Predictive Analysis of Academic Achievement in Information Studies: A Comparative Study Using Educational Data Mining Techniques

Knitchepon Chotchantarakun*
Department of Information Studies, Faculty of Humanities and Social Sciences,
Burapha University, Chonburi, Thailand
E-mail: knitchepon@go.buu.ac.th*

Received: March 19, 2025 / Revised: May 29, 2025 / Accepted: June 5, 2025

***Abstract*—Predicting students' academic achievement in the initial stages is beneficial for designing effective training programs to enhance success rates. Extracting knowledge from student data is a fundamental aspect of Educational Data Mining (EDM). This study aims to analyze the predictive factors influencing the outcomes of graduates from the Information Studies program. The results not only contribute to improving student performance but also aid in constructing a better curriculum. A dataset is utilized within five classification models to categorize students into four target classes. The datasets are grouped into three types: Demographic information, course grades, and early-stage GPA. This study addresses the issue of the imbalanced dataset by applying the Synthetic Minority Over-sampling Technique (SMOTE). The findings indicate that early-stage GPA (90.5%) is the most significant predictor, particularly when applying the Naive Bayes classifier on a balanced dataset. In contrast, demographic information (58.0%) and core course grades (87.5%) show lower predictive influence. The findings support learning strategies and enhancing curriculum design to improve final academic outcomes.**

***Index Terms*—Classification, Educational Data Mining, Feature Selection, Imbalanced Dataset, Machine Learning**

I. Introduction

With the transition from paper-based documentation to digital formats, educational institutions generate vast volumes of electronic data. Transforming this extensive data into meaningful knowledge is essential in decision-making, improving learning quality, and providing information for institutional planning to maximize efficiency. The extraction of such knowledge has been facilitated by advancements in computer technology, particularly in Artificial Intelligence (AI), Data Mining (DM), and Machine Learning (ML). These progressions enable the application of predictive modeling, clustering, and association rules mining to identify patterns and correlations within educational data. These techniques contribute to the refinement of academic curricula, enhancement of teaching methodologies, and improvement of student learning outcomes.

The DM process encompasses a wide range of techniques and algorithms that are applied in various domains, including medicine, marketing, industry, finance, and education. The specific application of data mining to educational data is referred to as Educational Data Mining (EDM) [1]. Baker and Yacef identified five primary approaches within EDM: Prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. Recent works usually integrate a combination of prediction, clustering, and data distillation for human judgment.

EDM has emerged as a significant area of research, driven by advancements in educational database management systems. It emphasizes on developing specialized and quantitative methodologies to analyze large volumes of data collected from institutions across various educational levels. By applying DM techniques to student datasets, educators can gain deeper insights into student behavior and performance. While academic achievement is traditionally assessed through metrics such as course grades and Grade Point Average (GPA), demographic factors also play a crucial role in shaping educational outcomes. Consequently, the application of DM to academic and demographic data enhances the program design at the department, faculty, and institution levels.

Student academic success is a key indicator of the teaching and learning quality of the institutions. Early classification of students during the initial stages of their study is an effective approach to minimizing the possibility of dropout. Additionally, this strategy contributes to improved academic performance by facilitating optimal resource allocation within institutions, ensuring that resources are utilized

effectively. Furthermore, it supports the development of potential students, enabling them to gain higher academic achievements.

Although data obtained from the university's Student Information System (SIS) is increasingly utilized in higher education, many institutions still face challenges in developing effective predictive models. Normally, static indicators such as course grades or test scores are used to estimate the student's learning outcome. However, they cannot reflect the complexity and dynamics of learning behavior influencing academic success. As a result, opportunities for the timely identification and support of low-performance students are often missed. There is a critical need for predictive models that are both data-driven and interpretable, particularly those that leverage early academic indicators such as demographic information, course grades, and GPA. These models must provide prominent levels of predictive accuracy and be meaningful enough to enable educators to implement early support and enhance instructional strategies.

This study investigates the application of machine learning techniques to identify the key factors influencing students' final academic achievement during the initial stages of their educational journey. The prediction is conducted using five classification models within the EDM framework. Experiments are performed using undergraduate students' data from the Information Studies (IS) program, combined with the Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset. This research contributes to enhancing students' final academic outcomes and assisting with the curriculum design.

The paper is organized as follows: Section II reviews relevant literature. Section III outlines the research methodology employed in this study. Section IV presents the experimental results along with explanatory analysis. Section V offers a detailed discussion of the findings. Finally, Section VI concludes the paper with a summary of key findings and significant issues contributing to the learning strategies.

II. Literature Review

Early prediction of student outcomes enables educators and administrators to make informed, timely decisions to enhance course effectiveness. It facilitates the development of specialized training programs aimed at increasing student success rates. Advancements in EDM have shown the application of DM techniques across various educational domains [2]. Research objectives in this field can be defined at multiple levels, including degree, academic year, course, and examination levels. Studies in this area commonly employ classification techniques, such as predicting student outcomes as Pass or Fail, as well as regression techniques, such as estimating the Cumulative Grade Point Average (CGPA). Typically, CGPA prediction utilizes students' Grade Point Averages (GPA) from their first two years to forecast their final CGPA at graduation.

EDM plays a crucial role in uncovering patterns and insights related to educational phenomena and learning processes [5], as well as in understanding students' academic performance. EDM has been widely applied to predict academic outcomes across various domains, including academic performance [4], student retention [5], study success [6], academic satisfaction [7], and dropout rates [8].

Key factors influencing education, commonly explored in EDM, include prior academic achievement, student demographic characteristics, e-learning activities, psychological aspects, and the learning environment. A study [9] highlighted that 69% of research in this field utilizes pre-university academic performance and demographic characteristics of learners. Among the most frequently used predictors of academic achievement are student assessments and cumulative Grade Point Averages (GPAs). Both pre-university information and data collected during the study period, such as semester grades and GPAX, significantly influence the prediction of academic achievement [10]. These elements, derived from students' academic journeys, are critical in forecasting their overall academic success throughout their educational tenure.

In this study [11], DM techniques were employed to analyze the academic achievement of 210 undergraduate students. The authors developed a predictive model to estimate students' final academic performance and explored the relationship between their academic outcomes and progress during their course of study. The variables utilized in the analysis were exclusively related to scores or grades. A decision tree algorithm was applied to construct a classification model, which categorized the dataset based on four information criteria: Information Gain, Gini Index, Accuracy, and Gain Ratio. Additionally, the X-means algorithm, using Euclidean distance and the Bayesian Information Criterion (BIC), was employed to group students into categories reflecting high and low academic performance. The DM process was implemented using RapidMiner. The findings from this predictive model offer early intervention opportunities for students identified in the low-performing groups while also providing guidance and opportunities for those demonstrating strong academic performance.

In 2022, a study [12] conducted a comprehensive review and analysis of emerging literature on the application of Artificial Neural Network (ANN) to predict academic achievement among university students. The article highlighted that ANN techniques are frequently combined with other data mining

methods to identify patterns and assess academic performance. EDM research focuses on the university level, as most researchers are affiliated with universities and have easier access to student data. Furthermore, ANN often demonstrates higher accuracy in evaluating model performance compared to other algorithms. CGPA was identified as a commonly used predictive factor, while other factors were found to have a minimal impact on the model's performance.

Reference [13] conducted a study on the academic performance of 635 master's degree candidates across diverse faculties, such as Business Administration, Engineering, and Information Technology. The study employed six machine learning algorithms, utilizing CGPA as the principal predictor. Evaluation of model efficacy utilized Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics to gauge the disparity between predicted and actual GPA scores. Findings indicated that the Neural Network (NN) algorithm exhibited superior performance, attaining the lowest error rates and demonstrating heightened predictive precision compared to other models.

A comprehensive survey and synthesis of 402 articles related to EDM and Learning Analytics (LA) were examined in a review study [14]. The study highlighted that EDM and LA techniques are effective in addressing various learning-related challenges. The application of DM techniques across these studies included Classification (26.23%), Clustering (21.25%), Image Mining (15%), Statistical Analysis (14.25%), Association Rule Mining (14%), Regression (10.25%), and Sequential Pattern Mining (6.5%), etc. These techniques contributed to the development of improved learning strategies for students. Additionally, the review introduced researchers to appropriate methodologies for conducting research in different educational contexts and provided essential tools and information to enhance university education systems. Reference [15] evaluated the effectiveness of ML techniques in predicting student performance using various ML algorithms. The research considered factors such as data quality, feature selection, and model complexity. The findings demonstrated that certain ML methods are particularly effective in forecasting student performance, thereby offering valuable insights for decision-making and academic planning.

Previous research has demonstrated that predicting student academic achievement can achieve high levels of accuracy, particularly when EDM utilizes classification techniques with a limited number of categories. For example, predicting binary outcomes such as pass/fail status or categorizing students based on satisfactory performance often results in even greater predictive accuracy. However, the effectiveness of these predictive models is influenced by several factors, including the choice of algorithms, the selection of variables, and the size of the dataset. In addition, demographic attributes such as age, gender, religion, place of residence, family background, employment status, and past GPA have been identified as significant contributing factors that enhance the accuracy of academic achievement predictions [16], [17].

Regarding the application of SMOTE, this research [18] demonstrated that applying SMOTE before splitting the dataset into training and testing sets enhanced the accuracy of the ANN model. Specifically, it yielded accuracy improvements ranging from 1.94% to 3.98% across multiple datasets, highlighting its capability to address class imbalance and improve the overall performance of classification models. This study [19] presented the improvement in the accuracy of ANN on imbalanced datasets using SMOTE by generating synthetic samples to balance the class distribution. However, this process may introduce noise into the dataset. The proposed method addresses this limitation by incorporating an Autoencoder, which helps filter out noise and enhances the overall classification performance.

Despite the growing impact of EDM in improving student outcomes across higher education, current research disproportionately focuses on STEM (Science, Technology, Engineering, and Mathematics) disciplines and general education courses. In contrast, specialized academic domains such as the IS program remain significantly underrepresented, which presents a critical gap in the literature. This program emphasizes interdisciplinary knowledge, including critical thinking, system analysis, information behavior, digital literacy, and information retrieval. These competencies are typically assessed through project work and collaborative assignments rather than focusing only on the numerical scores and exam-based evaluations commonly used in STEM. As a result, existing EDM models may not capture the detailed indicators of academic success relevant to information studies students.

Additionally, the demographic and academic profiles of students in this program may differ from those in traditional STEM fields due to diverse academic backgrounds. This research explores this gap by applying EDM methods to analyze academic performance within the IS program. The findings are expected to contribute to the growing body of EDM literature while generating actionable insights for improving teaching, learning, and student support in the field related to information science. Consequently, the use of predictive approaches on these types of datasets is still limited, indicating a clear research gap. This study aims to address this underexplored area by analyzing this specific population, which may yield novel insights to inform curriculum design and enhance student support within these academic domains.

III. Research Methodology

A. Educational Data Mining Process

The EDM process [20] comprises six key stages as shown in Fig. 1. The data collection phase involves gathering information from multiple sources, including pre-enrollment records, demographic details, students' learning environments, academic performance, and psychological attributes. These data are primarily obtained from the university's Student Information System (SIS), supplemented by student surveys. Once collected, the data are prepared for subsequent analysis.

During the initial preparation stage, the raw data are transformed into a structured format through a series of processes, including 1) Selection, 2) Cleaning, and 3) Derivation of new variables. This phase is particularly critical and often requires the most time to complete. Following this, statistical analysis serves as an initial exploratory step, offering an overview of the dataset and assisting researchers in understanding key characteristics before proceeding with data mining. This analysis typically includes descriptive statistics such as frequency, mode, median, mean, standard deviation, variance, range, and correlation.

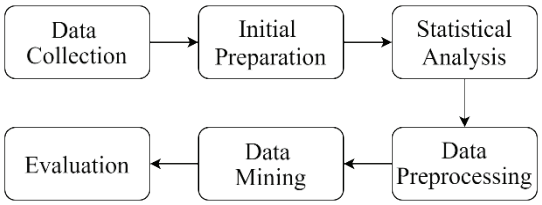

Fig.1. EDM Stages Framework

The data preprocessing stage consists of two essential components: Data transformation and feature selection. Feature selection focuses on identifying a subset of relevant variables while eliminating less significant or redundant ones. This process enhances the accuracy of predictive models and optimizes computational efficiency by reducing processing time. The DM process involves creating various types of models to evaluate and select the most appropriate model to summarize the research results. A confusion matrix is commonly used to determine the performance of these models and provides a detailed evaluation of their accuracy and effectiveness.

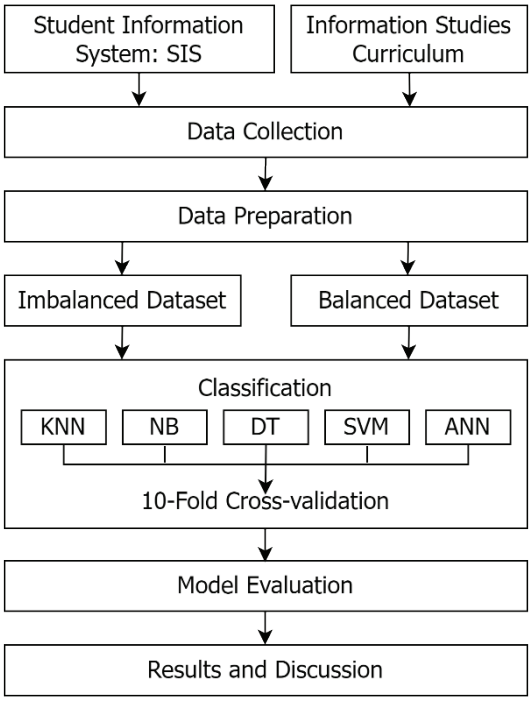

Fig.2. Research Framework

Fig.2 presents our research framework relevant to the EDM process. In the DM step, we create a model for predicting academic achievement using supervised ML to estimate the expected values of dependent variables based on the characteristics of the independent variables. Classification is the most popular method, and the most common classification algorithms are the Bayesian method, Neural Network, and Decision Tree. Apart from those three well-known algorithms, this study also includes the K-Nearest Neighbor and the Support Vector Machine for our model formation. As for data mining tools, WEKA is the most used tool for building predictive models since it has the functionality to answer all types of DM problems. RapidMiner is another widely adopted tool, ranking as the second most popular in the field [21].

B. Data Preparation

The volume of data is expanding rapidly, benefiting various academic disciplines, particularly in decision-making processes supported by computer technology and information management. The IS program is one of the key disciplines focused on information management through the application of modern technologies. It integrates the theories of information management and information technology. The program is designed to equip students with the knowledge and skills required to work in information-centric organizations, such as information agencies, libraries, and the broader information technology sector. This study utilizes data from the IS students of Burapha University, in conjunction with the IS curriculum.

Our study employs secondary data sourced from the SIS. The dataset is categorized into three types of variables: Demographic information, course grades, and early-stage GPA from the first six semesters. It includes data from 275 students who graduated from the Department of Information Studies between 2020 and 2023. Demographic information encompasses details such as students' home region, gender, number of siblings, parents' occupation and income, and high school GPA (SGPA). These variables are analyzed to identify patterns and factors influencing academic achievement among IS students.

This study investigates students' academic achievement over their four-year study period. The courses are categorized into three fundamental areas: General Education (GE), Information Science (IS), and Information Technology (IT). The IS and IT courses are further divided into core and elective courses, allowing a detailed analysis of students' academic performance. The dataset includes 12 GE courses, 18 core courses, and 10 elective courses, with each category split equally between IS and IT courses. The dataset is in its original form and prepared for the initial data preparation phase. Table I presents the variables used in the predictive models.

TABLE I
VARIABLES IN THE DATASETS

Type	Variables	Domain
Demographic	SGPA	{Excellent, Very Good, Good, Fair}
	Region	{North, South, Northeast, East, Central}
	Gender	{Male, Female}
	Sibling	{Yes, No}
	Father Income	{High, Medium, Low, None, Undefine}
	Father Occupation	{Government Officer, State Enterprise, Private Employee, Personal Business, Agriculture, Undefined}
	Mother Income	{High, Medium, Low, None, Undefined}
	Mother Occupation	{Government Officer, State Enterprise, Private Employee, Personal Business, Agriculture, Undefined}
Academic	GE Courses	{A, B+, B, C+, C, D+, D}
	IS Courses	{A, B+, B, C+, C, D+, D}
	IT Courses	{A, B+, B, C+, C, D+, D}
	GPA1-GPA6	{Excellent, Very Good, Good, Fair}
	AGPA2-AGPA6	{Excellent, Very Good, Good, Fair}

Academic achievement upon graduation is categorized into four classes based on the final GPA: Excellent, Very Good, Good, and Fair, as shown in Table II. After removing noise and missing values, the dataset includes information from 263 students out of an initial 275. This data cleaning improves the efficiency of the classification model and simplifies the computational process. Furthermore, to address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to generate additional samples, expanding the dataset from 263 to 400 instances.

TABLE II
TARGET CLASSES

GPA	Classes	No. of Instances (Original Dataset)	No. of Instances (Balanced Dataset)
3.50 – 4.00	Excellent	35	100
3.00 – 3.49	Very Good	102	100
2.50 – 2.99	Good	100	100
2.00 – 2.49	Fair	26	100
Total		**263**	**400**

An imbalanced dataset occurs when the instances of the target class are not evenly distributed. In this study, the dataset contains 35 samples for the "Excellent" class, 102 for "Very Good", 100 for "Good", and 26 for "Fair". The unequal distribution of instances in these classes results in an imbalanced dataset [22], [23], [24], which could potentially lead to inaccurate results. To address this issue, the SMOTE is applied to oversample the minority classes. As a result, the original dataset of 263 instances is expanded to 400 instances, with each target class containing 100 instances, thereby ensuring a balanced distribution across all classes.

Feature selection [25], [26] is a crucial pre-processing step in the DM process, aiming to identify and rank the importance of variables. Various ranking techniques are classified into filter-based, wrapper-based, and hybrid methods. In our study, we adopted a wrapper-based method, using the best classifier identified in our experiments to rank the variables based on their classification accuracy. The variable with the highest accuracy is considered the strongest correlation with the final academic performance.

C. Prediction Models

EDM models are broadly categorized into two types: predictive and descriptive models. Predictive models are employed to forecast outcomes using supervised learning techniques, while descriptive models aim to identify patterns that describe the underlying structure and relationships within unsupervised data. In this study, predictive models are applied to the datasets, utilizing five widely recognized ML algorithms: K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN). These algorithms are among the most used in EDM research [27], [28]. The selection of these

models for implementation is driven by the goal of identifying the most accurate predictive approach.

1) *K-Nearest Neighbor (KNN)* – KNN is a method of classifying data by comparing the data with sample data in the dataset. It refers to Instance-based Learning, which means the learned data is stored as an instance in the dataset. While classifying the new data, KNN will find *K* samples of data in the dataset that most resemble the new data (neighbors) and use them together to decide what class of new data should be. The principle behind this algorithm is to find the similar characteristics of the new data with the nearby dataset. Deciding the class of new data can be done using a method called "Majority Vote"; that is, new data will be the same class as the larger number of the neighbor data. Euclidean distance is a common measurement used to calculate the distance between data points. Classification using the KNN method can provide much accuracy depending on many factors, such as the completeness of the sample data used to represent the entire data or how much noise is in the data. Selecting a small value of K, such as $K = 1$ or 2, may cause misclassification because the closest data may be noise. Therefore, the value of *K* should be defined appropriately. This study has selected $K = 5$ for the experiments.

2) *Naive Bayes (NB)* – The NB method is a fundamental classification technique that utilizes probability theory based on Bayes' theorem. It classifies data into predefined groups by applying probabilistic principles. The NB algorithm assigns each data instance to the class with the highest posterior probability, ensuring optimal classification based on the given probabilistic framework.

3) *Decision Tree (DT)* – The DT method is a classification technique based on the concept of divide and conquer. Initially, the dataset is divided into smaller parts based on the values of the variables. The collection of decision nodes is connected by branches extending from the root node to the leaf node. Each node contains a condition that uses one of the data variables to decide on one child node. Decision-making starts at the root node and then moves on to the child nodes until reaching the leaf nodes, which are class nodes. The depth of the trees is related to how fast the model can make decisions. The selection of variables and conditions must be justified to obtain a tree that can classify the data as accurately as possible. This study uses the C4.5 algorithm to build the DT model.

4) *Support Vector Machine (SVM)* – SVM is a supervised learning algorithm designed for building classification models, particularly well-suited for datasets characterized by small sample sizes and a high dimensionality of features. The fundamental concept of SVM revolves around the creation of decision boundaries, referred to as hyperplanes, which partition the feature space to distinguish between different classes. The primary objective of SVM is to identify the optimal hyperplane by maximizing the margin, which is defined as the aggregate of the shortest distances from the hyperplane to the nearest data points of each class. This approach ensures enhanced generalization and robust classification performance. Regarding the experiment, the SVM model is configured with a regularization parameter C set to 1.0, balancing the trade-off between maximizing the margin and minimizing classification errors. A polynomial kernel is selected to capture complex and nonlinear relationships within the data. Normalization is applied to ensure that the features are on a consistent scale, which enhances the model's convergence and stability. The tolerance for the stopping criterion is 0.001, allowing the training process to terminate once improvements fall below this threshold.

5) *Artificial Neural Network (ANN)* – ANN represents a branch of AI whose structure and functionality resemble the neural networks of biological organisms. This technique is suitable for a non-linear fitting method. The ANN Model adjusts itself in response to the input associated with the learning rules. The feed-forward network, which consists of multiple layers of neurons, provides parameter calculation in several iterations to get the best configuration. The neurons' connections between layers are fully connected. Each neuron sends its computation results to every neuron in the next layer. Each link between neurons consists of a weight value that magnifies the value and passes it over its link by multiplying it by this weight. The result is transmitted to the next layer of neurons. Repeat this process from the input to the output layer.

In the experiment, the ANN model is initialized using default parameters included in the mining tool. The hidden layer is dynamically optimized by automatically adjusting the number of neurons based on the input data. A learning rate of 0.3 regulated the weight updates during training, while a momentum coefficient of 0.2 is incorporated to enhance convergence speed and minimize oscillations in the gradient descent process. The training process is conducted over 500 iterations to ensure adequate learning and model stability.

D. Models Evaluation

A common technique to evaluate the goodness of the DM algorithm is to apply the confusion matrix shown in Table III.

TABLE III
CONFUSION MATRIX

Observation	Positive	Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The class value of True Positive (TP) has the predicted class as YES and is YES, while the class value of False Negative (FN) has the predicted class as NO and is YES. Similarly, False Positive (FP) has the predicted class as YES and is NO, while True Negative (TN) has the predicted class as NO and is actually NO. These terms (TP, FN, FP, TN) represent frequency values and are used to construct a confusion matrix, which is a valuable tool for evaluating the performance of predictive models. The confusion matrix enables the calculation of various performance metrics, as shown in Table IV.

TABLE IV
MEASUREMENT

Performance Criteria	Formula
Accuracy	(TP+TN) / (TP+TN+FP+FN)
Precision	TP / (TP+FP)
Recall	TP / (TP+FN)
F-Measure	2 x (Precision x Recall) / (Precision + Recall)

Model evaluation is a critical step in assessing the effectiveness of a predictive model. To identify the best-performing model, it is necessary to compare multiple models using various evaluation metrics. Accuracy is the performance criterion that is relevant to all parameters. The higher the TP and TN rates, the more accurate the prediction is, which means better model performance. On the other hand, high FP and FN rates present a signal of incorrect prediction. Regarding precision and recall, the TP rate directly affects the performance. Accordingly, we focus on the TP rate and the classification accuracy to get the optimal prediction.

IV. RESULTS

In the experiment, the dataset is partitioned into two subsets: A training set and a testing set. Using the training set to build models by applying various algorithms, while the testing set is employed for model validation. Our EDM process incorporates the *K*-fold cross-validation technique, with the value of *K* set to 10. This choice of *K* is widely used in EDM research due to its effectiveness in balancing bias and variance. Weka is selected as the primary tool due to its comprehensive functionality for implementing and evaluating data mining models.

A. Demographic Information

Table V illustrates that academic achievement predictions based on demographic information have minimal influence on overall performance across all classifiers. The highest accuracy for the imbalanced dataset is achieved using the DT model, highlighted in bold, with an accuracy of 39.2%. Other models yield slightly lower accuracies, with the NB model performing the worst at 33.5% accuracy. When using a balanced dataset, the impact of demographic information on prediction accuracy remains limited, regardless of the model applied. The best-performing model in this case is the ANN model, achieving an accuracy of 58%, while other models produce slightly lower results. Similar to the imbalanced dataset scenario, the NB model demonstrates the lowest accuracy, at 51.8%. These findings suggest that demographic data contribute minimally to academic performance prediction, indicating that demographic attributes are not a significant factor for effective classification.

TABLE V
PERFORMANCE BASED ON DEMOGRAPHIC INFORMATION

Classifiers	Accuracy (%)	
	Imbalanced	Balanced
KNN	36.9	52.3
NB	33.5	51.8
DT	**39.2**	56.3
SVM	35.4	57.5
ANN	35.7	**58.0**

In Table VI, the variable prioritization results for demographic datasets are analyzed using a wrapper-based technique. The DT classifier is applied to the imbalanced dataset, while the ANN is used for the balanced dataset, as both classifiers yield the highest accuracy. The analysis reveals that the SGPA achieves the highest accuracy, particularly on the balanced dataset, with a score of 45.8%. In contrast, the remaining demographic factors demonstrate significantly lower accuracy scores. Consequently, variables related to demographic data are found to have minimal impact on student performance and are deemed insignificant in predicting academic outcomes.

TABLE VI
VARIABLE IMPORTANCE BASED ON DEMOGRAPHIC INFORMATION USING DT FOR IMBALANCED DATASET AND ANN FOR BALANCED DATASET

Order	Variables	Accuracy (%) (Imbalanced)	Accuracy (%) (Balanced)
1	SGPA	41.1	45.8
2	Father Occupation	40.7	28.0
3	Mother Occupation	39.5	25.0
4	Mother Income	38.8	34.5
5	Region	38.4	26.0
6	Father Income	37.3	26.3
7	Gender	37.3	27.5
8	Sibling	36.1	32.3

B. Grade Information

This section evaluates the classification performance across three datasets: General Education (GE) courses (12 courses), core courses (18 courses), and elective courses (10 courses). Additionally, the final column presents the classification results based on all courses combined (40 courses), as shown in Table VII.

TABLE VII
PERFORMANCE BASED ON GRADES DATASETS FROM EACH COURSE

	Classifiers	Accuracy (%)			
		GE	Core	Elec.	All
Imbalanced	KNN	59.7	67.7	62.0	73.8
	NB	67.3	**81.4**	74.5	84.4
	DT	53.6	59.3	56.7	61.6
	SVM	63.1	68.8	65.8	76.8
	ANN	61.6	74.5	68.1	80.6
Balanced	KNN	68.0	76.3	74.3	78.0
	NB	75.8	**87.8**	84.8	91.3
	DT	69.0	74.5	72.5	75.0
	SVM	77.3	79.0	81.8	85.0
	ANN	76.5	84.3	81.3	87.8

The reported results represent the estimated accuracy of five models applied to both balanced and imbalanced datasets. For the imbalanced dataset, the NB model achieves the highest accuracy of 81.4% when applied to core courses. However, the balanced dataset yields improved accuracy, reaching 87.8% using the same NB model. In contrast, elective and GE courses exhibit lower predictive performance compared to core courses. Among all the classification techniques, the DT model demonstrates the lowest accuracy. The implementation of the SMOTE enhances classification accuracy, indicating its effectiveness in addressing data imbalance. This suggests that institutions can leverage this technique to classify students' academic performance and formulate policies for improvement. The NB classifier consistently delivers optimal results and outperforms other models when applied to the balanced dataset.

Further analysis is conducted by separating core and elective courses into Information Science (IS) and Information Technology (IT) datasets. The datasets are divided into IS core and IT core, each consisting of 9 courses. Similarly, the elective courses are split into IS elective and IT elective datasets, each containing 5 courses.

TABLE VIII
PERFORMANCE BASED ON GRADES FROM IS AND IT COURSE

	Classifiers	Accuracy (%)			
		IS Core	IT Core	IS Elec.	IT Elec.
Imbalanced	KNN	69.6	65.8	55.1	60.1
	NB	**75.3**	74.5	61.2	68.4
	DT	64.3	59.3	49.8	59.3
	SVM	70.3	65.0	64.3	65.0
	ANN	72.6	63.9	55.9	63.9
Balanced	KNN	77.5	73.5	66.5	72.3
	NB	**85.8**	83.0	76.0	80.0
	DT	77.5	74.0	68.8	74.0
	SVM	83.0	78.5	76.3	79.3
	ANN	80.5	79.0	71.5	76.3

The results presented in Table VIII align with the findings in Table VII, indicating that the NB model achieves the highest accuracy of 85.3% when applied to IS core courses using the balanced dataset. This suggests that IS courses have a greater influence on the target class compared to IT courses across all classifiers. However, when considering only elective courses, IT electives exhibit better predictive performance than IS electives. Since elective courses are typically taken in the later stages of study, greater emphasis is placed on core courses. Experimental findings suggest that core courses consistently yield the highest classification accuracy for both balanced and imbalanced datasets. Among all classifiers, the NB model demonstrates the best performance, making it the most effective classification technique for predicting academic achievement based on course performance.

TABLE IX
VARIABLE IMPORTANCE BASED ON GRADE FROM IS COURSES

Order	Courses	Accuracy (%)
1	Organization of Information Resources	63.5
2	Information and Reference Services	61.8
3	Library of Congress Classification	60.3
4	Library Automation Systems	59.8
5	Cataloging of Information Resources	59.3
6	Information Science	56.5
7	Reading for Information Professional	54.3
8	Collection Development	53.8
9	Management of Information Institutes	48.3

Additionally, Tables IX and X provide a ranking of core courses using the NB classifier as an indicator. The classification of IS and IT courses is based on course descriptions. IS courses primarily focus on information management, whereas IT courses emphasize the application of modern technology to information-related tasks, encompassing principles, theories, and software tools. Among the IS courses, *Organization of Information Resources* has the highest predictive impact, achieving an accuracy of 63.5%. This is followed by *Information and Reference Services* and *Library of Congress*, which yield accuracies of 61.8% and 60.3%, respectively. These findings highlight the significance of IS courses in predicting academic performance, emphasizing their role in shaping students' overall achievement.

TABLE X
VARIABLE IMPORTANCE BASED ON GRADE FROM IT COURSES

Order	Courses	Accuracy (%)
1	Research and Statistics in Information Studies	70.3
2	Information Systems Analysis and Design	63.8
3	Electronic Information and Record Management	61.5
4	Programming in Information Work	55.8
5	Information Technology	53.8
6	Database Management for Information Work	51.3
7	Web Design for Information Work	51.3
8	Seminar on Current Issues and Trend in Information Science	44.0
9	Presentation and Training in Information Work	43.3

On the other hand, the IT dataset includes 9 compulsory courses. Among them, *Research and Statistics in Information Studies* and *Information Systems Analysis and Design* demonstrate superior predictive performance compared to IS courses, achieving accuracies of 70.3% and 63.8%, respectively. Our analysis suggests that individual IT courses exert a stronger influence on academic achievement prediction. However, when considering the entirety of courses within each dataset, IS courses exhibit a closer correlation with the target outcomes. This indicates that while specific IT courses contribute significantly to performance prediction, the overall impact of IS courses remains more substantial in determining academic success.

C. Early-stage GPA Information

Table XI presents a comparison of the five models, incorporating factors such as GPA from the first semester to the sixth semester (GPA1-GPA6) and the average GPA (AGPA) from the second to the sixth semester (AGPA2-AGPA6). Given that each academic year comprises two semesters, the analysis considers GPA information up to the end of the third year. The results indicate that the NB classifier achieves the highest accuracy of 76.8% for GPA4 when applied to the imbalanced dataset. Additionally, GPAs and AGPAs yield consistent performance across all predictive models throughout the six semesters. For AGPA datasets, classification accuracy steadily improves from the end of the second semester to the end of the sixth semester across all classification techniques. Furthermore, in the balanced dataset, GPA5 achieves the highest accuracy of 77.8% using the NB classifier. Therefore, the findings suggest that the NB model provides the most effective prediction, with GPA4 being the best predictor for the imbalanced dataset and GPA5 for the balanced dataset.

TABLE XI
PERFORMANCE BASED ON EARLY-STAGE GPA FROM EACH SEMESTER

	GPA	Accuracy (%)				
		KNN	NB	DT	SVM	ANN
Imbalanced	GPA1	59.7	59.7	59.7	59.7	57.4
	GPA2	64.3	64.3	64.3	64.0	64.3
	GPA3	62.4	62.4	62.4	62.4	62.4
	GPA4	71.1	**76.8**	71.1	71.1	71.1
	GPA5	68.4	68.4	68.4	68.4	68.4
	GPA6	64.3	64.0	64.3	64.3	65.8
	AGPA2	60.1	60.1	60.1	60.1	59.3
	AGPA3	62.4	63.5	63.5	63.5	63.5
	AGPA4	65.0	64.3	65.0	65.0	65.8
	AGPA5	66.5	65.0	66.5	66.5	66.9
	AGPA6	72.2	72.2	72.2	72.2	72.2
Balanced	GPA1	57.5	57.5	57.5	56.0	59.5
	GPA2	58.0	58.0	58.0	58.0	58.0
	GPA3	63.5	63.5	63.5	63.5	62.5
	GPA4	74.0	74.0	74.0	74.0	74.0
	GPA5	**77.8**	**77.8**	**77.8**	**77.8**	**77.8**
	GPA6	76.0	76.0	76.0	76.0	76.0
	AGPA2	62.0	62.0	62.0	62.0	62.0
	AGPA3	62.8	63.5	63.5	63.5	63.5
	AGPA4	68.5	68.5	68.5	68.5	68.5
	AGPA5	68.8	68.8	68.8	68.8	68.8
	AGPA6	71.0	71.0	71.0	71.0	71.0

TABLE XII
VARIABLE IMPORTANCE BASED ON EARLY-STAGE GPA USING NB CLASSIFIER

Order	Imbalanced		Balanced	
	Variables	Accuracy (%)	Variables	Accuracy (%)
1	GPA4	76.8	GPA5	77.8
2	AGPA6	72.2	GPA6	76.0
3	GPA5	68.4	GPA4	74.0
4	AGPA5	65.0	AGPA6	71.0
5	GPA2	64.3	AGPA5	68.8
6	AGPA4	64.3	AGPA4	68.5
7	GPA6	64.0	GPA3	63.5
8	AGPA3	63.5	AGPA3	63.5
9	GPA3	62.4	AGPA2	62.0
10	AGPA2	60.1	GPA2	58.0
11	GPA1	59.7	GPA1	57.5

The ranking of variables using GPA information, as determined by the NB classifier, is presented in Table XII. The GPA from the fifth semester, based on the balanced dataset, yields the highest accuracy compared to other variables. Factors related to GPA and AGPA consistently align with previous findings, demonstrating that as GPA or AGPA values approach the completion of academic studies, the prediction accuracy increases. This indicates that predictions become more reliable closer to graduation. However, this study's objective is to provide institutions with early insights into student performance before course completion. The results suggest that academic achievement can be estimated with an accuracy of 77.8% using GPA from the fifth semester (GPA5), offering a valuable tool for early intervention and support.

TABLE XIII
PERFORMANCE BASED ON THE INCREMENTAL INCLUSION OF GPA FROM EACH SEMESTER

	GPA	Accuracy (%)				
		KNN	NB	DT	SVM	ANN
Imbalanced	GPA1	**59.7**	**59.7**	**59.7**	**59.7**	57.4
	GPA2	63.1	63.5	63.9	**65.8**	64.6
	GPA3	68.4	72.2	**73.8**	71.5	70.3
	GPA4	77.9	**79.1**	76.4	75.7	76.0
	GPA5	81.4	**83.7**	73.8	81.4	79.5
	GPA6	83.3	**85.6**	75.7	84.4	82.1
Balanced	GPA1	57.5	57.5	57.5	56.0	**59.5**
	GPA2	63.8	**65.5**	64.5	64.3	65.0
	GPA3	74.0	75.3	76.5	**77.5**	76.3
	GPA4	81.0	83.5	**84.3**	80.0	83.5
	GPA5	83.5	**86.5**	82.3	84.5	81.5
	GPA6	87.3	**90.5**	86.0	89.8	87.3

Table XIII presents the results obtained by incrementally adding GPA datasets from GPA1 to GPA6. For instance, GPA1 consists of a single factor, while GPA2 includes both GPA1 and GPA2. Similarly, GPA3 comprises GPA1, GPA2, and GPA3, continuing this pattern until GPA6 incorporates all preceding GPA values. The results indicate that adding consecutive GPAs leads to an improvement in accuracy. Especially, the performance of the model on the balanced dataset shows a steady increase in accuracy from GPA1 to GPA6, reaching a maximum accuracy of 90.5%. While these outputs differ slightly from those of the AGPA datasets in Table XI, both results follow a similar trend, where predictive accuracy improves as the dataset includes more GPA information. The NB classifier achieved the highest accuracy, indicating that its combination with SMOTE enhances prediction performance.

V. DISCUSSION

The results reveal that student demographics have a negligible impact on academic performance. The most relevant factor is SGPA, which can be a useful indicator. Among the least influential factors are gender and siblings. Meanwhile, things like parents' occupation and incomes, or the religion of the students, have an even smaller effect. Hence, a student's background has little influence on their final academic achievement.

Grade information for GE courses demonstrated strong predictive accuracy, with the SVM model

achieving 77.3% accuracy, followed by the ANN model at 76.5%. Therefore, the GE course grade data is valuable for prediction.

When analyzing courses in the fields of IS and IT, we discovered that among the core courses, the IS group demonstrated slightly higher predictive accuracy than the IT group, with accuracy rates of 85.8% and 83%, respectively, using the same NB model. Conversely, for elective courses, the IT group performed slightly better than the IS group, achieving 80% accuracy with the NB model, compared to 76.3% accuracy with the SVM model. Overall, those two groups exhibited similar performance, which is beneficial to the curriculum design by considering the courses between both IS and IT.

Early-stage GPA information provides highly accurate predictions and is a valuable dataset for forecasting final academic achievement. In the first three years, the NB model achieved a prediction accuracy of 90.5%, followed by the SVM model with 89.8% accuracy. This type of dataset is especially useful because it covers the earlier stage of the educational journey, unlike individual course grades, which may vary depending on the semester in which students enroll. Some courses are taken later in the program making them less useful for early predictions. When examining GPA per semester, the GPA5 shows the most promising results, with 77.8% accuracy. Analyzing cumulative GPA over the first six semesters further increases the accuracy score. Interestingly, however, the accuracy of cumulative GPA predictions is lower than that of semester-based GPA predictions. This is likely because the GPA of the first year are not a strong predictor for long-term academic success since students were in the process of adapting to get familiar with university life.

The reason that NB outperforms the other models is due to its special characteristics. NB naturally handles categorical features without complex preprocessing, which may complicate other models. In addition, NB works well with the independence feature that has small datasets. Its probabilistic nature and low complexity make it less prone to overfitting than models like decision tree or neural network. For instance, to predict whether a student will be categorized into which class, NB may achieve higher accuracy than more complex models when the dataset is small, imbalanced, or contains mostly categorical data like the dataset used in this study.

The predictive insights derived from student performance data can serve as a foundation for evidence-based educational interventions. For example, if we can identify low-achievement students early through the model, institutions can focus on supporting mechanisms, such as tutoring or academic counseling, to help that group of students. Furthermore, consistent underperformance in specific courses can indicate a weakness in the curriculum. Curriculum committees may have to revise the prerequisite course structure or integrate supplemental instruction into difficult courses. At the strategic level, this knowledge empowers institutions in making decisions for allocating resources more effectively, prioritizing academic support services, and implementing retention strategies. This not only enhances student success rates but also contributes to long-term outcomes such as graduation rates and institutional accountability.

The five different ML models yielded varying results. Some models performed well with certain datasets but not with others, which is why we need to apply multiple models to determine the most effective prediction. Among the models, NB consistently delivered the highest accuracy across most types of datasets. Meanwhile, ANN and SVM performed well but slightly lower than NB. DT and KNN models showed minor accuracy compared to the others. Therefore, regarding this particular case study, if a single predictive model were chosen, NB would be the best option for predicting the final academic performance on graduation.

VI. Conclusion

This study has explored factors that affect the students' academic achievement using EDM techniques. We applied various datasets, including demographic information, course grades, and early-stage GPA from each semester, in the experiments. Course grades were categorized into three areas: GE, IS, and IT. Moreover, we discovered the distinction between core and elective courses within the area of information science and information technology.

To address the imbalanced dataset, the SMOTE technique was applied to balance the number of instances across each class. We generated classification models using five different classifiers: KNN, NB, DT, SVM, and ANN. The results demonstrate that early-stage GPA provides the highest predictive accuracy among other factors. GPA from the fifth semester had the most significant impact on the prediction. Among the classifiers, the NB model outperforms the others. Among the various types of datasets analyzed, demographic information demonstrated a relatively moderate impact on the performance. The highest accuracy achieved using this data type was 58.0%, obtained with the ANN algorithm, indicating its limited effectiveness compared to other dataset types. On the other hand, grades from core courses with the NB algorithm produced a significant effect on the learning outcomes by achieving an accuracy of 87.8%. The IS core courses provide better predictions than the IT core courses, while IT elective courses slightly underperform compared to their IS counterparts.

The incremental inclusion of early-stage GPA data from GPA1 to GPA6 on the balanced dataset

yielded the highest accuracy, reaching 90.5%. This indicates that having access to GPA information from additional early semesters enhances the accuracy of predicting academic achievement. The application of the SMOTE technique to balance the datasets led to improved performance across all dataset types. Consequently, EDM offers meaningful insights that facilitate early intervention, guide curriculum enhancement, and support student counseling efforts to raise final academic outcomes.

Future research is recommended to focus on integrating additional data sources, such as questionnaires, online learning activities, and other educational engagement metrics, to enhance predictive accuracy. Furthermore, exploring a wider range of machine learning algorithms, including Random Forest, Deep Neural Networks, Linear Regression, and Evolutionary Algorithms, could provide deeper insights and enhance the prediction models. These suggestions are capable of improving student performance prediction and supporting more effective academic decision-making. This research approach, which employs EDM techniques, can be applied to educational datasets with similar characteristics across various curriculum types.

Acknowledgment

This research was funded and supported by the Faculty of Humanities and Social Sciences, Burapha University.

References

[1] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min*., vol. 1, no. 1, pp. 3-16, Jan. 2009, https://doi, org/10.5281/zenodo.3554657

[2] V. L. Migueis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36-51, Nov. 2018, https://doi.org/10.1016/j.dss.2018.09.001

[3] M. Anoopkumar and A. M. J. M. Z. Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," in *Proc. 2016 Int. Conf. Data Min. Adv. Comput. (SAPIENCE),* 2016, pp. 122-133, https://doi.org/10.1109/SAPIENCE.2016.7684113.

[4] W. Xing, "Exploring the influences of MOOC design features on student performance and persistence," *Distance Educ*., vol. 40, no. 1, pp. 98-113, Dec. 2019, https://doi.org/10.1080/01587919.2018.1553560

[5] J. D. Parker, M. J. Hogan, J. M. Eastabrook, A. Oke, and L. M. Wood, "Emotional intelligence and student retention: Predicting the successful transition from high school to university," *Pers. Individ. Diffe*r., vol. 41, no. 7, pp. 1329-1336, Nov. 2006, https://doi.org/10.1016/j.paid.2006.04.022

[6] A. Richard-Eaglin, "Predicting student success in nurse practitioner programs," *J. Am. Assoc. Nurse Pract.*, vol. 29, no. 10, pp. 600-605, Oct. 2017, https://doi.org/10.1002/2327-6924.12502

[7] E. Alqurashi, "Predicting student satisfaction and perceived learning within online learning environments," *Distance Educ.*, vol. 40, no. 1, pp. 133-148, Dec. 2018, https://doi.org/10.1080/01587919.2018.1553562

[8] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," in *Proc. Int. Conf. Comput. Sci. Appl*., 2018, pp. 111-125, https://doi.org/10.1007/978-3-030-03023-0_10

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414-422, 2015, https://doi.org/10.1016/j.procs.2015.12.157

[10] H. Almarabeh, "Analysis of students' performance by using different data mining classifiers," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 8, pp. 9-15, Aug. 2017, https://doi.org/10.5815/ijmecs.2017.08.02

[11] A. Raheela, M. Agathe, A. A. Syed, and G. H. Najmi, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177-194, Oct. 2017, https://doi.org/10.1016/j.compedu.2017.05.007

[12] Y. Baashar et al., "Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs)," *Appl. Sci.*, vol. 12, no. 3, pp. 1-16, Jan. 2022, https://doi.org/10.3390/app12031289

[13] Y. Baashar et al., "Evaluation of postgraduate academic performance using artificial intelligence models," *Alexandria Eng. J*., vol. 61, no. 12, pp. 9867-9878, Dec. 2022, https://doi.org/10.1016/j.aej.2022.03.021

[14] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telemat. Inform*., vol. 37, pp. 13-49, Apr. 2019, https://doi.org/10.1016/j.tele.2019.01.007

[15] B. Owaidat, "Exploring the Accuracy and Reliability of Machine Learning Approaches for Student Performance," *Appl. Comput. Sci.*, vol. 20, no. 3, pp. 67-84, Sep. 2024, https://doi.org/10.35784/acs-2024-29

[16] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, "Analyzing students' academic performance using educational data mining," *Comput. Educ.: Artif. Intell.*, vol. 7, p. 100263, Dec. 2024, https://doi.org/10.1016/j.caeai.2024.100263

[17] J. Zimmermann, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," *J. Educ. Data Min*., vol. 7, no. 3, pp. 151-176, Oct. 2015, https://doi.org/10.5281/zenodo.3554733

[18] S. Alex, J. J. V. Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl. Soft Comput.,*vol. 156, p. 111491, May 2024, https://doi.org/10.1016/j.asoc.2024.111491

[19] S. A. Alex, "Classification of imbalanced data using SMOTE and autoencoder based deep convolutional neural network," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.,* vol. 31, no. 3, pp. 437-469, 2023. https://doi.org/10.1142/s0218488523500228

[20] E. Alyahyan and D. Dustegor, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 3, pp. 1-21, Feb. 2020, https://doi.org/10.1186/s41239-020-0177-7

[21] S. Jayaprakash, "A Survey on academic progression of students in tertiary education using classification algorithms," *Int. J. Eng. Technol*., vol. 8, no. 6, pp. 111-115, Feb. 2019.

[22] W. Intayoad, C. Kamyod, and P. Temdee, "Synthetic minority over-sampling for improving imbalanced data in educational web usage mining," *ECTI Trans. Comput. Inf. Technol.*, vol. 12, no. 2, pp. 118-129, Feb. 2019, https://doi.org/10.37936/ecti-cit.2018122.133280

[23] A. AL-Ashoor and S. Abdullah, "Examining techniques to solving imbalanced datasets in educational data mining systems," *Int. J. Comput*., vol. 21, no. 2, pp. 205-213, Jun. 2022, https://doi.org/10.47839/ijc.21.2.2589

[24] S. Aliga, A. S. Gaafar, and A. K. Hamoud, "Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection," *Informatica*, vol. 47, no. 1, pp. 11-20, 2023, https://doi.org/10.31449/inf.v47i1.4519

[25] K. Sutha and J. J. Tamilselvi, "A review of feature selection algorithms for data mining techniques," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 6, pp. 63-67, Jun. 2015.

[26] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70-79, Jul. 2018, https://doi.org/10.1016/j.neucom.2017.11.077

[27] M. Zafari, A. S. Niaraki, S. M. Choi, and A. Esmaeily, "A practical model for the evaluation of high school student performance based on machine learning," *Applied Sciences*, vol. 11, no. 23, p. 11534, Dec. 2021, https://doi.org/10.3390/app112311534

[28] M. R. Islam, A. M. Nitu, M. A Marjan, M. P. Uddin, M. I. Afjal, and M. A. A. Mamun, "Enhancing tertiary students' programming skills with an explainable educational data mining approach," *PLoS ONE*, vol. 19, no. 9, e0307536, Sep. 2024, https://doi.org/10.1371/journal.pone.0307536

Knitchepon Chotchantarakun is a lecturer in the Department of Information Studies, Faculty of Humanities and Social Sciences at Burapha University (BUU), Thailand. He received his Ph.D. in Computer Science from the Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Thailand, in 2021. He earned his M.Sc. in Computer Science from Chulalongkorn University (CU), Thailand, in 2006, and his B.Sc. in Computer Science with second-class honors from Mahidol University International College (MUIC), Thailand, in 2003. His research interests include evolutionary algorithms, feature selection, optimization, data mining, and machine learning.