

การวิเคราะห์และการคาดการณ์จำนวนนักศึกษาปริญญาตรีที่ออกจากระบบโดยใช้การเรียนรู้ของเครื่อง

Analysis and prediction of undergraduate student dropouts using machine learning

จักรพันธ์ สาตมณี¹, เอกชัย เนาวนิช² และ ธนพร ปฏิกรณ์³

Jakkapan Satmune¹, Ekachai Naowanich² and Thanaporn Patikorn³

Received: 22 December 2024 ; Revised: 7 February 2025 ; Accepted: 9 April 2025

บทคัดย่อ

การศึกษานี้จะนำเสนอการวิเคราะห์และเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องแปดแบบ ได้แก่ Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, AdaBoost, Hist Gradient Boosting, XGBoost และ Light Gradient Boosting ในการระบุนักศึกษาปริญญาตรีที่มีความเสี่ยงในประเทศไทย การศึกษานี้มุ่งเน้นไปที่การวิเคราะห์ข้อมูลทางวิชาการจากนักศึกษาในมหาวิทยาลัยเพื่อทำความเข้าใจปัจจัยที่มีอิทธิพลต่อความสำเร็จหรือความล้มเหลวทางวิชาการ และการพัฒนาโมเดลการทำนายเพื่อช่วยให้อาจารย์และผู้บริหารระบุแนวโน้มการเรียนรู้และนำกลยุทธ์การสอนที่มีประสิทธิภาพไปใช้ ชุดข้อมูลประกอบด้วยบันทึกจากนักศึกษาปริญญาตรี 2,104 คนในสาขาวิทยาศาสตร์และเทคโนโลยีตั้งแต่ปี 2020 ถึง 2024 ในประเทศไทย โดยมีคุณลักษณะข้อมูล 36 รายการ ในระหว่างการเตรียมข้อมูล ค่าที่ขาดหายไปจะถูกใส่โดยใช้เมธอด k-nearest neighbors (k=5) และความไม่สมดุลของคลาสได้รับการแก้ไขโดยใช้เทคนิคการสุ่มตัวอย่างชนกลุ่มน้อยแบบสังเคราะห์ (SMOTE) การกำจัดคุณลักษณะแบบเรียกซ้ำ (RFE) ถูกใช้เพื่อระบุคุณลักษณะที่สำคัญ ประสิทธิภาพของแบบจำลองได้รับการประเมินโดยใช้การตรวจสอบแบบไขว้โดยใช้ตัวชี้วัดหลักสี่ประการ ได้แก่ ความถูกต้อง ความแม่นยำ การเรียกคืน และคะแนน F1 ผลลัพธ์บ่งชี้ว่าแบบจำลอง Random Forest มีประสิทธิภาพดีที่สุด โดยบรรลุความถูกต้อง 84.32% ความแม่นยำ 0.85 การเรียกคืน 0.84 และคะแนน F1 0.85 ผลลัพธ์บ่งชี้ถึงศักยภาพที่สำคัญของแบบจำลองนี้ในระบบการศึกษา โดยเฉพาะอย่างยิ่งในการจัดเตรียมการแทรกแซงที่ตรงเป้าหมายสำหรับนักเรียนที่มีความเสี่ยงเพื่อลดอัตราการลาออกและเพิ่มความสำเร็จทางวิชาการ การวิจัยในอนาคตควรสำรวจปัจจัยที่มีอิทธิพลเพิ่มเติมและตรวจสอบแบบจำลองการเรียนรู้ของเครื่องจักรอื่น ๆ ที่อาจให้ประสิทธิภาพที่ดีขึ้น รวมทั้งเปรียบเทียบผลลัพธ์ในกลุ่มนักเรียนที่หลากหลายเพื่อให้มั่นใจถึงความน่าเชื่อถือและความสามารถในการนำไปใช้ที่กว้างขึ้น

คำสำคัญ: การวิเคราะห์และการคาดการณ์, การเรียนรู้ของเครื่องจักร

Abstract

This study presents an analysis and comparison of the performance of eight machine learning models, namely Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, AdaBoost, Hist Gradient Boosting, XGBoost, and Light Gradient Boosting, in identifying at-risk undergraduate students in Thailand. This study focuses on analyzing academic data from university students to understand the factors influencing academic success or failure, and developing predictive models to help instructors and administrators identify learning trends and implement effective

¹ อาจารย์, สาขาวิชาเทคโนโลยีดิจิทัล, คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ สุพรรณบุรี

² อาจารย์, สาขาวิชาเทคโนโลยีดิจิทัล, คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ นนทบุรี

³ อาจารย์, สาขาวิชาเทคโนโลยีดิจิทัล, คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ นนทบุรี

¹ Department of Digital Technology, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Suphan Buri

² Department of Digital Technology, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi

³ Department of Digital Technology, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi

* Corresponding author, e-mail: jakkapan.s@rmutsb.ac.th

teaching strategies. The dataset consists of records from 2,104 undergraduate students in science and technology from 2020 to 2024 in Thailand, with 36 data features. During data preparation, missing values were imputed using the k-nearest neighbors ($k=5$) method, and class imbalance was corrected using the synthetic minority sampling technique (SMOTE). Recursive feature elimination (RFE) was used to identify important features. The model performance was evaluated using cross-validation using four main metrics: accuracy, precision, recall, and F1 score. The results indicate that the Random Forest model performs the best. It achieved an accuracy of 84.32%, a precision of 0.85, a recall of 0.84, and an F1 score of 0.85. The findings suggest the significant potential of this model in the educational system, particularly in providing targeted interventions for at-risk students to reduce dropout rates and increase academic success. Future research should explore additional influencing factors and investigate other machine learning models that may provide improved performance, as well as compare the results across diverse student populations to ensure reliability and broader applicability.

Keywords: Analysis and prediction, machine learning

Introduction

In today's era, where information technology and artificial intelligence (AI) play pivotal roles in every aspect of human life, education is no exception (AI Thailand, 2022). The education system is rapidly adapting to technological advancements, particularly at the undergraduate level, where students are expected to develop specialized knowledge and skills within their chosen fields (Rastrollo-Guerrero *et al.*, 2020). Beyond the focus on teaching and learning, analyzing and predicting student performance has gained considerable attention, as it can be applied to improve teaching methods and provide preventative support for students who may struggle academically (Cahaya Hidayati *et al.*, 2024; Hashem *et al.*, 2024; Weist *et al.*, 2024). In the past, analyses of student performance largely relied on traditional statistical methods, primarily focused on retrospective data analysis. While informative, these methods fall short of accurately forecasting future trends or supporting proactive decision-making. Machine learning (ML), a technology capable of learning from data and predicting future outcomes (Baabdullah, 2024; Ibrahim Adedeji Adeniran *et al.*, 2024), presents new opportunities for more efficient analysis and prediction of student performance.

Machine learning can process complex and diverse data, including academic records, learning behaviors, participation in activities, and even personal information, to generate precise predictions. Applying this technology in educational systems enhances not only

performance forecasting but also improves teaching methods, tailors learning recommendations to individual students, and fully supports the development of each student's potential (Al-Sulami *et al.*, 2023; Nayak *et al.*, 2023; Venkatesan *et al.*, 2024). Additionally, performance analysis and prediction help educators and administrators gain insights into each cohort and major, as well as identify factors influencing student outcomes, such as social, economic, and university environmental factors (Hofmann *et al.*, 2024; Shoaib *et al.*, 2024). These forecasts enable strategic planning to improve educational effectiveness and facilitate timely interventions for at-risk students.

In the research context, machine learning holds the potential to drive significant change in the education sector. This study, focused on "Analysis and Prediction of Undergraduate Student Performance Using Machine Learning," (Altabrawee *et al.*, 2019) emphasizes applying ML techniques to analyze university student performance data. It explores factors affecting academic success or failure and develops prediction models to help instructors and administrators recognize learning trends and optimize teaching approaches (Alshamaila *et al.*, 2024; Hussain & Khan, 2023; Rabbi *et al.*, 2024). Furthermore, predicting student performance aids in designing personalized learning pathways, as students with different academic results may require unique learning strategies (Asselman *et al.*, 2023; Ibrahim *et al.*, 2023; Pamir *et al.*, 2022; Singh *et al.*, 2020). With machine learning, systems can customize learning experiences to meet each student's

needs and provide ongoing tracking and assessment of their progress.

This research aims to present a model for analyzing and predicting student performance that can be practically applied within the context of university education (Altabrawee *et al.*, 2019; Gašević *et al.*, 2016; Rastrollo-Guerrero *et al.*, 2020; Shoaib *et al.*, 2024). It is expected to benefit students, educators, and academic administrators by enhancing educational effectiveness and contributing to long-term improvements in student achievement and support strategies.

Materials and methods

1. Related work

Research in Educational Data Mining (EDM) has garnered continuous interest and development over the years, covering a wide range of fields such as Data Mining, Machine Learning, Deep Learning, and Big Data management. The primary objective of these research approaches is to develop predictive models capable of forecasting academic outcomes, predicting grades, ranking institutions, and recommending suitable institutions for students, using efficient data analysis tools and techniques. Significant studies in this field provide deeper insights and demonstrate practical methods that can be effectively implemented in educational systems.

Research on predicting student academic performance using Machine Learning (ML) is a key area within Educational Data Mining (EDM). Evaluating student performance is inherently complex, influenced by various factors such as academic history and personal characteristics (Hussain & Khan, 2023). This study employs a supervised learning approach, utilizing a dataset provided by the Board of Intermediate and Secondary Education (BISE) of Khyber Pakhtunkhwa, which includes data from various regions. The research focuses on using regression models and Decision Tree (DT) classifiers to predict students' scores and grades, with the data preprocessed for cleaning and organization prior to model training. The results demonstrate that ML technology is effective in accurately predicting student performance, offering valuable insights for enhancing educational quality and management within academic institutions.

Predicting student performance is a critical topic within the context of learning environments, such as schools and universities (Rastrollo-Guerrero *et al.*, 2020). This prediction can contribute to the development of effective mechanisms for improving academic outcomes and reducing dropout rates. Researchers benefit from leveraging technologies capable of managing large datasets derived from students' routine activities, which are collected through technology-enhanced software tools. In-depth data analysis reveals insights into the relationships between students and their academic tasks. This information can be used to create algorithms that accurately predict academic performance. In this study, the researchers reviewed over 70 studies utilizing advanced techniques, including machine learning, collaborative filtering, recommendation systems, and neural networks all within the field of artificial intelligence to enhance the effectiveness of predicting student performance.

This study primarily aims to identify students who need additional support to enhance their academic performance by utilizing four machine learning techniques: Artificial Neural Networks (ANN), Naïve Bayes, Decision Trees, and Logistic Regression (Altabrawee *et al.*, 2019). These models are applied to create a classifier for predicting student performance in computer science courses at Al-Muthanna University (MU). The research highlights the impact of using the internet as a learning resource and the time students spend on social networks on their academic performance, based on data from surveys and student gradebooks. Experimental results show that the ANN model (Fully Connected Feed Forward Multilayer ANN) performs best in predicting academic outcomes, with an efficiency score of 0.807 and an accuracy of 77.04%. Additionally, the Decision Tree model identifies five key factors that significantly influence students' academic performance.

Performance Factors Analysis (PFA) is a key approach in Knowledge Tracing (KT) and has been utilized to create adaptive educational hypermedia systems. PFA demonstrates high predictive accuracy compared to other KT methods (Asselman *et al.*, 2023). However, most developments have focused on instructional

aspects and have not incorporated machine learning techniques that could effectively enhance prediction accuracy. This study introduces a new approach in PFA by integrating machine learning models, such as Random Forest, AdaBoost, and XGBoost, to improve prediction accuracy of student performance. Experimental evaluations on three datasets showed that XGBoost outperformed other models, significantly enhancing performance predictions compared to traditional PFA algorithms. In conclusion, the new PFA, developed with machine learning techniques, effectively improves predictions of student academic performance, especially through the use of the XGBoost model, which demonstrates higher accuracy in analyzing performance factors among students.

Predicting applicants likely to enroll is a crucial aspect of college and university admissions management (Singh *et al.*, 2020). This research aims to assist educational institutions in optimizing the admissions process and reducing related costs by leveraging machine learning technology to analyze key parameters, such as high school grades, tutor selection, extracurricular activities, and reasons for applying. In this study, comparative analysis techniques were employed using various algorithms, including logistic regression, neural networks, and Light Gradient Boosting Machine (LightGBM), to predict applicants with a high likelihood of enrollment. Experimental results demonstrated that these algorithms achieved an accuracy of up to 98.5%. Applying these technologies not only enhances admission efficiency but also enables institutions to focus on students with potential and to keep pace with the ever-changing educational market demands.

This study proposes an approach for detecting electricity theft in smart grids by utilizing data from smart meters and focusing on machine learning (ML) and deep learning strategies. (Pamir *et al.*, 2022) While these technologies hold promise for enhancing detection accuracy, existing frameworks face limitations in performance and the handling of imbalanced data. To address these issues, the article introduces a Multi-Layered Boosted Classifier Stacking Model (MLBCSM), which employs the Adaptive Synthetic Sampling (ADASYN)

technique for data preprocessing, balancing, and classification. Simulation results indicate that MLBCSM outperforms other models in electricity theft detection by effectively combining the strengths of techniques such as AdaBoost, XGBoost, and CatBoost.

User authentication is crucial for securing digital services and preventing unauthorized access (Ibrahim *et al.*, 2023). Many organizations have shifted from single-factor authentication (SFA) to multi-factor authentication (MFA), which uses two or more methods of identity verification to enhance security. One innovative method is keystroke dynamics, a behavioral biometric technique that analyzes users' typing patterns to confirm their identity. This approach is cost-effective, as it requires no special hardware and is easy to implement, demanding minimal additional effort from users. This research introduces a new approach using quantile transformation to achieve uniform data distribution, thereby reducing the impact of outliers and extreme values. The primary classification method employed is Histogram Gradient Boosting, tested using the Carnegie Mellon University (CMU) keystroke dataset. Results indicate that the proposed method achieves an impressive average accuracy of 97.96% and an equal error rate (EER) of 0.014, outperforming previous machine learning and deep learning techniques in terms of performance.

2. Data and methodology

Predicting student academic performance at a national level is a crucial task, especially on online platforms. Our goal is to develop a predictive model that considers demographic and academic characteristics of undergraduate students across Thailand, as illustrated in Figure 1. We hope that educational institutions in various countries will be able to use these predictive insights to implement appropriate interventions. Although a student's environment and past academic performance impact their likelihood of dropping out, we applied machine learning algorithms after collecting, preprocessing, selecting, and transforming the data set. Consequently, the statistical analysis conducted is supported by hypothesis testing on performance indicators that we developed and validated.

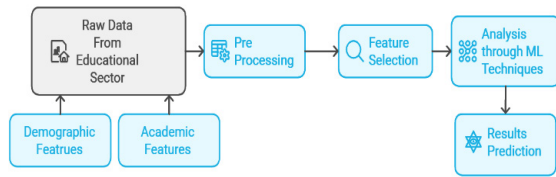


Figure 1 Conceptual model

2.1 Data

The data used for processing is presented in Table 1, which contains information on undergraduate students studying science and technology in Thailand from 2020 to 2024. This dataset includes 2,104 students and consists of 36 columns with various types of information. The original dataset contains numerous NaN values, anomalies, and unscaled data, making it unsuitable for model training with missing values, outliers, and unscaled data (Pamir *et al.*, 2022). Therefore, pre-processing is necessary to handle missing data by using the k-nearest neighbors' method. The missing values will be calculated based on the average of the nearest neighbors, with $k=5$, using Python, where the selection in KNN was determined based on empirical testing and reference to previous studies indicating that a smaller value of fit of KNN often provides better generalization. Smaller values may lead to overfitting, while larger values may smooth out important patterns in the data. Previous research suggests that smaller values often effectively balance bias and variance in classification tasks involving educational datasets (Peterson, 2009). This calculation for missing values is based on the mean or weighted average of the nearest k neighbors, which can be summarized by the following fundamental equation

A simple calculation formula of KNN for filling in missing values (without weights).

$$\hat{x}_i = \frac{1}{k} \sum_{j \in N(i)} x_j$$

KNN weighted average calculation formula.

$$\hat{x}_i = \frac{\sum_{j \in N(i)} \omega_j x_j}{\sum_{j \in N(i)} \omega_j}$$

Where : $\hat{x}_i$ It the value to be filled.

x_j It is the value from the neighbors.

ω_j It is the weight value.

$N(i)$ is the set of k nearest neighbors of the data point i .

Table 1 The data attributes

No.	Attribute	Description
1	GENDER_ID	Gender (int)
2	PROVINCE_ID	Province (int)
3	LEV_ID	value CLASS (int)
4	CURR_ID	Program (int)
5	FAC_ID	Course Code (int)
6	REF_PROGRAM_ID	Program Code (float)
7	CURR_TYPE_ID	Status current (int)
8	STUDY_TYPE_ID	Program Type (int)
9	STUDY_TIME_ID	Program Format (int)
10	CURR_INTER_ID	value current (float)
11	REGULAR_YEAR	Regular Year Level (int)
12	CURR_REG_ID	Academic Status Code (int)
13	CLASS	Grade Level (float)
14	GRAD_STATUS_ID	Pre-Entry Level (float)
15	REF_STD_STATUS_ID	Status Code (float)
16	ACC_GPA	GPA (float)
17	DEFORM_ID	Physical disability status (int)
18	SON_NUM	Float Child
19	SON_STDNUM	Float Siblings
20	FAT_STATUS_ID	Father's Life Status (float)
21	FAT_REVENUE_ID	Father's Income (float)
22	FAT_OCCUP_ID	Father's Marital Status (float)
23	MOT_STATUS_ID	Mother's Life Status (float)
24	MOT_REVENUE_ID	Mother's Income (float)
25	MOT_OCCUP_ID	Mother's Marital Status (float)
26	PAR_STATUS_ID	Parent's Life Status (float)
27	PAR_REVENUE_ID	Parent's Income (float)
28	PAR_OCCUP_ID	Parent's Marital Status (float)
29	RELIGION_ID	Religion (float)
30	Target	Status (int)
31	BLOOD	Blood Type (int)
32	HEIGHT	Height (float)
33	WEIGH	Weight (float)
34	RELIGION_NAME	Religion (int)
35	ADMISSION	Application Form (int)
36	CAMPUS	Area Center (int)

2.2 Data balancing

Initially, the dataset was analyzed to determine the class distribution between dropout and non-dropout students. The original dataset included 2,104 students, with 456 students (21.67%) classified as dropouts and 1,648 students (78.33%) classified as non-dropouts, clearly indicating class imbalance. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed (Rahayu *et al.*, 2024). SMOTE generates synthetic data points for the minority class (dropouts) by interpolating between existing dropout instances and their nearest neighbors, effectively balancing the representation between dropout and non-dropout cases and mitigating

bias toward the majority class during predictive modeling. Following data balancing, Recursive Feature Elimination (RFE) was conducted to identify and retain only the most relevant attributes for predicting student dropout. RFE iteratively selects impactful variables based on their predictive significance and correlation with the target variable (Barzani *et al.*, 2024). The variables with the highest significance for predicting student dropout status identified through RFE were ACC_GPA, REGULAR_YEAR, RELIGION_ID, GENDER_ID, and FAT_REVENUE_ID, highlighting clear associations with demographic and academic characteristics affecting student dropout risk. Consequently, these selected attributes were utilized

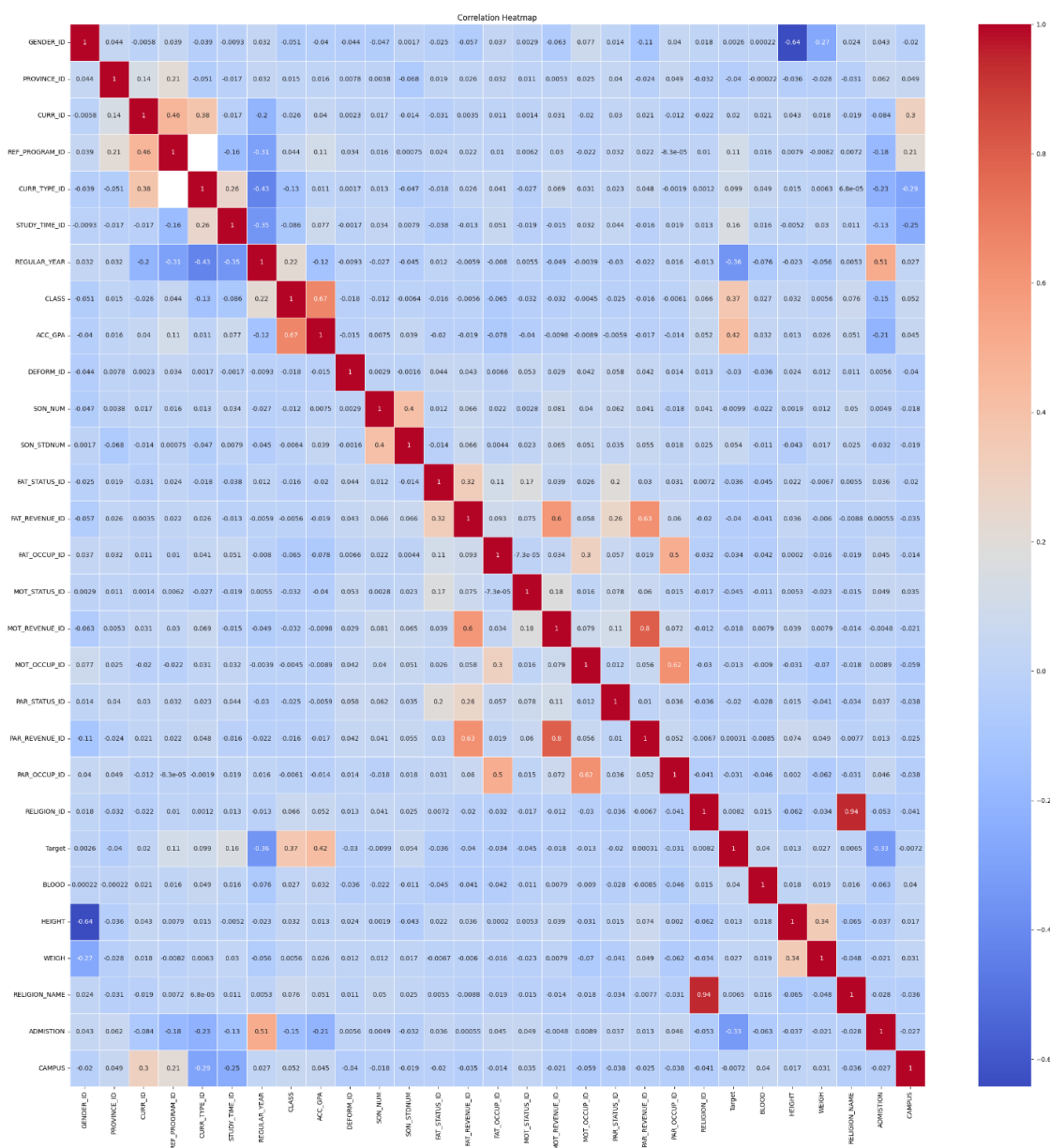


Figure 2 Heatmap of correlation matrix

to construct and refine the Machine Learning models used in this research.

The relationships between dataset attributes were further explored through correlation analysis, visualized in Figure 3 as a correlation matrix. This matrix presents correlation coefficients between pairs of attributes, utilizing color gradients to represent the strength and direction of relationships. Darker red indicates strong positive correlations, meaning variables increase together, while darker blue indicates strong negative correlations, suggesting an inverse relationship. Lighter shades of red and blue represent weaker or statistically insignificant relationships.

The correlation matrix in Figure 3 includes 29 attributes from the original set of 36. Several attributes were omitted from the correlation analysis due to considerations such as excessive missing data, minimal variability, or preliminary analyses indicating low relevance to dropout prediction. These omissions aimed to enhance clarity, interpretability, and overall effectiveness of the correlation visualization. Future research should explicitly document and justify attribute exclusions to maintain transparency and reproducibility.

Analysis of the correlation matrix identified four influential attributes closely associated with dropout status: RELIGION_ID, RELIGION_NAME, REGULAR_YEAR, and GENDER_ID. RELIGION_ID showed the highest correlation, suggesting religiosity might influence student retention. REGULAR_YEAR also demonstrated significant correlation, highlighting the importance of enrollment year in predicting dropout likelihood. Lastly, GENDER_ID exhibited meaningful correlation, emphasizing potential gender-based differences in student attrition. These findings underline the significance of demographic and enrollment factors in developing effective predictive models for student dropout.

3. Algorithm

The selection of an appropriate model for predicting student dropout depends on various factors such as interpretability, efficiency, and processing complexity. This study employs several machine learning models proven effective for classification tasks. Decision

Tree offers interpretability but is prone to overfitting, while Random Forest mitigates this issue by combining multiple decision trees. Logistic Regression is well-suited for binary classification with linear relationships, whereas K-Nearest Neighbors (KNN) leverages distance-based calculations, making it effective for small datasets. Ada-Boost enhances weak classifiers through boosting, Hist Gradient Boosting optimizes structured data learning using histogram-based techniques, XGBoost provides a highly efficient gradient boosting implementation, and LightGBM delivers faster performance, particularly for large datasets.

3.1 Decision tree

This model divides datasets into subsets based on various attributes to make decisions, making it easy to use and understand (Li *et al.*, 2024; Perez *et al.*, 2018). The tree structure shows a sequence of events that can be followed, providing insight into the primary factors that might lead students to drop out. However, it is prone to overfitting when data is highly complex. The Decision Tree model divides data into different nodes according to attributes, using decision criteria such as Gini Impurity or Entropy to select the best split path. The formula used to calculate node purity is as follows.

$$\text{Gini} = 1 - \sum_{i=1}^n p_i^2$$

$$\text{Entropy} = - \sum_{i=1}^n p_i \log(p_i)$$

Where : p_i is the proportion of data that is in class j .
 n is the total number of classes.

3.2 Random Forest

It combines multiple decision trees to create a single model, reducing the risk of overfitting by randomly selecting features and training multiple trees. This approach is suitable for high-accuracy predictions and works well with complex data, such as predicting student dropouts, which may depend on various factors (Kumar *et al.*, 2024).

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

Where : $\hat{y}$ It is the result of the model.

$f_t(x)$ This is the prediction result from each tree.

3.3 Logistic regression

The statistical model used to predict the probability of outcomes in binary groups finds the relationship between predictor variables and outcomes through logistic regression (Ayyalasomayajula *et al.*, 2024; Perez *et al.*, 2018). This model is suitable for forecasting student dropout when the relationship between predictor variables and outcomes is linear, and the data is not complex.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_nX_n)}}$$

Where : β is the coefficient of the variable
 X is a predictive variable.

3.4 K-Nearest Neighbors (KNN)

KNN is a model that uses the distance between data points to predict outcomes by examining the k most similar neighboring samples (Halder *et al.*, 2024). It is suitable for relatively simple datasets with a small number of samples. However, this model may not perform well when there are many variables or a large amount of data.

$$\hat{y} = \frac{1}{k} \sum_{i \in N(x)} y_i$$

Where : $\hat{y}$ It is the predicted value.

k It is the number of adjacent data used in the calculation.

y_i It is the value of the target variable in the nearest neighbor data.

$N(x)$ is the set of nearest neighbors points that are closest to the data to be predicted

3.5 AdaBoost

It is an efficiency-enhancing model that improves accuracy by training multiple simple models and

combining their outputs to create a more robust model (Xing *et al.*, 2024). This model performs well on less complex data but can also handle imbalanced data, such as cases where the number of employees who resign is lower than those who do not.

$$F(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

Where : α_t is the weight of the base model at t .

$h_t(x)$ This is the basic model.

Hist gradient boosting

It is a Gradient Boosting technique that processes data quickly by using histograms for parameter calculation and optimization (Wijayanti *et al.*, 2024). This method is suitable for large and complex datasets. The use of Hist Gradient Boosting can achieve high accuracy while requiring less processing time compared to standard Gradient Boosting methods.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Where : γ_m It is learning rate.

$h_m(x)$ It is a newly created tree in round m .

3.7 XGBoost

It is an enhanced version of Gradient Boosting that operates quickly and efficiently, reducing overfitting through regularization and allowing for detailed learning adjustments (Yu *et al.*, 2024). This model is particularly suitable for predicting student attrition when the data is complex and abundant. It has gained significant popularity due to its accuracy and effectiveness in handling large datasets.

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

Where : η It is learning rate.

3.8 LightGBM

Another technique is Gradient Boosting, which focuses on high efficiency and faster processing than XGBoost by utilizing leaf-wise data splitting and handling large datasets (Omotehinwa *et al.*, 2024). It is

particularly well-suited for large datasets and diverse features, such as student data related to dropout rates. LightGBM can enhance the prediction process to be both rapid and accurate.

$$F_m(x) = F_m - 1(x) + \gamma_m h_m(x)$$

Where : γ_m is the appropriate learning rate.

$h_m(x)$ It is a newly created tree.

Results

In preparing data for building a machine learning model, we start by separating the features from the target. Here, X represents the data containing only the features, which is obtained by removing the 'Target' column, while y is the target data. Next, we split the data into a training set and a testing set using the train_test_split function, designating %80 of the data for training and %20 for testing.

For performance evaluation, metrics from the confusion matrix are used, including Accuracy, Precision, Recall, and F1 score, which are commonly used measures in classification problems. The confusion matrix summarizes the results of binary predictions by displaying four cases: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Arefin *et al.*, 2024; Omotehinwa *et al.*, 2024; Venkatesan *et al.*, 2024; Wijayanti *et al.*, 2024; Yu *et al.*, 2024). Accuracy is the proportion of total outcomes that the model predicts correctly. Precision measures the accuracy of the model in predicting the positive class. Recall reflects the model's ability to identify all instances of the positive class. Because Precision and Recall are often inversely related, the F1 Score serves as the harmonic mean of the two.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2 Cross-validation results for identifying at-risk students using early detection models.

Method	Acc	Precision	Recall	F1
Decision Tree	79.81	0.80	0.79	0.80
Random Forest	84.32	0.85	0.84	0.85
Logistic Regression	66.03	0.60	0.66	0.63
K-Nearest Neighbors	52.73	0.53	0.52	0.53
AdaBoost	60.81	0.70	0.60	0.65
Hist Gradient Boosting	83.14	0.83	0.83	0.83
XG boost	81.71	0.80	0.82	0.81
Light Gradient	82.42	0.81	0.82	0.82

The results of the performance testing of machine learning models in identifying at-risk students were assessed using cross-validation techniques. This research compared the performance of eight models: Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, AdaBoost, Hist Gradient Boosting, XGBoost, and LightGBM. Model performance was evaluated using four main metrics: Accuracy, Precision, Recall, and F1 Score.

The Random Forest model demonstrated the best outcomes, achieving an accuracy of 84.32%, a precision of 0.85, a recall of 0.84, and an F1 score of 0.85. Further analysis using Recursive Feature Elimination (RFE) identified critical factors influencing student dropout, including ACC_GPA, REGULAR_YEAR, RELIGION_ID, GENDER_ID, and FAT_REVENUE_ID, with GPA identified as the most significant predictor. These results highlight the effectiveness of the Random Forest model in accurately identifying at-risk students. Therefore, the Random Forest model can be effectively employed by educational institutions to proactively identify students at risk and implement targeted interventions to reduce dropout rates.

Discussion and conclusion

This research study examined and compared the performance of eight machine learning models, namely Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, AdaBoost, Hist Gradient Boosting, XGBoost, and LightGBM for identifying at-risk

undergraduate students using data from students in Thailand. Prior to model training, we addressed the imbalance between dropout and non-dropout students using the Synthetic Minority Oversampling Technique (SMOTE). SMOTE effectively balanced the dataset, reducing model bias toward the majority class and improving predictive accuracy.

Additionally, we employed Recursive Feature Elimination (RFE) as a feature selection method to identify the most critical attributes for predicting dropout status. RFE highlighted key attributes such as ACC_GPA, REGULAR_YEAR, RELIGION_ID, GENDER_ID, and FAT_REVENUE_ID, which significantly contributed to model performance. Incorporating these selected features enhanced the interpretability and accuracy of the predictive models.

The evaluation results, assessed using cross-validation techniques and four primary metrics (Accuracy, Precision, Recall, and F1 Score), indicated that the Random Forest model yielded the best outcomes, with an accuracy of 84.32%, precision of 0.85, recall of 0.84, and an F1 score of 0.85. These results underscore the Random Forest model's capability to accurately identify at-risk students.

These findings suggest that the Random Forest model has potential applications for developing educational systems, (Kumar *et al.*, 2024; Xu & Yin, 2021) particularly in providing appropriate support and interventions for at-risk students, which could help reduce dropout rates and promote greater academic success. However, future research should consider a more in-depth analysis of factors and variables related to student risk, as well as exploring other machine learning models that may perform better. In addition, comparative studies of the performance of models and advanced techniques such as hybrid models or adding other data using data from more diverse samples should be conducted to ensure their reliability and applicability in a wider context.

Acknowledgment

The researchers would like to express their gratitude to Rajamangala University of Technology Suvarnabhumi for providing the dataset.

References

- AI Thailand. (2022). *National AI action plan for Thailand's development (2022–2027)*. <https://ai.in.th/about-ai-thailand/>
- Al-Sulami, A., Al-Masre, M., & Al-Malki, N. (2023). Predicting at-risk students' performance based on LMS activity using deep learning. *International Journal of Advanced Computer Science and Applications*, 14(6), 567–574. <https://doi.org/10.14569/IJACSA.2023.0140661>
- Alshamaila, Y., Alsawalqah, H., Aljarah, I., Habib, M., Faris, H., Alshraideh, M., & Salih, B. A. (2024). An automatic prediction of students' performance to support the university education system: A deep learning approach. *Multimedia Tools and Applications*, 83(15), 46369–46396. <https://doi.org/10.1007/s11042-024-18262-4>
- Altabrawee, H., Ali, A. B. J., & Ajmi, Q. (2019). Predicting students' performance using machine learning techniques. *Journal of Physics: Conference Series*, 1234(1), 012010. <https://doi.org/10.1088/1742-6596/1234/1/012010>
- Arefin, S., Chowdhury, M., Parvez, R., Ahmed, T., Abrar, A. F. M. S., & Sumaiya, F. (2024). Understanding APT detection using machine learning algorithms: Is superior accuracy a thing? *2024 IEEE International Conference on Electro Information Technology (EIT)*, 532–537. <https://doi.org/10.1109/EIT60633.2024.10609886>
- Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31(6), 3360–3379. <https://doi.org/10.1080/10494820.2021.1928235>
- Ayyalasomayajula, M. M. T., Agarwal, A., & Khan, S. (2024). Reddit social media text analysis for depression prediction: Using logistic regression with enhanced term frequency-inverse document frequency features. *International Journal of Electrical and Computer Engineering*, 14(5), 5998–6005. <https://doi.org/10.1591/ijece.v14i5.pp5998-6005>

- Baabdullah, A. M. (2024). The precursors of AI adoption in business: Towards an efficient decision-making and functional performance. *International Journal of Information Management*, 75, Article 102745. <https://doi.org/10.1016/j.ijinfomgt.2023.102745>
- Barzani, A. R., Pahlavani, P., Ghorbanzadeh, O., Gholamnia, K., & Ghamisi, P. (2024). Evaluating the impact of recursive feature elimination on machine learning models for predicting forest fire-prone zones. *Fire*, 7(12), Article 440. <https://doi.org/10.3390/fire7120440>
- Cahaya Hidayati, C., Phyu Thet, E. W., & Nouanthong, P. (2024). Evaluation: Media booklet about education on prevention of flour albus for teenage girl. *Journal Evaluation in Education (JEE)*, 5(2), 50–60. <https://doi.org/10.37251/jee.v5i2.928>
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education*, 28, 68–84. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1), 1–55. <https://doi.org/10.1186/s40537-024-00973-y>
- Hashem, R., Ali, N., Zein, F. E., Fidalgo, P., & Khurma, O. A. (2024). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research and Practice in Technology Enhanced Learning*, 19, Article 19023. <https://doi.org/10.58459/rptel.2024.19023>
- Hofmann, P., Lämmermann, L., & Urbach, N. (2024). Managing artificial intelligence applications in healthcare: Promoting information processing among stakeholders. *International Journal of Information Management*, 75, Article 102728. <https://doi.org/10.1016/j.ijinfomgt.2023.102728>
- Hussain, S., & Khan, M. Q. (2023). Student-Performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*, 10(3), 637–655. <https://doi.org/10.1007/s40745-021-00341-0>
- Ibrahim Adedeji Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., & Abhulimen, A. O. (2024). Integrating business intelligence and predictive analytics in banking: A framework for optimizing financial decision-making. *Finance & Accounting Research Journal*, 6(8), 1517–1530. <https://doi.org/10.51594/farj.v6i8.1505>
- Ibrahim, M., Abdelraouf, H., Amin, K. M., & Semary, N. (2023). Keystroke dynamics based user authentication using histogram gradient boosting. *International Journal of Computers and Information (IJCI)*, 9(2). <https://ijci.journals.ekb.eg/>
- Kumar, P., Senthil Pandi, S., Kumaragurubaran, T., & Rahul Chiranjeevi, V. (2024). Human activity recognitions in handheld devices using random forest algorithm. *2024 International Conference on Automation and Computation (AUTOCOM)*, 159–163. <https://doi.org/10.1109/AUTOCOM60220.2024.10486087>
- Li, H., Song, J., Xue, M., Zhang, H., & Song, M. (2024). A survey of neural trees: Co-evolving neural networks and decision trees. *IEEE Transactions on Neural Networks and Learning Systems*. Advance online publication. <https://doi.org/10.1109/TNNLS.2024.3446891>
- Nayak, P., Vaheed, S., Gupta, S., & Mohan, N. (2023). Predicting students' academic performance by mining the educational data through machine learning-based classification model. *Education and Information Technologies*, 28(11), 14611–14637. <https://doi.org/10.1007/s10639-023-11706-8>
- Omotehinwa, T. O., Oyewola, D. O., & Mounq, E. G. (2024). Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. *Informatics and Health*, 1(2), 70–81. <https://doi.org/10.1016/j.infoh.2024.06.001>
- Pamir, Javaid, N., Akbar, M., Aldegheishem, A., Alrajeh, N., & Mohammed, E. A. (2022). Employing a machine learning boosting classifiers based stacking ensemble model for detecting non technical losses in smart grids. *IEEE Access*, 10, 121886–121899. <https://doi.org/10.1109/ACCESS.2022.3222883>

- Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. <https://doi.org/10.1109/COLCACI.2018.8484847>
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), Article 1883. <https://doi.org/10.4249/scholarpedia.1883>
- Rabbi, F., Ullah, N., Hossain, I., Ashour, O. M., & Ashour, O. (2024). *Predicting management skills of undergraduate students by leveraging different machine learning classifiers* [Preprint]. ResearchGate. <https://www.researchgate.net/publication/380880140>
- Rahayu, W., Jollyta, D., Hajjah, A., Nora Marlim, Y., & Desnelita, Y. (2024). Synthetic Minority Oversampling Technique (SMOTE) for boosting the accuracy of C4.5 algorithm model. *Journal of Artificial Intelligence and Engineering Applications*, 3(3). <https://ioinformatic.org/>
- Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3), Article 1042. <https://doi.org/10.3390/app10031042>
- Shoaib, M., Sayed, N., Singh, J., Shafi, J., Khan, S., & Ali, F. (2024). AI student success predictor: Enhancing personalized learning in campus management systems. *Computers in Human Behavior*, 158, Article 108301. <https://doi.org/10.1016/j.chb.2024.108301>
- Singh, S. P., Singh, P., & Mishra, A. (2020). Predicting potential applicants for any private college using LightGBM. *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*. <https://doi.org/10.1109/ICITIIT49094.2020.9071525>
- Venkatesan, R. G., Karmegam, D., & Mappillairaju, B. (2024). Exploring statistical approaches for predicting student dropout in education: A systematic review and meta-analysis. *Journal of Computational Social Science*, 7(1), 171–196. <https://doi.org/10.1007/s42001-023-00231-w>
- Weist, M. D., Garbacz, A., Schultz, B., Bradshaw, C. P., & Lane, K. L. (2024). Revisiting the percentage of K-12 students in need of preventive interventions in schools in a “peri-COVID” era: Implications for the implementation of tiered programming. *Prevention Science*, 25(3), 481–487. <https://doi.org/10.1007/s11121-023-01618-x>
- Wijayanti, E. B., Rosal, D., Setiadi, I. M., & Setyoko, B. H. (2024). Dataset analysis and feature characteristics to predict rice production based on eXtreme Gradient Boosting. *Journal of Computing Theories and Applications*, 1(3), 299–310. <https://doi.org/10.62411/jcta.10057>
- Xing, H. J., Liu, W. T., & Wang, X. Z. (2024). Bounded exponential loss function based AdaBoost ensemble of OCSVMs. *Pattern Recognition*, 148, Article 110191. <https://doi.org/10.1016/j.patcog.2023.110191>
- Xu, Q., & Yin, J. (2021). Application of Random Forest algorithm in physical education. *Scientific Programming*, 2021, Article 1996904. <https://doi.org/10.1155/2021/1996904>
- Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S., & Meng, S. (2024). *Advanced user credit risk prediction model using LightGBM, XGBoost and Tabnet with SMOTEENN* [Preprint]. arXiv. <https://arxiv.org/abs/2408.03497v1>