offered a variety of support measures, such as academic supplements and grants, these methods may be inappropriate in terms of timing and may not be well prepared to reduce the ratio of withdrawals/at-risk students and to increase the number of high-performing students. Consequently, university policy makers have expressed a strong demand for early performance prediction measures (López-Zambrano et al., 2021) to help increase the number of outstanding students and reduce the number of underperforming students during the early stages of study, before students reach the final semester. They have also placed high expectations on an early warning system to identify influential factors associated with high- and low-achievers before the start of the first academic term in order to predict students' learning trends, plan accordingly, and implement timely incentives and preventive measures.

An early warning system alone is not enough to effectively enhance students' academic performance, since understanding the influential factors affecting students' academic performance at the early stage of their educational process is complex (Alhazmi & Sheneamer, 2023). To better foster students' academic performance at the early stages, the academic performance, especially outstanding and underperforming performance, among freshmen, should receive greater attention, as achievement in the first year plays a crucial role in shaping students' attitudes and performance in subsequent academic years (Martins et al., 2021). Additionally, freshmen as newcomers frequently need more assistance because they face a variety of challenges during their transition to university life (Meehan & Howells, 2018), including high school-university transition (Mulaudzi, 2023), university expectations-reality mismatch (Zaťková et al., 2025), low academic achievement (Rodríguez et al., 2017), adaptation to an independent learning environment (Cameron & Rideout, 2022), and economic difficulties (Brooker et al., 2017). First-year students are also more likely to drop out of their university programs (Azmitia et al., 2018).

Owing to the digitalization of academic processes, universities can gain, store, and analyze students' big data, which is available in school databases. If these data are transformed into meaningful knowledge, they can be useful for the quality of education and helping students achieve their academic objectives at the early stages (Alhazmi & Sheneamer, 2023). Data mining and educational data mining through Artificial Intelligence (AI) algorithms can aid universities in addressing students' performance challenges (López-Zambrano et al., 2021; Alhazmi & Sheneamer, 2023). In recent years, data mining and AI-driven models have been effectively employed to predict student academic performance in studies such as those by Kaunang and Rotikan (2018), Musso et al. (2020), Mengash (2020), López-Zambrano et al. (2021), Alhazmi and Sheneamer (2023), Kassaw and Demareva (2024), and Çırak et al. (2024). Yet, previous studies reviewed in this study primarily relied on during- or post-course/semester factors, such as degree completion (Musso et al., 2020) and e-learning activities (Moreno-Marcos et al., 2020) to uncover influential factors affecting students' learning performance. Additionally, freshmen have received less attention among scholars. These methods cannot help universities in general, and vocational universities in Taiwan in particular, gain sufficient time to implement proactive measures and provide timely support for students with outstanding or poor academic performance. Thus, this study aims to:

1) Explore the factors that most strongly influence excellent and underperforming academic performance among freshmen

2) Develop AI-driven models that serve as an early detection system for freshmen' academic performance

For practical purposes, this work makes a major contribution to improving the academic performance of students in Taiwanese vocational universities. This work provided AI-driven models that can serve as an early warning system. By using enrollment information factors, which can be obtained as soon as students enroll at universities, this study can aid school leaders in identifying outstanding and underperforming students, as well as the influential factors, before they begin their first term. It is expected to assist educational policy makers and stakeholders in developing proactive measures to reduce the number of underperforming/at-risk students from the early stages of their studies.

2. LITERATURE REVIEW

2.1. Data Mining in Education

Data mining is an automated step in the process of knowledge discovery in databases, focusing on discovering new and useful knowledge from big data (Shu et al., 2023). It has been effectively applied in the social sciences (Shu et al., 2023), healthcare (Southall et al., 2019), and tourism (Shapoval et al., 2018). Recently, data mining approaches through machine learning or AI algorithms have increasingly inspired diverse researchers to analyze educational data to explore potential and practical solutions to academic problems and resolve educational research issues (Arcinas et al., 2021). This novel approach is known as educational data mining or data mining in education (Romero & Ventura, 2007). The process of data mining in education comprises several steps: data collection, data preprocessing, data mining, interpretation and evaluation, and knowledge extraction, as displayed in Figure 1 (Romero & Ventura, 2007).

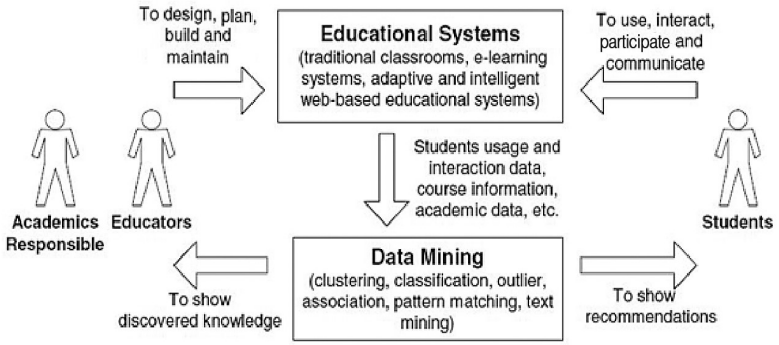


Figure 1. The Process of Data Mining in Education (Romero & Ventura, 2007)

Educational data mining is able to retrieve meaningful knowledge from analyzing large-scale data collected in educational systems to better understand students' learning trends and conditions (Sarra et al., 2019; Arcinas et al., 2021). The purposes of data mining in education are to (a) predict students' academic performance, (b) explore and upgrade current domain models, (c) analyze the influences of diverse types of instructional support, and (d) improve scientific knowledge (López-Zambrano et al., 2021). In order to add value, data mining in education must assist educational policy makers in enhancing students' academic performance and future success for sustainable development (Sarra et al., 2019; Huynh-Cam et al., 2022). This study was carefully conducted based on the theoretical background of data mining in education and followed its guidance in order to add value to educational systems.

2.2. Students' Academic Performance

For decades, applications of data mining approaches in education have successfully supported university policy makers in their efforts to enhance students' academic performance and success. Researchers' concerns have ranged from entry levels to graduation levels across the university timeline. For instance, at the entry level, as part of an early warning system, Lau et al. (2019) used socio-economic background variables and scores in Chinese, English, Mathematics, Comprehensive Science, and the Proficiency Test in the national university entrance examinations to predict students' performance using statistical methods and a Neural Network (NN) algorithm. They concluded that the NN model performed better, with an accuracy of 84.8%, a recall of 94.8%, a precision of 86.3%, and an Area Under the ROC Curve of 86%. They also confirmed that the most influential predictors of grade point average across four university years were gender (female students), English scores, and students' mothers' background characteristics.

Mengash (2020) employed high-school GPA, Scholastic Achievement Admission Test scores, and General Aptitude Test scores to predict the GPA for the first two semesters of study among 2,039 university students majoring in computer science in the Kingdom of Saudi Arabia, using NN, Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB) algorithms. The study found that the NN algorithm surpassed the others, with an accuracy of 79.22%, precision of 81.44%, recall of 78.03%, and an F1-score of 79.70%. The findings also indicated that early predictors of students' performance after the first two semesters included students' Scholastic Achievement Admission Test scores and pre-admission information.

Interestingly, Musso et al. (2020) applied GPA, academic retention, and degree completion to predict academic performance from the second- to the final-academic year of 655 Argentinean students from introductory psychology classes (Psychology, Communication, Business, and Marketing), using a NN algorithm. They found that the NN model achieved maximum overall accuracy, precision, recall, specificity, and F1-scores of 100%. They also found that learning strategies were the strongest predictors of students' GPA, whereas background information was the strongest predictor of students' academic retention decisions, and coping strategies were the strongest predictors of degree completion.

A review of AI-driven models indicates that many previous studies surveyed in the present work have successfully employed DT and NN supervised learning algorithms, using students' academic performance as the class label.

The DT algorithm has been successfully utilized for prediction and classification in the domain of machine learning (Albreiki et al., 2021). DT is a model-based method that uses a tree-shaped graph. Factors appearing in the tree are considered important; in contrast, factors that do not appear in the tree are considered unimportant (Matzavela & Alepi, 2021; Huynh-Cam et al., 2022). Compared to other supervised learning algorithms, DT can provide readable knowledge rules, which are helpful for university decision-making processes (Huynh-Cam et al., 2022). Evidently, DT has been successfully used in educational fields due to its high prediction accuracy, ease of use, and ease of understanding and interpretation for knowledge representation (Matzavela & Alepi, 2021; Ahmad et al., 2025). This success supports the selection of the DT model for this study.

NN is a computational model that imitates the neural structures and processes of the human brain. Its structure includes an input layer, a hidden layer, and an output layer. The input layer receives data, the hidden layer processes the data, and the output layer produces the final output. Several previous studies, such as Mengash (2020) and Musso et al. (2020), successfully used NN to predict learning performance. Since NN is the foundation of AI, which was originally designed to solve problems that are impossible for humans to carry out, this study applied NN as a comparison base.

In short, due to the successful applications and significant attention given to AI-driven approaches mentioned above, this study followed this trend by employing AI-driven methods to forecast freshmen's academic performance at a Taiwanese vocational university using enrollment information factors.

2.3. Feature Selection Methods

Feature selection is a method used to select a subset of relevant features from a larger set of original datasets in order to address overfitting problems and enhance predictive performance (Chandrashekar & Sahin, 2014). This process helps researchers select suitable features for the training dataset, identify the most relevant features for predicting target data, and determine which features should be eliminated or retained for further analysis, thereby improving predictive performance. One of the most popular feature selection methods is the recursive feature elimination (RFE) algorithm, since RFE tends to remove redundant and weak features while retaining independent features (Priyatno & Widiyaningtyas, 2024). RFE starts with the entire set of features and removes the least informative feature from the total feature set one by one in each iteration. This study used the RFE method to select relevant features for constructing AI-driven models.

3. MATERIALS AND METHODS

There are six steps employed in this research as displayed in Figure 2. Each step is described in detail in the following subsections.

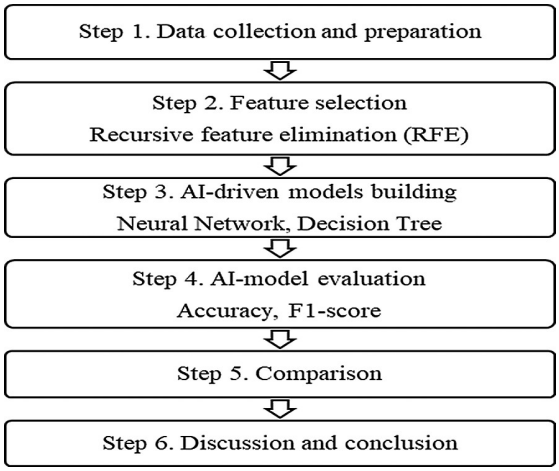


Figure 2. Research Procedure Employed in this Study

(1) Step 1. Sample, data collection and preparation

The sample used in this pilot study included 1,928 freshmen enrolled in a technical and vocational university in Taiwan at the end of the first term of the 2020/2021 academic year. They were randomly selected from ten majors at the target university via the school database. Students' identities were kept anonymous for ethical purposes. The input (independent) factors included thirteen enrollment information variables (Table 1), grouped into three major dimensions-demographic, socio-economic, and family background. The output (dependent) factor was the grade point average (GPA) score for students' first term of the target year.

After collecting raw data, Microsoft excel 2019 was used for preprocessing. *Firstly,* several irrelevant factors, such as names, ID numbers, birth dates, and birthplaces, were removed. Next, all students with missing values-for example, unknown GPA scores and students who dropped out or were suspended before the end of the first term-were excluded. This means that these samples had no output factor (GPA). In supervised machine learning methods used to predict GPA, AI models learn by mapping input factors to a target output-GPA. If a specific sample is missing the GPA factor, it has no correct answer for AI models to learn from. Consequently, when building the training set for this problem, samples containing missing values in the target output (GPA) are often removed because they lack the

necessary labeled output. *Secondly,* as computers can only understand numerical values, all categorical values were transformed into numerical values. Table 1 describes the input and output factors utilized in this study and their corresponding transformed values. Figure 3 shows the correlation matrix among the employed input and output factors. Next, the preprocessed data were normalized using Eq. (1).

$$X_N = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X_N is the normalized-, X_max is the maximum-, and X_min is the minimum values.

Table 1. Description of the Input and Output Factors Employed in this Study

ID No.	Feature description	Values and their transformed values
Output	Academic performance (class label)	Excellent (EX) = 5; Very good (VG) = 4; Good (G) = 3; Average (AVG) = 4; Poor (PR) = 1
X1	Major	Finance = 1; Business administration = 2; Architecture = 3; Applied chemistry = 4; Leisure service management = 5; Industrial design = 6; Applied English = 7; Information Management = 8; Industrial engineering management = 9; Civil and construction engineering = 10
X2	Gender	Men=1, women=2
X3	Students' home address	Not applicant = 0; North = 1; South = 2; Central = 3; East = 4
X4	University admission type	General admission=1; International-student admission=2
X5	Student's parents' average monthly income (in Taiwan dollars)	Very low: ≤ 25,000 = 1 Low: Above 25,000~40,000 = 2 Medium: Above 40,000~60,000 = 3 High: Above 60,000~100,000 = 4 Very high: Above 100,000 = 5
X6	Student lives in dormitory or not	Outside = 0; In-dorm = 1
X7	Students' living expenses support	Parents/family provided=1; Self-earning=3; Studying loans=4; Full-time job=5; Part-time job=6; Scholarships/grants in- and out school=7
X8	Studying loans in- and out- school	Yes = 1, No = 0
X9	Free/reduced tuition fee	Free/reduce = 1; Not free/reduce = 0
X10	Student fathers' career	State officer = 1; Teacher = 2; Laborer = 3; Businessmen = 4; Agriculture = 5; House husband = 6; Others =7
X11	Father's highest education status	PhD = 6; Master = 5; Specialist = 4; Bachelor = 3;
X12	Student mothers' career	State officer = 1; Teacher = 2; Laborer = 3; High school = 2; Below high school = 1 Businessmen = 4; Agriculture = 5; Housewife = 6
X13	Mother's highest education status	Specialist = 4; Bachelor = 3; High school = 2; Below high school = 1

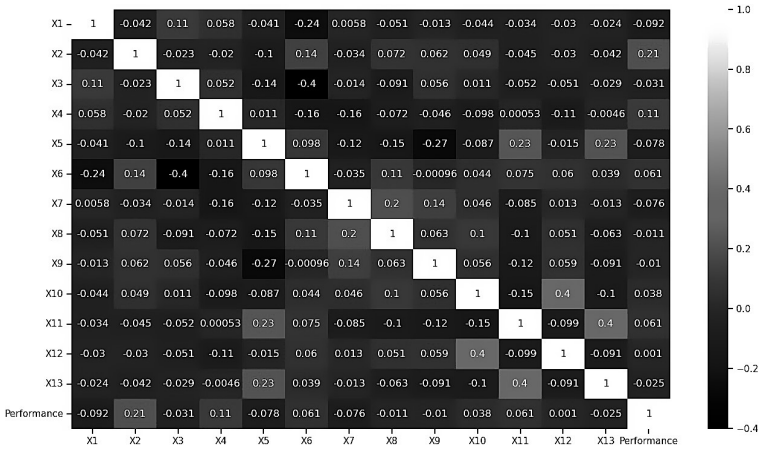


Figure 3. Confusion Matrix Heatmap of the Employed Input and Output Factors

Finally, we used the output factor "GPA scores" as a class label to classify the data based on Taiwanese grading score system (100-point scale). Table 2 shows the class distribution of employed samples by GPA scores.

Table 2. Class Distribution of Employed Samples by GPA Scores (n=1928)

No.	GPA score	Class distributed	Number of students	Percentage (%)
1	90-100 points	Excellent (EX)	71	3.68
2	80 - 89 points	Very Good (VG)	707	36.67
3	70-79 points	Good (G)	770	39.94
4	60-69 points	Average (AVG)	295	15.30
5	Below 60 points	Poor (PR)	85	4.41

As displayed in Table 2, it is obvious that compared to other classes, the class "EX" and "PR" are minority with a small number of samples 71(3.68%) and 85 (4.41%), respectively; yet they are meaningful for university decision makers. This means that class imbalance was clearly present. Consequently, we conducted three classification cases to handle class imbalance problem as follows.

- Case A, namely Origin, used original dataset with five classes (Table 2) as a base comparison.

- Case B, namely Focus and Combine (FC), emphasized two minority class "EX" and "PR". For the research purposes, predictions on minority classes "EX" and "PR" were prior for analysis. Thus, at first, we removed the class "G". Then we combined 2 classes: "EX" and "VG" to create a new class "nEX". Similarly, we combined 2 classes: "AVG" and "PR" to create a new class "nPR". Yet, after combining, the imbalance problem was present between two new classes "nEX" and "nPR". Additionally, the AI models could not gain expected prediction performance. Thus, we used a resampling method - random oversampling, adapted from Chang et al. (2021) and Huynh-Cam et al. (2022) in Case C to solve the imbalanced data problem between two new classes "nEX" and "nPR".

- Case C, namely Oversampling (OS), randomly oversampled the minority class "nPR" until it approximately equals to the majority class "nEX". We used this resampling method to handle class imbalance problems in this work because Chang et al. (2021) and Huynh-Cam et al. (2022) illustrated that oversampling is one of the effective solutions for handling class imbalance problems.

Table 3. Numbers of Samples in Case B and Case C

Case	No.	GPA score	Class	Number of students	Distribution
Case B: Focus and Combine (FC)	1	80 ~100 pts.	nEX	778 (40.35%)	Combine 2 class: "EX" & "VG"
	2	69 ~ below 60 pts	nPR	380 (19.71%)	Combine 2 class: "AVG" and "PR"
Case C: Oversampling (OS)	1	80 - 100 pts	nEX	778 (40.35%)	Remain the same
	2	69 ~ below 60 pts	nPR	760 (39.42%)	Oversampled

(2) Step 2. Feature selection

The prepared data was divided into two datasets for building AI models. Dataset 1 (D1) used all full 13 factors (Table 1) as a base comparison. Dataset 2 (D2) used 10 RFE-selected factors, including X1, X2, X4, X5-X8, X10, X11, and X13.

(3) Step 3. AI-driven models building

This study utilized two AI models: NN and DT to predict and explore the most influential factors for excellent and poor academic performance of freshmen. We utilized DT due to its wide application, simplicity, ease of use and ease of interpretation. DT can be visualized in an understandable manner and be clearly interpreted through visualized models. We applied NN as a comparison base since it is the foundation of AI model.

The study deployed the experiments on Windows Operating System with a 3.80 GHz Intel(R) Xeon(R) E-2174G CPU and 64 GB of RAM. The Python 3 programming language, using Jupyter Notebook software (version 6.5.4) with Scikit-learn packages was used to build the models and analyze the data. Each AI model was constructed, employing a 5-fold cross validation (CV) process (Browne, 2000) with five different training-testing data (80:20). This means that Dataset 1 (D1) and Dataset 2 (D2) were split into five different data folds with the ratio 80:20 (Table 4). Each data fold took turn to be used as a training dataset to build NN and DT models. The rest were used as the testing models. The mean value and standard deviation (SD) of five data folds (5-fold CV) were used for comparing prediction performance among AI models for each dataset. Table 5 shows parameters used for these AI models.

Table 4. Data Spit for Neural Network and Decision Tree Model

Sample	Number of samples	Percentage
Training set	1542	80%
Testing set	386	20%
Total	1928	100%

Table 5. Parameters Used for Constructing AI Models

Models	Parameters	Values
DT	Criterion	Gini for measuring the impurity and avoiding overfitting*
	Max depth	3-5 to prune the tree
	Random state	None
NN	Hidden layer size	17, 10
	Activation function	relu, sigmoid
	Learning rate	0.01
	Optimizer	adam
	Loss	Binary cross entropy
	Epochs	100

* Its value is between 0-1, where 0 is perfectly pure. Further explanation is available at https://youtu.be/sEIrZ66Pj0E?si=TIbbXeOn-otpiMTL

(4) Step 4. AI-model evaluation

After building AI models (Step 3), five evaluation metrics - accuracy, precision, recall, F1, and ROC-AUC, which were adjusted from (Huynh-Cam et al., 2024) and (Chang et al., 2021), were used to assess AI models' performance through Eqs. (2)-(5). These metric values are 0-1, where 1 means perfect performance and 0 refers to poor. True Positive (TP) means truly "EX" students are correctly classified as "EX"; True Negative (TN) means truly "PR" students are correctly classified as "PR". False Positive (FP) and False Negative (FN) are errors, in which truly "EX" and "PR" students are misclassified.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1} = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (5)$$

(5) Step 5. Comparison

After evaluation, we compared and selected the AI model and data fold, having the highest important scores, to extract the top influential factors for first year students' academic performance.

(6) Step 6. Discussion and conclusion

After finalizing the most influential factors, we summarized the research outcomes, suggested several practical solutions for university leaders, and drew a basic map for future studies.

4. RESULTS

4.1. Results of Model Evaluation

This study used two AI models - NN and DT - to classify excellent and poor academic performance of freshmen using two datasets. Dataset 1 (D1) employed fully 13 original input factors without using any feature selection methods as a base comparison. Dataset 2 (D2) used 10 factors selected using the RFE method (Step 2). We used overall accuracy, precision, recall, F1, and ROC-AUC to assess AI-models' performance across two datasets. Firstly, we judged the results of the first four evaluation metrics. Table 6 compares these metrics.

In Case A (Origin), it is obvious that the overall accuracy, precision, recall, F1 results of DT and NN algorithms are very low. In particular, for D1, DT performs better than NN with overall accuracy of 50.0% (2.6), precision of 25.7% (25.0), recall of 26.3% (25.1), and F1 of 23.7% (20.5), respectively. For D2, its accuracy, precision, recall, and F1 scores are 52.3% (0.6), 23.0% (1.4), 24.0% (0.0), and 23.3 (0.6), respectively. This means that in Case A, the minority classes "EX" and "PR" cannot be detected correctly, indicating this classification results are meaningless. The low accuracy may result from 5-class classification. Consequently, Case B (FC) only emphasized two minority classes: "EX" and "PR".

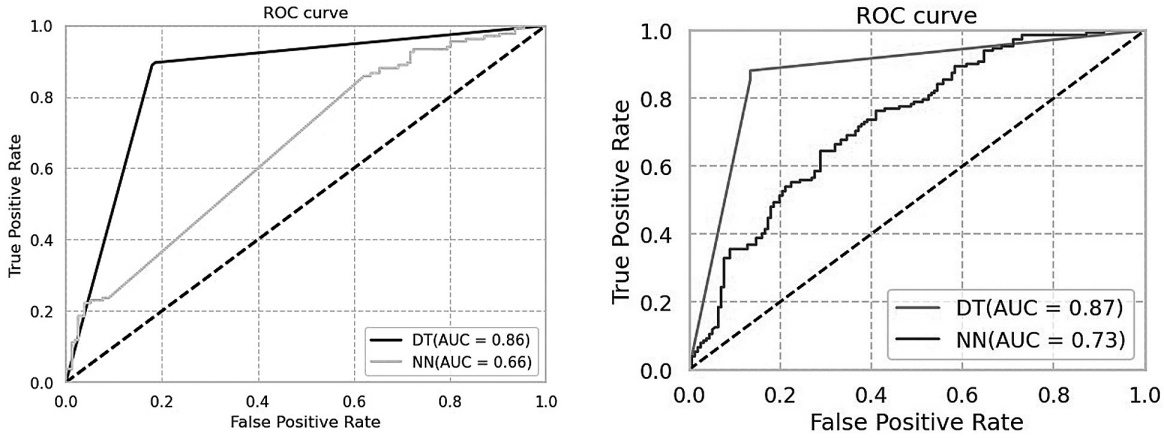
In Case B, all evaluation metrics of DT and NN models constructed in both D1 and D2 increase. In particular, for D1, the accuracy, precision, recall, and F1 scores of DT model rise to 65.0% (0.7), 60.5% (6.4), 59.0% (0.0), and 60.0 (2.8), respectively. For D2, accuracy increases up to 63.7% (0.6), precision rises to 53.3% (0.6), recall is up to 57.0% (1.7), and F1 increases to 55.0% (1.0), respectively. These values for NN model built in D1 and D2 do not much higher than those in Case A.

In Case C, it is clear that the accuracy, precision, recall, F1 values of DT and BB models built in Case C (OS), significantly and stably increase. Between two models, DT model surpasses NN with all values of 86.0% for D1 and of approximately of 90.0% for D2. This advisable that Case C should be used for further analysis.

Table 6. A Comparative Model Evaluation between Two Datasets

Dataset	Classification	Model	Performance (%) (Mean, SD)			
			Accuracy	Precsion	Recall	F1
Dataset 1: Full factors	Case A: Origin	DT	50.0 (2.6)	25.7 (25.0)	26.3 (25.1)	23.7 (20.5)
		NN	45.0 (5.3)	19.3 (16.7)	21.3 (25.8)	18.7 (18.5)
	Case B: FC	DT	65.0 (0.7)	60.5 (6.4)	59.0 (0.0)	60.0 (2.8)
		NN	47.0 (2.8)	33.5 (4.9)	15.0 (8.5)	20.5 (9.2)
	Case C: OS	DT	86.0 (0.0)	86.0 (7.1)	86.0 (5.7)	86.0 (0.0)
		NN	69.0 (0.0)	68.5 (3.5)	71.5 (0.7)	67.5 (0.7)
Dataset 2: RFE-selected factors	Case A: Origin	DT	52.3 (0.6)	23.0 (1.4)	24.0 (0.0)	23.3 (0.6)
		NN	43.3 (1.5)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Case B: FC	DT	63.7 (0.6)	53.3 (0.6)	57.0 (1.7)	55.0 (1.0)
		NN	51.7 (9.0)	35.0 (16.7)	23.7 (27.6)	26.0 (24.6)
	Case C: OS	DT	89.0 (2.8)	89.5 (3.5)	88.5 (2.1)	89.0 (2.8)
		NN	68.5 (2.1)	67.5 (2.4)	69.5 (4.9)	68.5 (3.5)

In the final step, we computed ROC-AUC metric for in-depth comparison. From Figure 5, it is clear that DT performs better and more stable than NN with an AUC of 86% (D1) and 87% (D2). This can be concluded that DT model can be able to correctly classify excellent and underperformed students better than the NN model. Thus, DT was used to extract the top influential factors for freshmen's academic performance.

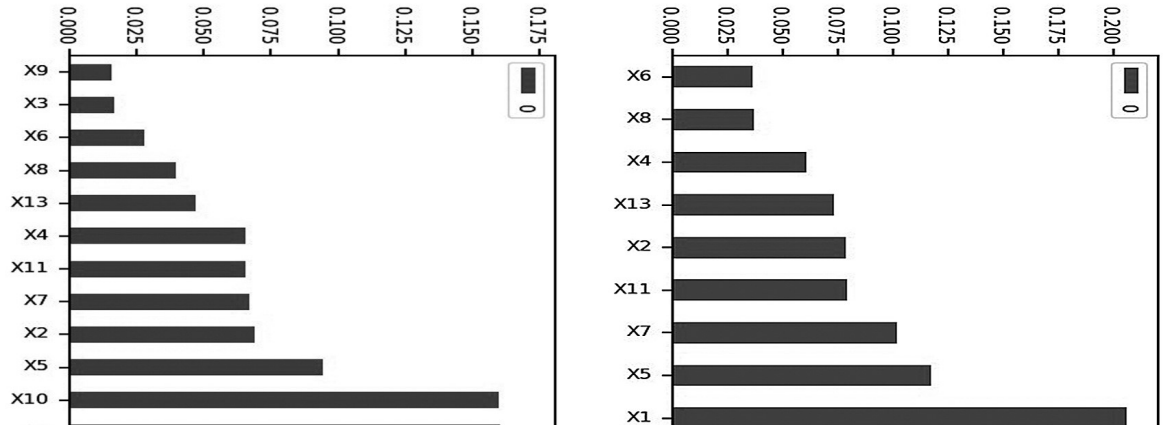


Dataset 1 (D1): Full factors **Dataset 2 (D2): RFE-selected factors**

Figure 5. Comparisons on ROC-AUC between Two Datasets (Case C: OS)

4.2. Results of Feature Selection

Figure 6 shows the rankings of feature importance retrieved from D1 and D2. Student mothers' career (X12) and student fathers' career (X10) are the most influential factors for excellent and underperformed academic performance of first year students. Free/reduced tuition fee (X9) and student lives in dormitory or not (X6) hardly affect students' excellent and poor performance, respectively.



(a) Dataset 1 (D1): Full factors **(b) Dataset 2 (D2): RFE-selected factors**

Figure 6. Feature Importance Rankings of Two Datasets

4.3. Comparative Results of Feature Importance

Table 7 lists the rank order of the most important features driving excellent and underperformed academic performance of freshmen. The top three factors associated with outstanding and underperformed academic performance comprise student fathers' career (X10), major (X1), and excellent and average monthly income of student's parents (X5).

Table 7. Comparisons of Feature Importance between Two Datasets (Decision Tree Model)

Rank order	Dataset 1: Full factors	Importance score	Dataset 2: RFE-selected factors	Importance score	The most influential factors finalzed*	
1	X12	0.172	X10	0.210	X10	Student fathers' career
2	X1	0.161	X1	0.206	X1	Major
3	X10	0.160	X5	0.117	X5	Student's parents' average monthly income
4	X5	0.094	X7	0.102		
5	X2	0.069	X11	0.079		
6	X7	0.067	X2	0.078		
7	X11	0.066	X13	0.073	N/A	
8	X4	0.065	X4	0.061		
9	X13	0.047	X8	0.037		
10	X8	0.039	X6	0.036		
11	X6	0.028				
12	X3	0.016		N/A		
13	X9	0.016				

** Note: Selected based on importance scores and frequency between two datasets.*

5. DISCUSSION

The research findings indicated that fathers' careers, students' majors, and parents' average monthly income significantly impact the academic performance of first-year students from Taiwanese vocational and technological universities. These results are aligned with the findings of Mengash (2020) and Musso et al. (2020). These findings suggest that collaboration between parents and schools plays an important role in improving the academic performance of freshmen in Taiwan. Educational stakeholders should take several practical and proactive actions to improve the academic performance of freshmen both internally and externally (Rodríguez et al., 2017; Brooker et al., 2017; Meehan & Howells, 2018; Mulaudzi, 2023).

Internally, at the school level, university policy makers should offer financial grants to aid students in overcoming financial difficulties and arrange academic support teams to address academic issues. These practical and proactive measures can help freshmen overcome the extreme challenges they face during the early stages of university life (Rodríguez et al., 2017; Brooker et al., 2017; Meehan & Howells, 2018; Mulaudzi, 2023). Students who are predicted early to have poor academic performance may receive timely interventions to improve their academic achievement (Alhazmi, & Sheneamer, 2023). For underperforming students, schools can implement early warning counseling measures and provide remedial teaching resources, individualized teaching assistance, and outside-the-classroom support to help them integrate into a new learning environment and community. These suggested measures can effectively prevent underperforming students from falling behind in their learning process and dropping out during the early stages of study, thereby increasing the number of outstanding students. For outstanding students, apart from elite and excellence scholarships, universities should offer advanced support programs such as special classes for professional and technical development, license examination training, entrepreneurial competitions, and other skill development opportunities.

It is illustrated that the second most influential factor associated with the academic performance of freshmen in Taiwanese vocational and technological universities is the major in which students enroll; hence, at the academic department level, several practical and proactive actions should be considered to help freshmen. For example, departments should create a more supportive, positive, friendly, and stimulating learning environment (Cameron & Rideout, 2022). Students often wish to be recognized; thus, school policy makers and teachers should pay more attention to outstanding students and provide appropriate encouragement. Such encouragement may take the form of positive feedback and recognition, as well as organized competitions and activities that promote academic excellence. Teachers should promptly provide feedback on students' assessments and assignments, which can support students' learning. Peer mentoring, in which an experienced student supports a less experienced student by providing useful knowledge and information, is also recommended because this method can help freshmen transition to the university environment (Chalapati et al., 2018). At universities, students encounter many difficulties. For example, learning and teaching methods, academic requirements, and learning materials are often quite different from those in high school. Students have just transitioned from high school to university, where they are expected to study independently. Their expectations of university life may not match reality (Zaťková et al., 2025). Additionally, they may lack experience in course registration and may not know how many credits are appropriate for their studies.

Externally, parents should actively monitor and support their children's academic progress and help them overcome difficulties during their early university years. As illustrated, the most influential factor affecting students' academic performance is fathers' careers. According to Taiwanese culture, students are often highly dependent on their parents. Family background significantly influences Taiwanese students' academic performance. As first-year students have just left home and begun living independently for the first time, they strongly need parental support, especially from their fathers, in terms of both guidance and financial assistance.

6. CONCLUSION

Student academic performance has been continuously examined by theorists and practitioners worldwide. Yet, the academic performance of freshmen as newcomers who often face greater challenges and require more assistance-has received limited attention from theorists and practitioners. Many previous related studies used during- or post-course/semester factors, such as degree completion and

e-learning activities, to uncover influential factors affecting students' learning performance. However, these methods cannot provide universities in general, and vocational universities in Taiwan in particular, with sufficient time to implement proactive measures and provide timely support for students with outstanding and poor academic performance. The present study used enrollment information factors to construct AI-driven models-NN and DT-and to explore the most influential factors affecting the academic performance of freshmen. The results confirmed that the DT model performed better than the NN model, with an accuracy of 91% and an F1-score of 89.5%. This study significantly contributed to the theories and practices related to improving students' academic performance.

For practical purposes, this work provided predictors of outstanding and underperforming academic performance among freshmen at Taiwanese vocational universities. These predictors aided school policy makers in identifying outstanding and underperforming students before they began their first term, as these factors can be obtained at the time of enrollment. The early prediction models can serve as an early warning system for vocational universities in Taiwan. Based on these findings, educational policy makers can have sufficient time to prepare additional support measures and proactive actions to improve students' academic performance. These predictors can also be used as benchmarks for other educational institutions.

For theoretical purposes, this study contributed to the integration of AI-driven models with the RFE feature selection method, serving as an early warning system for students' academic performance. Regarding data analysis, the findings indicate that, when using AI models, the oversampling method is effective in handling data imbalance issues because it can provide better classification performance. Most importantly, this study suggested that, when dealing with data collected directly from school databases, data containing personal information must be handled carefully for ethical reasons.

Although the present research offered major contributions to the theories and practices of predicting freshmen's academic performance, it still has some limitations. This research was conducted as a pilot study at a vocational university in Taiwan, and the research data were collected from a single vocational university in Taiwan. Hence, the results cannot be generalized to other universities with similar contexts. However, evidence suggests that data from other universities can also be used in future research to benchmark the findings of this study.

DECLARATIONS

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because all collected data were completely anonymized at the source, kept confidential, and used solely for statistical analysis.

Transparency: The author confirms that the manuscript presents an accurate, honest, and transparent account of the research. All essential aspects of the study, including the methodology, sample characteristics, and data analysis, have been transparently reported, and no important details have been omitted.

Competing Interests: The authors declare no conflict of interest.

Authors' Contributions: Conceptualization: Thao-Trang Huynh-Cam and Long-Sheng Chen, research design: Thao-Trang Huynh-Cam, Long-Sheng Chen, and Khai-Vinh Huynh, methodology: Thao-Trang Huynh-Cam and Long-Sheng Chen, software, Thao-Trang Huynh-Cam; validation, Thao-Trang Huynh-Cam, Long-Sheng Chen, and Khai-Vinh Huynh, formal analysis, Thao-Trang Huynh-Cam

and Khai-Vinh Huynh, investigation, Long-Sheng Chen; resources, Thao-Trang Huynh-Cam and Long-Sheng Chen, writing-original draft preparation, Thao-Trang Huynh-Cam, Long-Sheng Chen, and Khai-Vinh Huynh; writing-review and editing, Thao-Trang Huynh-Cam and Long-Sheng Chen; visualization, Khai-Vinh Huynh. All authors have read and agreed to the published version of the manuscript.

Disclosure of AI Use: The authors declare that no generative AI or AI tools have been used in any part of this work since all ideas and expressions are original to the writer.

REFERENCES

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *Ieee Access, 11*, 27579-27589.

Azmitia, M., Sumabat-Estrada, G., Cheong, Y., & Covarrubias, R. (2018). "Dropping out is not an option": How educationally resilient first-generation students see the future. *New Directions for Child and Adolescent Development, 2018*(160), 89-100.

Arcinas, M. M., Sajja, G. S., Asif, S., Gour, S., Okoronkwo, E., & Naved, M. (2021). Role of data mining in education for improving students performance for social change. *Turkish Journal of Physiotherapy and Rehabilitation, 32*(3), 6519-6526.

Ahmad, A., Ray, S., Tabrej Khan, M., & Nawaz, A. (2025). Student performance prediction with decision tree ensembles and feature selection techniques. *Journal of Information & Knowledge Management, 24*(02), 2550016.

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

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*(1), 108-132.

Bai, S., Hew, K. F., Sailer, M., & Jia, C. (2021). From top to bottom: How positions on different types of leaderboard may affect fully online student learning performance, intrinsic motivation, and course engagement. *Computers & Education, 173*, 104297.

Baashar, Y., Alkawsi, G., Ali, N. A., Alhussian, H., & Bahbouh, H. T. (2021, July). Predicting student's performance using machine learning methods: A systematic literature review. In *2021 International Conference on Computer & Information Sciences (ICCOINS)* (pp. 357-362). IEEE.

Brooker, A., Brooker, S., & Lawrence, J. (2017). First year students' perceptions of their difficulties. *Student Success, 8*(1), 49-62.

Cameron, R. B., & Rideout, C. A. (2022). 'It's been a challenge finding new ways to learn': first-year students' perceptions of adapting to learning in a university environment. *Studies in Higher Education, 47*(3), 668-682.

Chang, J. R., Chen, L. S., & Lin, L. W. (2021). A novel cluster based over-sampling approach for classifying imbalanced sentiment data. *IAENG International Journal of Computer Science, 48*(4), 1118-1128.

Çırak, C. R., Akıllı, H., & Ekinci, Y. (2024). Development of an early warning system for higher education institutions by predicting first-year student academic performance. *Higher Education Quarterly, 78*(4), e12539.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28.

Chalapati, S., Leung, R., & Chalapati, N. (2018). Exploring factors affecting first-year students' learning experiences: A case study of a private university in Taiwan. *Student Success, 9*(4), 25-39.

Huynh-Cam, T. T., Chen, L. S., & Huynh, K. V. (2022). Learning performance of international students and students with disabilities: Early prediction and feature selection through educational data mining. *Big Data and Cognitive Computing, 6*(3), 94.

Huynh-Cam, T. T., Chen, L. S., Nguyen, V. C., Nguyen, T. H., & Lu, T. C. (2024). Why first-year e-students are dissatisfied: Machine learning methods for enhancing retention. *International Journal of Applied Sciences Engineering, 21*, 2023532.

Kaunang, F. J., & Rotikan, R. (2018, October). Students' academic performance prediction using data mining. In *2018 Third International Conference on Informatics and Computing (icic)* (pp. 1-5). IEEE.

Kassaw, C., & Demareva, V. (2024, October). Predictor of low academic achievement among Dilla university students, southern Ethiopia, 2024. In *Frontiers in Education* (Vol. 9, p. 1438322). Frontiers Media SA.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences, 1*(9), 982.

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema, 33*(3), 456.

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021, March). Early prediction of student's performance in higher education: A case study. In *World Conference on Information Systems and Technologies* (pp. 166-175). Cham: Springer International Publishing.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *Ieee Access, 8*, 55462-55470.

Meehan, C., & Howells, K. (2018). 'What really matters to freshers?': evaluation of first year student experience of transition into university. *Journal of Further and Higher Education, 42*(7), 893-907.

Mulaudzi, I. C. (2023). Challenges faced by first-year university students: Navigating the transition to higher education. *Journal of Education and Human Development, 12*(2), 79-87.

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education, 80*(5), 875-894.

Moreno-Marcos, P. M., Pong, T. C., Munoz-Merino, P. J., & Kloos, C. D. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access, 8*, 5264-5282.

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. *Computers and Education: Artificial Intelligence, 2*, 100035.

Priyatno, A. M., & Widiyaningtyas, T. (2024). A systematic literature review: recursive feature elimination algorithms. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 9*(2), 196-207.

Rodríguez, M. S., Tinajero, C., & Páramo, M. F. (2017). Pre-entry characteristics, perceived social support, adjustment and academic achievement in first-year Spanish university students: A path model. *The Journal of Psychology, 151*(8), 722-738.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Sarra, A., Fontanella, L., & Di Zio, S. (2019). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research, 146*(1), 41-60.

Shapoval, V., Wang, M. C., Hara, T., & Shioya, H. (2018). Data mining in tourism data analysis: inbound visitors to Japan. *Journal of Travel Research, 57*(3), 310-323.

Southall, N. T., Natarajan, M., Lau, L. P. L., Jonker, A. H., Deprez, B., Guilliams, T., ... & Ardigò, D. (2019). The use or generation of biomedical data and existing medicines to discover and establish new treatments for patients with rare diseases-recommendations of the irdirc data mining and repurposing task force. *Orphanet Journal of Rare Diseases, 14*(1), 225.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research, 110*, 102817.

Zaťková, T. Š., Seberini, A., & Tokovska, M. (2025). First year university life: Expectations versus reality. *International Journal of Instruction, 18*(3), 335-352.