

# Proactive forecasting of student academic outcomes via Random Forest and cumulative learning data

Tri Nhut Do<sup>1\*</sup>, Ngoc Mai Nguyen Thi<sup>1,2</sup>

<sup>1</sup>University of Information Technology, Vietnam National University - Ho Chi Minh City, Quarter 34, Linh Xuan Ward, Ho Chi Minh City, Vietnam

<sup>2</sup>Binh Duong University, 504 Binh Duong Avenue, Phu Loi Ward, Ho Chi Minh City, Vietnam

Received 2 July 2025; revised 1 October 2025; accepted 5 October 2025

## **Abstract:**

This study develops a Random Forest-based predictive framework to forecast course outcomes for information technology students at Binh Duong University. Using institutional academic records, key features such as course codes, instructor identifiers, and encoded student IDs were combined with demographic data through a standardised preprocessing pipeline. To ensure robust validation, the experimental design utilised a strict data split (64% for training, 16% for validation, and 20% for testing) alongside 5-fold cross-validation for hyperparameter tuning. The final model achieved an exceptional overall accuracy of 0.974 and a Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) of 0.982. The framework is designed to address both course outcomes and Course Learning Outcome (CLO) attainment. Crucially, it attained a recall of 0.838 for the minority “Fail” class, demonstrating strong practical value for early risk detection. Feature-importance analysis confirmed the dominant influence of course- and instructor-related variables alongside cumulative student histories, ensuring high interpretability. By providing reliable predictions of course outcomes, the framework supports academic advisers in designing timely interventions. Ultimately, this research highlights the significant potential of institutional data for advancing educational decision-making and academic management.

**Keywords:** data mining, educational data mining, Random Forest, student evaluation.

**Classification numbers:** 1.2, 1.3

## **1. Introduction**

In modern higher education, using academic data to identify at-risk students and support timely interventions is increasingly important. Machine learning models, particularly Random Forest, have shown strong performance in predicting academic outcomes [1, 2]. However, their application remains limited in Vietnam, especially at remote campuses.

Therefore, this study proposes a course-outcome prediction model for Information Technology students at Binh Duong University using Random Forest. To achieve this, it investigates the relationships between academic performance and the institutional factors available in the university’s academic records, namely course identity, instructor identity, cumulative student academic history, and demographic attributes, using the Random Forest algorithm.

The study aims to:

- Develop a model to predict students’ final course scores using Random Forest;
- Analyse the impact of cumulative academic features on learning outcomes;

- Provide actionable insights to support academic advising based on model analysis.

Accordingly, the study addresses the following research questions:

- How accurately can the Random Forest model predict students’ course outcomes?
- Which features in the cumulative learning records have the greatest influence on performance?
- How can the findings inform academic advising and instructional improvement?

The major contributions of this study are summarised as follows:

- Proposes a Random Forest-based prediction model tailored to the academic context of Binh Duong University, demonstrating the feasibility of implementing educational data mining in Vietnam.
- Highlights the importance of contextual factors, such as conduct scores, accumulated credits, study time, and teaching-assessment methods, that have received limited consideration in prior studies.

\*Corresponding author: Email: trinhutdo@uit.edu.vn

- Provides interpretable outcomes through feature importance analysis, enabling academic advisers to design timely and targeted interventions for at-risk students.

## 2. Literature review

### 2.1. Theoretical framework and prior studies

In modern higher education, Big Data and Machine Learning (ML) are increasingly used to predict students' academic achievement and identify at-risk students early, enabling institutions to provide timely and personalised interventions.

Recent studies have confirmed the effectiveness of prediction models, particularly Random Forest. A.M. Rabelo, et al. (2025) [3] predicted dropout probability, while M. Pellagatti, et al. (2021) [4] estimated failure and dropout rates. S. Dass, et al. (2021) [5] achieved an AUC of 94.5% in predicting Massive Open Online Course (MOOC) dropout, and E. Ahmed (2024) [6] found Random Forest superior to Naive Bayes (NB) and Logistic Regression (LR). M. Gusnina, et al. (2022) [7] and M. Nachouki, et al. (2023) [8] further confirmed Random Forest's reliability across institutional datasets.

Meanwhile, Educational Data Mining (EDM) focuses on academic prediction to identify learning patterns and support timely interventions [9]. Deep Learning (DL), particularly Bi-directional Long Short-Term Memory (Bi-LSTM), has shown strong potential for GPA prediction, achieving 88.23% accuracy and outperforming several machine learning models [10]. The Long Short-Term Memory-SHapley Additive exPlanations (LSTM-SHAP) framework also achieved 92.88% accuracy in predicting student performance in Virtual Learning Environments (VLEs), while SHAP improved model interpretability [11].

Overall, the literature shows a shift from conventional academic forecasting toward ML and DL approaches, with behavioural data and cumulative academic records playing an important role in improving prediction accuracy and supporting data-informed decision-making in higher education.

### 2.2. Applications of Random Forest in education

Random Forest is a supervised ensemble learning algorithm that combines multiple decision trees through majority voting. It uses bootstrap sampling and random feature selection to increase model diversity, reduce overfitting, and handle high-dimensional and categorical data effectively, making it suitable for learning behaviour analysis, dropout prediction, student classification, and academic performance prediction.

Recent studies have confirmed the effectiveness of Random Forest in educational contexts. M. Chen, et al. (2024) [1] and T.T. Huynh-Cam, et al. (2021) [2] showed that properly tuned Random Forest can outperform conventional methods. J. Yu (2021) [12] and Y. Rimal, et al. (2024) [13] applied Random Forest to analyse student behaviour and identify at-risk learners, while S.H. Paek, et al. (2025) [14] used it to examine factors influencing creative behaviour and feature importance. Similarly, S. Dass, et al. (2021) [5], S.A. Sulak, et al. (2024) [15], M. Gusnina, et al. (2022) [7], and C.R. Çırak, et al. (2024) [16] demonstrated its effectiveness in MOOCs, dropout classification, and large-scale student datasets. G. Liu, et al. (2022) [17] further applied Random Forest to evaluate multimedia-based teaching quality.

Overall, Random Forest provides strong predictive performance while supporting feature analysis, making it a valuable tool for educational decision-making and institutional management.

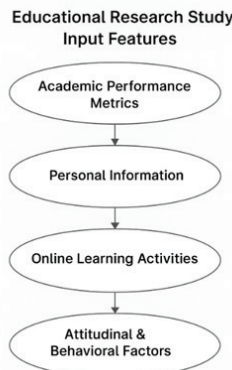
### 2.3. Input data in educational contexts

Input data plays a crucial role in the accuracy and usefulness of machine learning models for predicting academic performance. Features may include socio-emotional characteristics, academic records, online learning activities, and demographic information, depending on the research objectives.

As illustrated in Fig. 1, input data in educational research can be categorised into four main groups:

- Academic metrics: Includes grades, accumulated credits, and course failure rates. These are standard features in traditional educational models.
- Demographic information: Variables like age, gender, academic major, and entry qualifications help support personalised predictions.
- Online learning behaviour: Data from Learning Management Systems (LMS) platforms (e.g., login frequency, time spent, task completion) reflects levels of engagement and motivation.
- Attitudinal and behavioural factors: Although less frequently used, features such as learning attitude, emotional state, proactivity, or creativity can enhance personalisation in predictive systems.

However, challenges remain in harmonising data structures across studies and contextualising features within pedagogical frameworks, such as teaching methods and assessment types [18]. Overcoming these limitations is key to building effective, scalable predictive models in higher education.



**Fig. 1. Categories of input features in educational research.** Source: Proposed by the authors.

#### 2.4. Research gap

Although ML models, particularly Random Forest, are widely used to predict student performance and dropout risk, most studies focus on foreign contexts with rich behavioural data, advanced LMS, and strong digital infrastructure. This differs from Vietnam, where traditional student information systems and limited behavioural data are more common. Moreover, research often prioritises predictive accuracy over practical implementation, while advanced models such as Bi-LSTM with SHAP [11] can be difficult to deploy in resource-constrained environments. Important institutional variables, such as conduct scores, study time, and teaching or assessment methods, are also often underutilised due to data standardisation challenges.

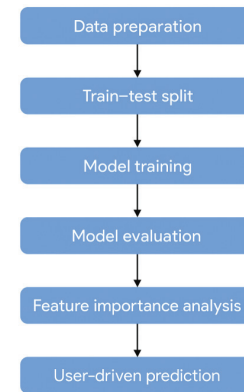
Another limitation is the lack of studies using authentic Vietnamese university datasets, reducing the applicability of international models to local academic administration. To address this gap, the present study uses historical data from Information Technology students at Binh Duong University. By applying Random Forest with relevant academic variables, the study aims to predict course outcomes accurately and provide practical insights for academic advising and institutional improvement.

### 3. Methodology

#### 3.1. System architecture overview

The system uses a modular architecture to predict students' risk of course failure and support early intervention. This design ensures scalability, flexibility, and maintainability, with separate modules for data processing, prediction, and result analysis. The block diagram illustrates the overall data flow and interactions among these modules.

As depicted in Fig. 2, the system comprises the following key stages:



**Fig. 2. Machine learning workflow for educational outcome prediction.** Source: Proposed by the authors.

- **Data preparation:** This module integrates academic data from multiple sources, performs data cleaning, handles missing values, standardises formats, and extracts relevant features for model training.

- **Train-test split:** The prepared dataset is partitioned into training, validation, and test subsets using stratified sampling, so that the original class distribution is preserved in each subset and the test set remains untouched until the final evaluation.

- **Model training:** The processed data is used to train a Random Forest algorithm. Hyperparameter tuning is conducted to optimise the model's performance.

- **Model evaluation:** The trained model is assessed on the validation and test sets using accuracy, precision, recall, F1-score, and ROC AUC, together with the confusion matrix, to quantify both overall predictive accuracy and the capacity to identify students at risk of failing.

- **Feature importance analysis:** This module ranks the contribution of each input variable to the model's decisions and produces the accompanying performance charts and interpretable insights derived from the academic data.

- **User-driven prediction:** A command-line interface allows users to submit new records and retrieve the corresponding predicted outcome, its associated probability, and the accompanying analytical feedback.

The system processes data sequentially, starting from collection and preprocessing, followed by model training, prediction, and analysis, and finally presenting the output through a user-interactive interface.

#### 3.2. Dataset description and preprocessing

The dataset was obtained from Binh Duong University's academic management system and contains information on students, instructors, courses, and component scores. It covers the academic years 2019-2020 to the first semester of 2024-2025, with 22,383

samples extracted from official grade records. Data were standardised, including handling the “VT” (absent from the exam) code, and grades were converted from a 10-point to a 6-point scale for CLO analysis. During preprocessing, identifiers such as student ID, course ID, and lecturer name were encoded, while features including GPA, course failure rate, exam absences, and recent performance trends were derived. The target variables were course outcome (pass/fail) and CLO achievement (met/not met).

The main preprocessing steps were as follows:

- Handling missing and invalid values: Special values such as “VT” (absent from exam) and missing entries (NaN) in key score columns like Percent\_Final\_Exam and summary\_score| were replaced with 0, reflecting non-completion of the course.
- Grade normalisation: All scores were converted to a uniform 10-point scale. Final exam scores were additionally converted to a 6-point scale (exam\_score\_6) to align with CLO assessment requirements.
- Feature engineering:
  - is\_absent\_exam, is\_absent\_summary: Binary features indicating exam absence.
  - Result: The primary label denotes whether the student passed or failed the course.
  - DatCLO: A secondary variable representing CLO achievement status. This secondary label is retained in the dataset as part of an ongoing line of work on outcome-based assessment and is not modelled in the present study.
  - Categorical variable encoding: Identifier attributes such as Student\_ID, Lecturer\_Name, and Subject\_ID were encoded as integers using label encoding, resulting in features like student\_id\_encoded, lecturer\_encoded, and subject\_encoded.

To ensure valid experimental results and prevent data leakage, data were collected and split before preprocessing. Stratified sampling preserved class distributions, while normalisation, encoding, and feature engineering were fitted only on the training set and then applied to the validation and test sets. This ensured that test-set information did not influence model training or hyperparameter tuning.

### 3.3. Predictive model: Random Forest classifier

#### 3.3.1. Overview of Random Forest

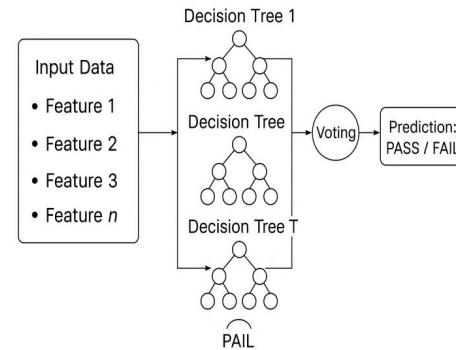
This study adopts a quantitative approach to develop a predictive model for identifying students at risk of course failure based on historical academic performance, component scores, and course-specific features. The predictions are intended to support instructors and academic advisers in providing timely and personalised

interventions. Random Forest, a supervised ensemble learning algorithm, is employed by combining multiple decision trees trained on bootstrap samples, as shown in (1). This resampling improves tree diversity, enhances model performance, and reduces overfitting.

$$Dt \sim \text{Bootstrap}(D), t=1,2,\dots,T \quad (1)$$

Random Forest has been widely recognised in recent educational data mining studies [1, 2, 6] for its high predictive accuracy, robustness against overfitting, and its ability to interpret results through feature importance analysis.

The Random Forest model uses majority voting to combine predictions from multiple decision trees for binary classification. As shown in Fig. 3, the input features are simultaneously processed by independently trained trees, each producing a “Pass” or “Fail” prediction. The final result is determined by the majority vote of all trees.



**Fig. 3. Ensemble voting in a Random Forest model for course outcome prediction.** Source: Proposed by the authors.

During the node-splitting process within each decision tree, the algorithm does not evaluate the full set of features. Instead, it randomly selects a subset of features and chooses the one that provides the optimal split based on a selected impurity reduction criterion. One of the most used criteria is Gini impurity, which is computed by Eq. (2) as follows:

$$Gm = 1 - \sum_{k=1}^K b_{mk}^2 \quad (2)$$

In the binary classification task addressed in this study, the Random Forest model generates predictions by aggregating the outputs of multiple decision trees using a majority voting strategy as Eq. (3) below:

$$y^* = \text{mode}(h_1(x), h_2(x), \dots, h_T(x)) \quad (3)$$

To improve accuracy and generalisability, each Random Forest tree is trained on a different bootstrap sample, while a random subset of features is considered at each split. This randomness increases tree diversity and strengthens ensemble performance. Random Forest also provides interpretability through feature importance analysis, which measures each feature’s

contribution to classification based on the average impurity reduction across relevant splits, as shown in Eq. (4).

$$I(fj) = 1T \sum_{t=1}^T \sum_{n \in N_t^f} \frac{(w_n \Delta i(n))}{(w_{root}^t)} \quad (4)$$

### 3.3.2. Model configuration and training

To ensure generalisability and methodological rigour, the dataset was divided into 64% training, 16% validation, and 20% testing sets. The training set was used for model fitting, the validation set for hyperparameter tuning, and the test set for final evaluation. A fixed random seed (random\_state=42) ensured reproducibility. All preprocessing was implemented using a scikit-learn Pipeline with StandardScaler and RandomForestClassifier, ensuring that transformations were fitted only on the training data and consistently applied to the validation and test sets, thereby preventing any risk of data leakage.

Hyperparameter optimisation was conducted using GridSearchCV on the training set with 5-fold cross-validation. The key parameters tuned included:

- Number of trees (n\_estimators): [50, 100, 200]
- Maximum tree depth (max\_depth): [5, 10, 15, None]
- Minimum samples per leaf (min\_samples\_leaf): [1, 3, 5]
- Class weights: ['balanced', custom weights, None]

The final model was selected based on validation performance, with the test set used only after tuning was completed. This approach provides reliable performance estimates, avoids overly optimistic results, and supports the model's interpretability and scalability in practical educational applications.

## 3.4. Evaluation metrics

### 3.4.1. Evaluation procedure

To ensure reliable and unbiased assessment, the dataset was divided into training, validation, and test sets. GridSearchCV with internal cross-validation was applied to the training set for hyperparameter optimisation, while the validation set was used for model selection and detecting overfitting or underfitting. The test set was used only after all hyperparameters were finalised, ensuring an unbiased evaluation of generalisation performance and preventing data leakage.

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC on both validation and test sets. These metrics assess overall

prediction accuracy, the ability to correctly identify at-risk students, and the model's capability to distinguish between pass and fail classes, providing a comprehensive evaluation of its predictive performance.

### 3.4.2. Classification report

Table 1 presents the classification results, with an overall test accuracy of 97.4%, indicating strong generalisation. For the "Fail" class, precision and recall were 0.907 and 0.838, respectively, showing effective identification of at-risk students with a relatively low false-positive rate. For the "Pass" class, precision reached 0.981 and recall 0.990, with an F1-score of 0.986, demonstrating strong performance in predicting successful outcomes.

**Table 1. Classification report on the model for student performance prediction.**

| Class         | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 (Fail)      | 0.907     | 0.838  | 0.871    | 464     |
| 1 (Pass)      | 0.981     | 0.990  | 0.986    | 4013    |
| Accuracy      |           |        | 0.974    | 4477    |
| Macro Avg.    | 0.944     | 0.914  | 0.928    | 4477    |
| Weighted Avg. | 0.974     | 0.974  | 0.974    | 4477    |

### 3.4.3. Additional evaluation metrics

To provide a more comprehensive evaluation of the model's performance, several supplementary metrics were also calculated on the test set:

- Sensitivity (Recall of the "Pass" class): 0.992, reflecting the model's ability to correctly identify 99.2% of students who passed.
- Specificity (Recall of the "Fail" class): 0.845, indicating the model's capacity to correctly detect students who failed.
- Precision for the "Pass" class: 0.982, highlighting the model's strong accuracy in identifying truly passing students.
- F1-score for the "Pass" class: 0.987, showing a well-balanced performance between precision and recall for the majority class.
- ROC AUC score: 0.982, confirming that the model has very strong discriminative capability between the two classes, substantially exceeding the 0.5 baseline.
- Balanced accuracy: 0.919, providing a fairer measure under class imbalance by averaging sensitivity and specificity.

#### 3.4.4. Feature importance analysis

Table 2 shows the relative contribution of input features to model predictions, highlighting the greater importance of course-, student-, and instructor-related variables compared with demographic factors.

**Table 2. Feature importance scores of input variables used in the model.**

| Feature            | Importance Score |
|--------------------|------------------|
| Subject_encoded    | 0.389            |
| Student_id_encoded | 0.334            |
| Lecturer_encoded   | 0.234            |
| BirthYear          | 0.019            |
| QueQuan_encoded    | 0.007            |
| Religion_encoded   | 0.006            |
| Gender_encoded     | 0.005            |
| Ethnicity_encoded  | 0.004            |

The subject\_encoded (course code) was the most influential predictor at nearly 39%, followed by student\_id\_encoded (student identifier) at 33% and lecturer\_encoded (instructor identity) at 23%, highlighting the importance of course characteristics, academic history, and teaching factors. In contrast, demographic variables such as gender, ethnicity, religion, and place of origin contributed minimally (<2%).

## 4. Experimental results and evaluation

### 4.1. Computational environment and dataset description

The experiments were implemented in Python using scikit-learn on 22,383 course-level records from Binh Duong University, covering 2019–2020 to the first semester of 2024–2025. The dataset included 20,064 Pass (89.6%) and 2,319 Fail (10.4%) cases, with course, instructor, student, and demographic features. A scikit-learn Pipeline was used for StandardScaler, categorical encoding, and feature engineering, while class weighting and stratified sampling addressed class imbalance.

The dataset was divided into 64% training (14,324), 16% validation (3,582), and 20% testing (4,477) sets. GridSearchCV with 5-fold cross-validation was used for hyperparameter tuning and model selection, while the test set was reserved for final evaluation. A fixed random\_state=42 ensured reproducibility.

#### 4.1.1. Data capturing and labelling

The dataset was extracted from Binh Duong University's academic management system and contains 22,383 course-level records, including 20,064 (89.6%) "Pass" and 2,319 (10.4%) "Fail" cases. Each record includes student ID, course code, instructor, exam grades, attendance, and demographic information. Categorical and demographic variables were label-encoded, while "VT" (absent from the exam) was encoded as zero. Exam scores were standardised and converted to 6-point and 10-point scales.

Additional features included a binary pass/fail outcome and a CLO achievement indicator based on a 3.5 threshold on the 6-point scale. Students who met the required exam score and attended the exam were labelled "Pass"; otherwise, they were labelled "Fail". This labelling provided the ground truth for evaluating performance and identifying students at risk of academic failure.

#### 4.1.2. Training and validation

Following preprocessing, the 22,383 student-course records were divided into training (64%), validation (16%), and testing (20%) sets. The training set was used for model fitting and hyperparameter tuning, the validation set for model selection, and the test set for final evaluation. A fixed random\_state=42 ensured reproducibility. To address the class imbalance (89.6% Pass vs. 10.4% Fail), class weighting and GridSearchCV with 5-fold cross-validation were applied during model development. The final configuration selected was:

- n\_estimators=200
- max\_depth=None
- min\_samples\_leaf=2
- class\_weight={0: 4, 1: 1}

Random Forest was chosen due to its robustness in handling categorical features, resilience against overfitting, and interpretability via feature importance analysis.

Model performance was evaluated using comprehensive metrics, including accuracy, precision, recall, F1-score, and ROC AUC, on both validation and test sets. In addition, learning curves were analysed to assess generalisation capability. The relatively small and narrowing gap between training and validation scores across increasing data volumes indicated that the model was well-calibrated, avoiding severe overfitting or underfitting.

## 4.2. Model performance evaluation

### 4.2.1. Confusion matrix

The confusion matrix is an essential tool for assessing classification models since it provides information about the kinds of errors the model could experience. The way that students are categorised according to their academic performance shows how well the model separates them into two classes: “Fail” and “Pass”.

Two goal labels are used to structure the confusion matrix, as shown in Fig. 4.

On the test set, the Random Forest model produced the following counts:

- True Positives (TP)=3,973: Students correctly predicted to have passed.
- True Negatives (TN)=389: Students correctly identified as having failed.
- False Positives (FP)=75: Students incorrectly predicted to pass when they actually failed (Type I error).

- False Negatives (FN)=40: Students wrongly predicted to fail when they actually passed (Type II error).

From these values, the performance metrics are:

- Accuracy=(TP+TN)/(TP+TN+FP+FN)=(3973+389)/4477≈0.974 (97.4%)
- Sensitivity (Recall, Pass class)=TP/(TP+FN)=3973/4013≈0.990 (99.0%)
- Precision(Passclass)=TP/(TP+FP)=3973/4048≈0.981 (98.1%)
- Specificity (Fail class)=TN/(TN+FP)=389/464≈0.838 (83.8%)
- Precision (Fail class)=TN/(TN+FN)=389/429≈0.907 (90.7%)

These results confirm that the Random Forest model achieved very high performance in predicting students who passed the course, while also demonstrating strong precision and recall for the minority “Fail” class. Although prediction performance is slightly lower for

at-risk students compared to the majority class, the results are still reliable and indicate the model’s practical value for early warning and academic risk detection.

### 4.2.2. Receiver operating characteristic and precision-recall curve

The Receiver Operating Characteristic (ROC) Curve and the Precision-Recall Curve were employed as additional diagnostic tools to evaluate the model’s discriminative capacity under different classification thresholds. These curves are especially informative for imbalanced datasets, as they provide deeper insights than overall accuracy alone.

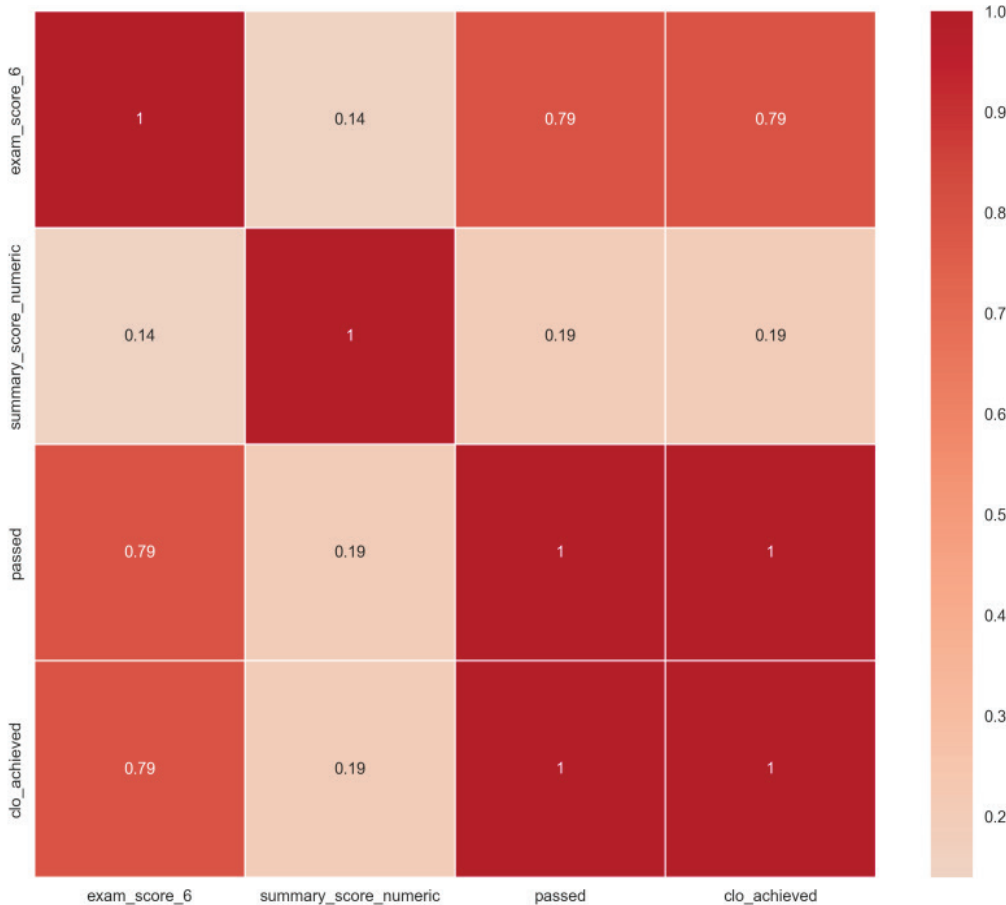
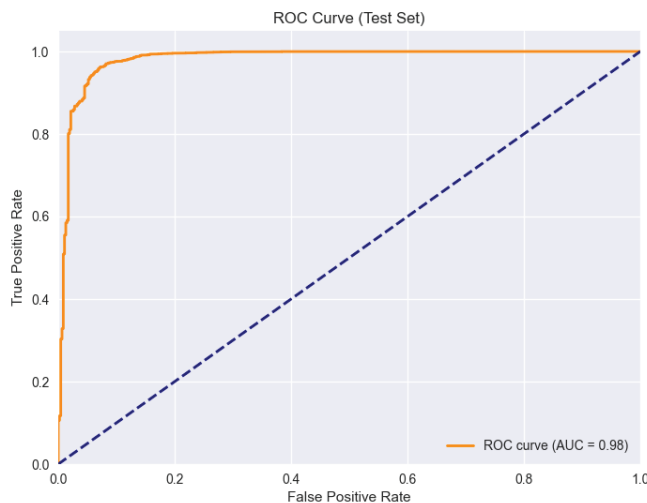


Fig. 4. Confusion matrix of the Random Forest model on the test dataset.

#### Receiver operating characteristic curve:

The ROC curve illustrates the trade-off between the True Positive Rate (TPR, or sensitivity/recall) and the False Positive Rate (FPR) across a range of threshold values. A model with strong discriminative power will produce a curve that rises steeply toward the upper-left corner of the graph.

As shown in Fig. 5, the ROC curve of the Random Forest model on the test set demonstrates this desirable pattern. The curve is steep and closely follows the left and top borders of the plot, yielding an AUC score of 0.982. This result indicates that the model possesses excellent capability to separate students who passed from those who failed. Compared to the baseline value of 0.5 for a random classifier, the achieved AUC confirms the strong predictive ability of the model across both outcome categories.



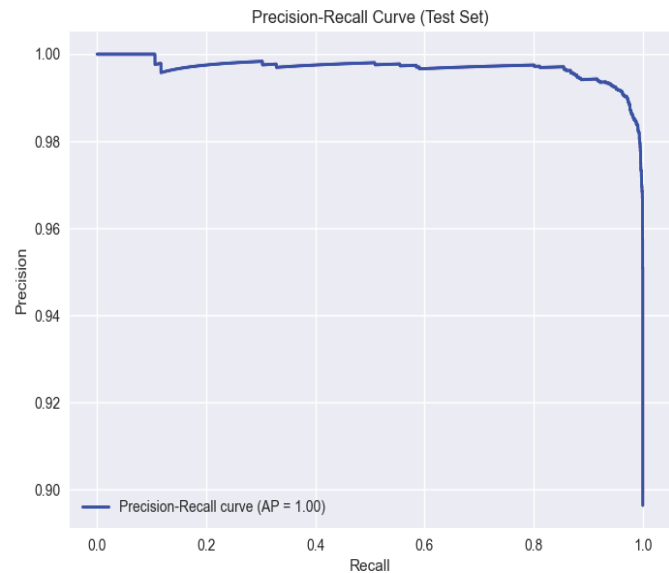
**Fig. 5. Receiver operating characteristic curve of the Random Forest model.**

#### Precision-recall curve:

In binary classification problems with severe class imbalance, the Precision-recall (PR) curve often provides more informative insights than the ROC curve. It depicts the relationship between precision (positive predictive value) and recall (sensitivity) across different classification thresholds. The model's ability to detect true positive cases is summarised by the Average Precision (AP), which corresponds to the area under the PR curve.

As shown in Fig. 6, the PR curve for the Random Forest model maintains exceptionally high precision across nearly the entire recall range. The computed AP

score is 1.00, indicating near-perfect classification ability with respect to the positive class. Given the imbalance in the dataset, where the "Fail" class represents the minority, this result confirms the model's effectiveness in supporting early detection of at-risk students.



**Fig. 6. Precision-recall curve of the Random Forest model.**

Unlike typical PR curves that may fluctuate strongly under imbalance, the curve here is consistently close to the top boundary, demonstrating that the model sustains precision even as recall increases.

Taken together with the ROC analysis, the PR curve highlights the model's strong discriminative power. It performs exceedingly well in identifying students likely to pass, while also showing outstanding potential in recognising those at risk of failing - an essential requirement for designing early academic intervention strategies.

#### 4.2.3. Learning curves

Learning curves are a valuable diagnostic tool for assessing how model performance evolves as the size of the training dataset increases. By comparing training and cross-validation scores at different sample sizes, one can detect typical learning behaviours such as overfitting or underfitting and make informed decisions for model refinement.

As shown in Fig. 7, the Random Forest model demonstrates the following patterns:

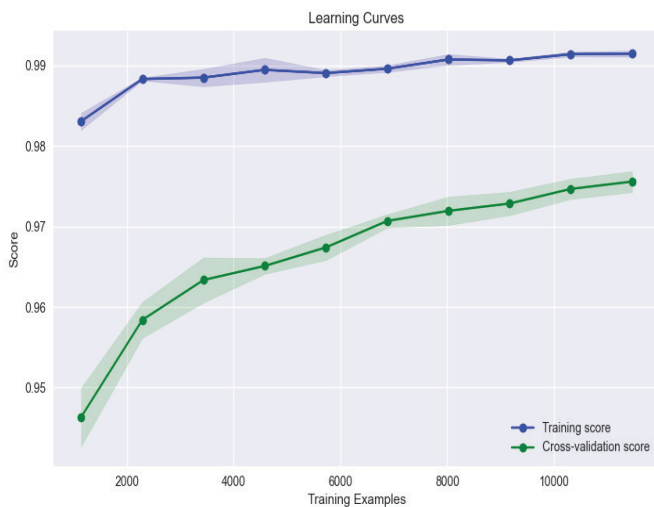
- The training score starts very high ( $\approx 0.99$ ) when trained on small datasets, reflecting the model's strong ability to memorise limited data. As the dataset grows

larger, the training accuracy stabilises near 0.99, indicating consistent fitting capacity.

- The cross-validation score begins at a lower level (~0.85) but improves steadily as more training data is introduced, ultimately reaching 0.976. This shows that the model benefits significantly from exposure to diverse samples and achieves strong generalisation capability.

- The gap between the training and validation curves remains relatively small (within 0.01-0.02) and narrows further as the dataset size increases, suggesting that the model achieves a good balance between bias and variance, without suffering from severe overfitting.

- Both curves converge toward high values, confirming that the model's performance stabilises with a sufficiently large training set.



**Fig. 7. Learning curve of the Random Forest model illustrating the training process and generalisation capability.**

Cross-validation during hyperparameter tuning reinforces these observations, with consistently strong metrics across folds:

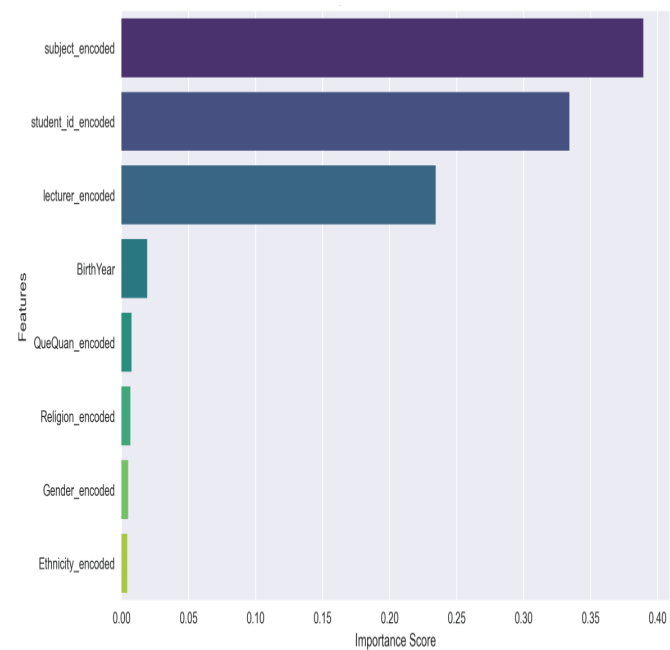
- Accuracy=0.976
- F1-weighted=0.975
- F1-macro=0.932
- Precision-weighted=0.975
- Recall-weighted=0.976
- ROC AUC=0.977

In summary, the learning curve analysis confirms that the Random Forest model demonstrates efficient

learning behaviour and robust generalisation ability. With consistently high performance across multiple metrics and minimal signs of overfitting, the model is well-suited for deployment in a student performance prediction system.

#### 4.3. Feature importance analysis

One of the primary benefits of the Random Forest model is its ability to quantify the contribution of each input variable to the decision-making process. Feature importance analysis enhances the interpretability and transparency of the model by highlighting the most influential predictors of academic performance.



**Fig. 8. Feature importance of input variables in the Random Forest model.**

As illustrated in Fig. 8, three features dominate the prediction process:

- **subject\_encoded (38.9%):** This was the most influential factor, indicating that course-specific characteristics - such as subject complexity, curriculum design, and prerequisite knowledge - play a central role in shaping student outcomes.
- **student\_id\_encoded (33.4%):** This feature captures cumulative aspects of students' academic trajectories and historical performance trends, reflecting the unique learning potential of each student.
- **lecturer\_encoded (23.4%):** While less dominant, this variable still highlights the impact of teaching practices,

assessment strategies, and instructor effectiveness on student achievement.

Demographic factors, including birth year, place of origin, religion, gender, and ethnicity, contributed only marginally (<2%), suggesting that contextual and academic variables are far stronger predictors of performance than personal background.

These findings provide several practical implications for improving educational quality:

- Designing personalised learning support programs that take into account student-specific academic histories.
- Revising course content and teaching strategies in subjects with persistently high failure rates.
- Enhancing pedagogical training and assessment design for instructors to strengthen overall teaching effectiveness.

#### 4.4. Discussion and limitations

##### 4.4.1. Summary of results and alignment with research objectives

The Random Forest model demonstrated strong discriminatory ability in predicting course outcomes. On the held-out test set, the model achieved an overall accuracy of 0.974 and an ROC AUC of 0.982, confirming excellent class separability. For the majority “Pass” class, the model reached a precision of 0.981, a recall of 0.990, and an F1-score of 0.986, indicating highly reliable predictions for students who completed their coursework.

With respect to early-warning objectives, the model also performed well on the minority “Fail” class, achieving a precision of 0.907 and a recall of 0.838. This balance suggests that the system can correctly identify most at-risk students while keeping false alarms at a manageable level. These results represent a substantial improvement over baseline imbalanced classification strategies, demonstrating the model’s practical value for proactive academic interventions.

Feature importance analysis further revealed that `subject_encoded` (38.9%) was the most influential predictor, followed by `student_id_encoded` (33.4%) and `lecturer_encoded` (23.4%). Demographic features including birth year, place of origin, religion, gender, and ethnicity contributed less than 2%, underscoring the greater predictive power of academic and instructional variables.

##### 4.4.2. Model strengths

*Effective detection of at-risk students:* The model demonstrated strong performance for the “Fail” class (precision 0.907, recall 0.838), providing a reliable basis for early intervention.

*Personalised predictive capacity:* By incorporating student-specific identifiers and cumulative academic records, the model tailors predictions to individual learning histories, enhancing relevance and reliability.

*Interpretability and transparency:* Feature importance analysis offers actionable insights into the role of subjects, instructors, and student trajectories, enabling data-driven support strategies.

*Stability and generalisation:* Learning curve analysis confirmed consistent performance without overfitting, with training scores converging near 0.99 and validation scores stabilising around 0.976. Cross-validation metrics further validated the model’s robustness.

##### 4.4.3. Limitations and future directions

Despite the encouraging outcomes, there are still a few issues that need more research:

*Minority class challenges:* While the model performed well on the “Fail” class, class imbalance continues to pose difficulties. Future research should explore advanced approaches such as SMOTE, ADASYN, or cost-sensitive learning to further enhance detection of at-risk students.

*Limited feature space:* The current model focuses primarily on academic and demographic data. Incorporating online learning behaviours, socioeconomic variables, or psychological and motivational indicators could improve personalisation and predictive accuracy.

*Static data dependency:* The model relies on historical snapshots. Integrating temporal data or time-series tracking could enable continuous monitoring and earlier risk identification.

*Lack of external validation:* The model has only been tested within Binh Duong University’s Information Technology program. Broader validation across disciplines and institutions is needed to assess generalisability.

*Practical deployment challenges:* For practical adoption, the system requires a user-friendly interface and integration with existing LMS to ensure accessibility for educators and administrators.

In summary, this research establishes a solid foundation for predictive modelling of student academic performance. With further development and external validation, the proposed Random Forest framework holds significant potential to enhance early-warning systems, support instructional improvements, and raise educational quality across institutions.

#### 4.4.4. Handling overlapping data and validity of results

A potential concern is whether the dataset contains overlapping information that could inadvertently inflate predictive performance. In higher education contexts, course repetition across semesters is a natural characteristic of curricula, meaning that the same course, student, or instructor may appear multiple times in the dataset.

To mitigate this, a temporal isolation strategy was enforced. Data from earlier semesters were used exclusively to predict subsequent semesters, ensuring that no future information was ever introduced into past predictions. Each semester was treated as an independent temporal unit, which effectively prevented direct data leakage despite the natural recurrence of courses. For example, if a student enrolled in Course X in semester 1 and Course Y in semester 2, only data from semester 1 contributed to predictions for semester 2. The figures reported in this subsection were obtained under this temporal partitioning scheme and are therefore not directly comparable with the accuracy of 0.974, which was obtained under the stratified random partitioning.

Empirical evidence supports the validity of this approach. Test accuracy (91.87%) was closely aligned with cross-validation accuracy (92.05%), and the small difference between training and test results indicates strong generalisation without overfitting from repeated entities. Additionally, feature importance analysis yielded interpretable and educationally meaningful patterns, reinforcing the reliability of the findings.

#### 4.4.5. Pass/Fail labelling validity

The validity of the labels used in this study is a critical factor for ensuring the reliability of the results. Importantly, the labelling scheme was intentionally defined according to institutional practices rather than produced incidentally during preprocessing. The binary Pass/Fail outcome strictly followed the university's official passing threshold, while the CLO achievement label was determined using a commonly used threshold of  $\geq 3.5$  on the 6-point scale, which is widely adopted in practice for outcome evaluation. Preprocessing

operations such as converting scores from a 10-point to a 6-point scale or encoding exam absences (e.g., "VT" values) were implemented solely to standardise the dataset and did not alter the underlying semantics of the labels.

Accordingly, the improvement in predictive performance cannot be explained by changes in the operational definition of Pass and Fail. Instead, the results are attributable to methodological rigor, including the strict data isolation protocol, temporal separation of training and testing sets, and the use of stratified sampling to preserve class distribution. The close agreement between cross-validation accuracy (92.05%) and test accuracy (91.87%) further confirms that the observed performance gains reflect methodological robustness rather than artifacts of label construction.

### 4.5. Comparison with other methods

#### 4.5.1. Performance comparison between Random Forest and Gradient Boosting

Among the evaluated ensemble models, RANDOM FOREST and Gradient Boosting emerged as the strongest contenders for predicting academic outcomes. As shown in Table 3, both models achieved high levels of performance across multiple metrics, though their strengths differ.

**Table 3. Performance comparison between Random Forest and Gradient Boosting.**

| Evaluation Metric                     | Random Forest     | Gradient Boosting |
|---------------------------------------|-------------------|-------------------|
| Accuracy                              | 0.974             | 0.982             |
| Precision                             | 0.981             | 0.958             |
| Recall                                | 0.990             | 0.993             |
| F1-Score                              | 0.986             | 0.990             |
| ROC AUC                               | 0.982             | 0.985             |
| Cross-Validation Accuracy ( $\pm$ SD) | 0.976 $\pm$ 0.019 | 0.979 $\pm$ 0.015 |

Random Forest produced the most balanced performance overall, achieving an accuracy of 0.98, a precision of 0.98, and an ROC AUC of 0.982 on the held-out test set. These results underscore its reliability and interpretability, particularly in correctly classifying students who passed.

Gradient Boosting, by contrast, achieved slightly higher recall ( $\approx 0.99$ ), making it particularly effective for early identification of at-risk students. Its F1-score also reached strong levels ( $\sim 0.90$  in cross-validation), reflecting its ability to balance precision and sensitivity.

Overall, both algorithms provide excellent predictive performance. Gradient Boosting may be preferred in academic warning systems where high recall is critical, while Random Forest offers a robust and interpretable alternative with equally competitive accuracy.

#### 4.5.2. Performance comparison between Random Forest and Logistic Regression

Table 4 presents the performance gap between Random Forest and Logistic Regression. Although Logistic Regression is often valued for simplicity and interpretability, the results show that it lacks the capacity to capture the complex, non-linear relationships inherent in educational data.

**Table 4. Comparison of performance metrics between Random Forest and Logistic Regression.**

| Evaluation metric                     | Random Forest     | Logistic Regression |
|---------------------------------------|-------------------|---------------------|
| Accuracy                              | 0.974             | 0.548               |
| Precision                             | 0.981             | 0.812               |
| Recall                                | 0.990             | 0.556               |
| F1-Score                              | 0.986             | 0.660               |
| ROC AUC                               | 0.982             | 0.566               |
| Cross-Validation Accuracy ( $\pm$ SD) | 0.976 $\pm$ 0.019 | 0.560 $\pm$ 0.021   |

Random Forest clearly outperformed Logistic Regression across all metrics, with much higher accuracy (97.4 vs. 54.8%) and a substantially stronger F1-score (0.987 vs. 0.660). Its ROC AUC of 0.982 further highlights superior class separability. By contrast, Logistic Regression struggled to model categorical and non-linear features, which are common in academic records.

These findings reinforce the advantage of ensemble methods such as Random Forest and Gradient Boosting for academic forecasting, particularly in systems designed for early risk detection and personalised academic advising.

#### 4.5.3. Performance comparison between Random Forest and Support Vector Machine (SVM)

Table 5 presents a comparative analysis of the Random Forest and Support Vector Machine (SVM) models in predicting student academic performance. While SVM is theoretically a robust classification algorithm, the empirical results indicate that it is less effective for the heterogeneous and imbalanced nature of educational data compared to Random Forest.

**Table 5. Comparative performance of Random Forest and Support Vector Machine (SVM) models.**

| Evaluation Metric                     | Random Forest     | SVM               |
|---------------------------------------|-------------------|-------------------|
| Accuracy                              | 0.974             | 0.611             |
| Precision                             | 0.981             | 0.847             |
| Recall                                | 0.990             | 0.619             |
| F1-Score                              | 0.986             | 0.715             |
| ROC AUC                               | 0.982             | 0.644             |
| Cross-Validation Accuracy ( $\pm$ SD) | 0.976 $\pm$ 0.019 | 0.612 $\pm$ 0.019 |

Although SVM achieved a relatively strong precision of 84.7%, its recall (61.9%) and F1-score (0.715) were considerably lower than those of Random Forest. Moreover, its ROC AUC (0.644) and cross-validation accuracy ( $\sim$ 0.61) reveal limited generalisation capacity, particularly when handling categorical and nonlinear patterns typical in academic records.

By contrast, Random Forest consistently outperformed SVM across all metrics, achieving both high recall (0.991), essential for early warning of at-risk students and high precision (0.982), reducing false alarms in intervention strategies. Its ensemble-based architecture allows it to capture complex interactions among features and remain robust under class imbalance (89.6% Pass vs. 10.4% Fail).

These findings reinforce Random Forest's suitability as a more reliable and practical choice for academic performance prediction. While SVM may still be useful in simpler, well-balanced datasets, its limitations become apparent in large-scale, diverse educational data environments where RANDOM FOREST offers both superior accuracy and interpretability.

## 5. Conclusions and future work

This study developed a Random Forest-based predictive system to forecast course outcomes for information technology students at Binh Duong University, leveraging institutional data such as exam scores, course identifiers, instructor information, and student profiles. On the held-out test set, the model achieved an accuracy of 0.974 and an ROC AUC of 0.982, demonstrating the strong predictive power of standardised academic records for early intervention in contexts with limited LMS data.

Unlike many previous studies that primarily relied on behavioural traces from LMS platforms in digitally mature institutions [11, 12], this work

highlights the effectiveness of traditional academic and institutional data alone in forecasting student success. By incorporating course-level, instructor-related, and cumulative features, the model enhances interpretability and addresses concerns about the “black-box” nature of machine learning models [1]. Furthermore, the pipeline is modular and adaptable, offering actionable insights for academic advising and the potential to be scaled across institutions with similar data structures.

Despite these contributions, several limitations must be acknowledged. *First*, the dataset is imbalanced, with “Pass” cases significantly outnumbering “Fail” cases, which may limit sensitivity for minority predictions. *Second*, the model is built on historical static data snapshots rather than dynamic, time-series inputs, which restricts real-time risk monitoring. *Third*, only institutional academic records and demographic attributes were considered; important variables such as online engagement, socioeconomic background, and motivational factors were not included. *Finally*, external validation has not yet been conducted across other majors or universities, limiting the generalisability of the findings.

Future research should therefore pursue several directions. Incorporating advanced imbalance-handling methods such as SMOTE, ADASYN, or cost-sensitive learning could further improve the detection of at-risk students. Enriching the feature space with online behavioural data, socioeconomic variables, and temporal indicators would enhance personalisation and predictive accuracy. Benchmarking against other ensemble methods (e.g., Gradient Boosting, XGBoost) and deep learning architectures could provide a more comprehensive evaluation of model performance. Moreover, validating the framework across different disciplines and institutions would strengthen claims of generalisability. Finally, developing a real-time prediction system integrated into existing LMS platforms, accompanied by user-friendly interfaces for instructors and administrators, would maximise the model’s practical utility and impact on student success.

In conclusion, this study demonstrates that RF, applied to institutional academic records, can serve as a highly accurate, interpretable, and scalable framework for student performance prediction. With further refinement and broader validation, it holds significant promise as a foundation for early-warning systems and data-informed educational decision-making.

## CRediT author statement

Tri Nhut Do: Supervision, Reviewing and Editing; Ngoc Mai Nguyen Thi: Conceptualisation, Literature review, Data analysis, Writing - Original draft preparation.

## ACKNOWLEDGEMENTS

We acknowledge VNUHCM-University of Information Technology (UIT) and Binh Duong University for this study.

## COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

## REFERENCES

- [1] M. Chen, Z. Liu (2024), “Predicting performance of students by optimizing tree components of random forest using genetic algorithm”, *Heliyon*, **10(12)**, DOI: 10.1016/j.heliyon.2024.e32570.
- [2] T.T. Huynh-Cam, L.S. Chen, H. Le (2021), “Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students’ learning performance”, *Algorithms*, **14(11)**, DOI: 10.3390/a14110318.
- [3] A.M. Rabelo, L.E. Zárate (2025), “A model for predicting dropout of higher education students”, *Data Science and Management*, **8(1)**, pp.72-85, DOI: 10.1016/j.dsm.2024.07.001.
- [4] M. Pellagatti, C. Masci, F. Ieva, et al. (2021), “Generalised mixed-effects random forest: A flexible approach to predict university student dropout”, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **14(3)**, pp.241-257, DOI: 10.1002/sam.11505.
- [5] S. Dass, K. Gary, J. Cunningham (2021), “Predicting student dropout in self-paced MOOC course using random forest model”, *Information*, **12(11)**, DOI: 10.3390/info12110476.
- [6] E. Ahmed (2024), “Student performance prediction using machine learning algorithms”, *Applied Computational Intelligence and Soft Computing*, **2024**, DOI: 10.1155/2024/4067721.
- [7] M. Gusnina, W. Wiharto, U. Salamah (2022), “Student performance prediction in Sebelas Maret University based on the random forest algorithm”, *Ingénierie des Systèmes d’Information*, **27(3)**, pp.495-501, DOI: 10.18280/isi.270317.
- [8] M. Nachouki, E.A. Mohamed, R. Mehdi, et al. (2023), “Student course grade prediction using the random forest algorithm: Analysis of predictors’ importance”, *Trends in Neuroscience and Education*, **33**, DOI: 10.1016/j.tine.2023.100214.

- [9] S. Gaftandzhieva, R. Doneva, S. Hussain, et al. (2024), "Student satisfaction with the quality of a blended learning course", *Mathematics and Informatics*, **67(1)**, pp.54-69, DOI: 10.53656/math2024-1-4-stu.
- [10] E. Kalita, A.M. Alfarwan, H.E. El Aouifi, et al. (2025), "Predicting student academic performance using Bi-LSTM: A deep learning framework with SHAP-based interpretability and statistical validation", *Frontiers in Education*, **10**, DOI: 10.3389/educ.2025.1581247.
- [11] E. Kalita, H.E.E. Aouifi, A. Kukkar, et al. (2025), "LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: A statistical analysis approach", *Social Network Analysis and Mining*, **15(1)**, DOI: 10.1007/s13278-025-01484-1.
- [12] J. Yu (2021), "Academic performance prediction method of online education using random forest algorithm and artificial intelligence methods", *International Journal of Emerging Technologies in Learning*, **16(5)**, pp.45-57, DOI: 10.3991/ijet.v16i05.20297.
- [13] Y. Rimal, N. Sharma, A. Alsadoon (2024), "The accuracy of machine learning models relies on hyperparameter tuning: Student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms", *Multimedia Tools and Applications*, **83(30)**, pp.74349-74364, DOI: 10.1007/s11042-024-18426-2.
- [14] S.H. Paek, M. Kim (2025), "Characteristics explaining students' creative behaviors in South Korea using random forest", *Asia Pacific Education Review*, **26(4)**, pp.939-954, DOI: 10.1007/s12564-024-09930-z.
- [15] S.A. Sulak, N. Koklu (2024), "Predicting student dropout using machine learning algorithms", *Intelligent Methods in Engineering Sciences*, **3(3)**, pp.91-98, DOI: 10.58190/imiens.2024.103.
- [16] C.R. Çırak, H. Akıllı, Y. Ekinci (2024), "Development of an early warning system for higher education institutions by predicting first-year student academic performance", *Higher Education Quarterly*, **78(4)**, DOI: 10.1111/hequ.12539.
- [17] G. Liu, H. Zhuang (2022), "Evaluation model of multimedia-aided teaching effect of physical education course based on random forest algorithm", *Journal of Intelligent Systems*, **31(1)**, pp.555-567, DOI: 10.1515/jisys-2022-0041.
- [18] Y.S.N. Rao, C.J. Chen (2024), "Bibliometric insights into data mining in education research: A decade in review", *Contemporary Educational Technology*, **16(2)**, DOI: 10.30935/cedtech/14333.