MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2023, 19(12), em2369
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/13863

Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning

Chayaporn Kaensar [1*], Worayoot Wongnin [1]

[1] Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, THAILAND

Received 27 July 2023 ▪ Accepted 01 October 2023

Abstract

Recently, many global universities have faced high student failure and early dropout rates reflecting on the quality of education. To tackle this problem, forecasting student success as early as possible with machine learning is one of the most important approaches used in modern universities. Thus, this study aims to analyze and compare models for the early prediction of student performance with six machine learning based on Thailand's education curriculum. A large dataset was collected from the admission scores of 5,919 students during 2011-2021 of 10 programs in the Faculty of Science at Ubon Ratchathani University. The methodology was carried out using Jupyter Notebook, Python 3, and Scikit-Learn to build the models for prediction. To obtain a higher result, we needed not only to find high-performance prediction models, but also to tune hyperparameter configurations consisting of 138 possible different patterns to identify the best-tuned model for each classifier. Furthermore, we investigated significantly important predictors affecting student success for 10 programs in our faculty. In the experiments, the process was divided into two parts: First, we evaluated effective models using a confusion matrix with 10-fold cross-validation. The results showed that random forest (RF) had the highest F1-measure of 86.87%. While predictive models using fine-tuned RF of 10 programs claimed accuracy of about 72% to 93%. Second, we computed the importance of each feature with fine-tuned RF classifiers. The result showed that national test scores (e.g., ONET-English, ONET-Math, ONET-Science, ONET-Social studies, ONET-Thai, and PAT2), entry type, and school grade (e.g., art, English, GPA, health, math, science, and technology) are highly influential features for predicting student success. In summary, these results yield many benefits for other relevant educational institutions to enhance student performance, plan class strategies and undertake decision-making processes.

Keywords: student performance prediction, hyperparameter tuning, feature importance, machine learning

INTRODUCTION

In recent years, many universities across the globe have been faced with challenging problems in the management of new students. In particular, during the COVID-19 pandemic, many institutes were forced to close their facilities and turn to online teaching modes rapidly, and these events critically affected all educational stakeholders including the university, faculty, teachers, students, and especially the first-year students such as early dropouts, high withdrawal rates, poor academic performance, inadequate learning competence, or ceasing studies at a very early stage of their program (Kornpitack & Sawmong, 2022). However, even though the impact of the pandemic seems to be decreasing, the amount of new student difficulties is still growing.

Similarly, the Faculty of Science at Ubon Ratchathani University (UBU), provides education in science and technology and is located in Northeast Thailand. From the academic record between 2011-2021, statistical information has been collected from the UBU Registrar's Office (REG UBU), which stores course registration data (Ubon Ratchathani University, 2010). Several problems have been reported, as follows: First, almost 21% of all early dropouts occur during the first academic year of

✉ chayaporn.k@ubu.ac.th **(*Correspondence)** ✉ worayoot.w@ubu.ac.th

Contribution to the literature

- This study not only provides early student performance predictions using six machine learning methods based on hyperparameter tuning, but also explores important predictors for student success in the first year of studies.
- This study further elevated the idea that all predictive models constructed should be tuned and tested according to different hyperparameter settings, resulting in the best performance used to compare and determine performance results.
- This study used 19 input score data of about 5,919 students from 2011-2021. Another finding identifies which factors affect first-year student success in each program at the Faculty of Science Aat UBU. This can benefit those programs that are considered for targeting additional supports to student at an early stage.
- This study assists other relevant educational institutes improving student retention, reducing early dropouts and failure rates, planning teaching strategies, determining criteria for future admissions and preparing appropriate readiness courses for new students.

studies, while about 37% are defined as having poor academic performance, but the remainder was still registered in the program. Due to this, the student failure rate tends to steadily increase every year, which affects the quality of education and budget losses. Second, our new students lack essential academic skills for studying programs in the field of science and technology. This makes them most likely to drop out early, delay graduation, or even cease their studies. Furthermore, many public universities in Thailand (including UBU), are also faced with challenges in determining and assessing student scores (e.g., the national test score and the school subject score) for the admission and recruitment process to ensure a student's suitability for a specified program (Kornpitack & Sawmong, 2022). This was negatively affected because if new students are not properly cared for and prepared for an educational program, it will prove to be a drawback for the university.

To minimize this problem, our faculty has provided many plans and aids: establish readiness preparation to improve student skills before starting the first year of studies, survey student profiles from online visualized reports to analyze and discover insights into their background knowledge and use human expert experience to define proper admission criteria, setting it to select suitable students who might fit with and succeed in the program. However, those existing approaches might be helpful for students, faculty and the university, but for efficient decision-making and prediction, it is still inadequate. So, providing additional educational tools serving as a support system to resolve the issues above is also required.

Consequently, the process of improving the educational system by providing intelligent and accurate predictions to predict student success early. Many modern universities consider this to be one of the most effective approaches (Roslan & Chen, 2020). In an educational context, machine learning techniques are a potential approach widely used to accomplish the problems of predicting education outcomes, forecasting student behavior and improving educational quality. This is primarily applied to predict and support decision-making such as analyzing student behavior demography (Bilal et al., 2022; Kaensar & Wongnin, 2023), predicting student performance (Usman et al., 2017; Yagci, 2022), and identifying the relationship between student data and their achievement (Chang & Wang, 2016; Qahmash et al., 2023). However, although those published works could provide educational benefits, those features, and the experimental results tend to be quite limited. For example, a lack of tuning parameters for implementing a model does not take into consideration school subject data, which is most impactful for predicting student performance, thus providing a low accuracy rate, small sample sizes and typically focusing on one department or even just at the course-level.

Thus, to fill this gap and differentiate from previous work by gaining insight from a large dataset over ten years. This study aims to create a machine learning model to discover and compare the best algorithm for predicting student performance and analyze which factors affect first-year student success in each program at the Faculty of Science at UBU. To compare models, six different algorithms such as a decision tree (DT), linear regression (LR), multi-layer perceptron (MLP), naïve Bayes (NB), random forest (RF), and support vector machine (SVM) were used in the experiment. Hence, to provide more accurate results, each predictive model ensured that possible hyperparameters would be also adjusted properly to find the best pattern for achieving good academic performance. Based on the results, we employed student score records in 10 different programs in the Faculty of Science at UBU in experiments and reports.

The large data used in this work was obtained from the UBU Registrar's Office, which comprised about 5,919 student records from 2011-2021, across 10 programs in the Faculty of Science at UBU, such as

(1) Biology (BIO),

(2) Chemistry (CHEM),

(3) Data Science and Software Innovation (DSSI),
(4) Environmental Science (ENV),
(5) Information and Communication Technology (ICT),
(6) Mathematics (MATH),
(7) Microbiology (MICRO-BIO),
(8) Occupational Health and Safety (OCC-HS),
(9) Physics (PHY), and
(10) Rubber and Polymer Technology (RUBBER).

In connection with this, data from 19 test scores based on the university's recruitment system and Thailand's core curriculum, were considered and used such as national admission test data (e.g., GAT, ONET, PAT1, and PAT2), school GPA and eight main school subject scores (art, foreign languages, health and physical education, mathematics, religion and culture, science, social studies, technology and career, and Thai language). To meet the challenges above, we asked research questions listed, as follows:

RQ1. What are the differences among the six machine learning techniques (DT, LR, MLP, NB, RF, and SVM) with fine-tuned parameters used for predicting first-year student performance?

RQ2. How was the best classifier with fine-tuned parameters used to enhance and predict student performance for each program at the Faculty of Science at UBU?

RQ3. How can the best classifier with fine-tuned parameters be used to identify the important attributes that affect student performance in each program at the Faculty of Science at UBU?

Based on this, this study provided benefits and significantly improved the university, faculty, program, teachers, and students in four ways: To begin with the proposed prediction models could guide new students in early prediction of their performance and aid in determining teaching strategies for teachers. Second, the program, faculty, and university can recognize some warning signs of students at risk early to take precautions and help their students. Third, institutions could use the results as decision support systems to define appropriate admission criteria and readiness preparation courses. Finally, this outcome could serve as a model for other relevant faculties and institutions in the field of science and technology.

The following are the remaining sections of this paper: We present a review of related work. Then, we describe the concept and definition. Next, we explain the research methodology. After that we discuss the implementation and experimental results. And finally, we summarize the findings and outlines future directions.

LITERATURE REVIEW

In the past few years, there has been an increasing number of students failing and even dropping out during their first year of graduation over time and this has an impact on teaching efficiency indicators or even additional costs to the institutions and the nation. Therefore, the management of universities is to improve the quality of education for all stakeholders by studying and analyzing the prediction of student performance and identifying factors that affect student performance. In this part, we review related work from two areas, detailed, as follows:

Applying Machine Learning to Predict Student Performance

A traditional statistical method was used in several studies, such as Kemda and Murray (2021) and Mothial et al. (2018), to predict student performance in the pre-intelligent system era. However, they still struggled with large datasets and became less reliable for prediction. Conversely, the machine learning model proved its worth in terms of high efficiency (Ko & Leu, 2021; Raschka, 2015; Sathe & Adamuthe, 2021).

In the first stage of the literature review, we found that only a single technique was used to analyze student educational data in many approaches that are frequently used. For example, Usman et al. (2017) tools were developed to predict student entry tests using regression technique. They used demographic data obtained from 5,042 students at the University of Engineering and Technology, Pakistan such as academic data, age, gender, and interests. However, this study was implemented by using the MATLAB 2015. Qahmash et al. (2023) applied MLP to analyze pre-admission tests (e.g., GAT-General aptitude test, HSP-High school percentage, and SAAT-Standard achievement admission test) of medical colleges to predict student performance in the first two years. The results proved that both GAT and SAAT scores are very strong variables for the prediction of medical student performance. Bengesai and Pocock (2021) used a DT to capture data consisting of demography and academic performance scores (AP score) in a school of 1,370 students studying at a South African university from 2012 to 2013, to predict whether the student remains at the university or not. The results showed that AP scores and financial status were the most important variables related to student perseverance, while personal data such as gender, school, and residence had less significance for classifying students at risk of dropping out. Subsequently, Santosa et al. (2021) employed k-means clustering to predict students' grades using university entrance test scores and English scores. The output model of this prediction had an accuracy of 78.59%. Moreover, it is notable that English skill variables were applied to assess admissions. While Rajagopal (2020) predicts student

admission to university using logistic regression (LOR) and factor variables such as admission chance scores, experience rating, GRE scores, GPA, and TOEFL. This study is useful for many universities for predicting admissions, selecting candidates, and planning timelines efficiently. The model resulted in an accuracy value of 87.50%, and GRE and TOEFL scores have a clear impact on student performance. Dabaliz et al. (2017) applied LR to analyze the pre-admission variables (e.g., GAT, high school score, IELTS, NAT-National achievement test, and TOEFL) for 737 students to predict medical student performance in Saudi Arabia. This proved that NAT and TOEFL scores are very important predictors during the preclinical year.

In addition, comparing several machine learning techniques based on prior student academic data from schools to predict their performance in higher education has been widely proposed. For example, Nurhachita and Negara (2021) applied and compared three machine learning techniques (deep learning [DL], NB, and RF) to predict student performance at the University of Indonesia based on their school's features. The results showed that NB provides more accuracy than the others at 99.97%. Correspondingly, neural network (NN) and RF classifiers gave the best prediction results in Mengash (2020) and Singh et al. (2020) at 79% and 96%, respectively. Next, Maksimova et al. (2022) explored four classifiers (LOR, NB, NN, and SVM) to analyze data collected from pre-university test scores such as essays, interviews, GPA, and mathematics. This work sought to produce and compare models for predicting the early dropout of computer science students. Also, they identified whether pre-university data have a significant effect on the failure rate. As a result, it showed that SVM could predict dropout rates with more than 70% accuracy. Adekitan and Noma-Osaghae (2019) applied six techniques including Ada boost, DT, LOR, NB, NN, and RF to predict new student success at the University of Nigeria using collected admission data such as GPA, joint admissions and matriculation board, scholastic aptitude test (SAT), and West African examinations council scores, but the regression model provided the highest result but with an accuracy of only 51.90%.

Although many projects offer great benefits for constructing and comparing models using machine learning techniques in general, considering key attributes to predict students' success or failure while meeting educational needs is still required, and is explored in the next section.

Identifying Important Attributes Used to Predict Student Performance

This section will analyze factors associated with machine learning models to identify which affect student performance. Recently, many researchers focused on important influential factors and prediction methods. The criteria of research are shown, as follows:

Devi and Ratnoo (2022) used 330 students' data (e.g., family background, personal profile, and school score) to determine what factors are important for dropout students using RF. In these cases, their results obtained 86% accuracy and showed the performance in high school subjects, income and father's education are relevant to student dropout. Yang et al. (2022) discovered factors affecting academic performance in the blended learning of freshmen using LR for predictions at Central China Normal University. The results indicated that online behavior like the number of posts, the number of replies, and the amount of learning time, are significant for student success. That is, student performance will be high for those who frequently post a question and reply with answers to course material through online learning. Subsequently, Gutierrez et al. (2022) used 255 students' entry data to predict first-year student performance at the Faculty of Engineering, University of Northeast Mexico, based on admission data such as school GPA, SAT, and SHA (study habits and attitudes test). They presented results of prediction using a correlation test that determines the significance of school GPA and SAT as important variables in predicting academic performance. Likewise, Holladay et al. (2020) applied simple and multiple linear regression obtained from 113 students at the College of Veterinary Medicine, University of Georgia from 2015 to 2017 to predict students' first-year performance. Various variables of admission data such as GPA, GPASci (GPA in science courses), GPALast45hr (GPA for the last 45 credit hours), GRE-QV (quantitative and verbal reasoning measures), and GRE-AW (analytical writing measures) were gathered and used for experimentation. It has been proved that all attributes in the admission score dataset influence academic performance for first-year students, particularly attributes like GPALast45hr, GPASci, and GRE-QV were found to be the most impactful attributes.

However, to cover recent advances and challenges, published research combined many tasks like constructing predictions, creating comparisons, and discovering important attributes to predict future student performance. For example, Ko and Leu (2021) applied seven classifiers (e.g., association rule [AR], Bayesian network, DT, KNN, LOR, MLP, NB, and SVM) and used 215 students' data, finding that NB provides greater accuracy at 83.26%, over other methods. While AR addressed important factors, tracking weekly progress and believing in self-efficacy, have been proven successful receiving good grades. However, the limitations are concerns about the small size of the data. Later, many researchers evaluated different machine learning techniques with the same objectives, but they used larger datasets. For example, Huynh-Cam et al. (2022) collected 4,036 students to predict learning

performance by classifying student groups such as local, international, and disabled students. As a result, there was a significant difference in performance based on course credits and elective credit requirements, parental income, and education levels. Similarly, Huynh-Cam et al. (2021) used 2,407 students' records collected from 12 departments at Taiwan Vocational University. They found that influential features for predicting student performance are their parents' occupation, the department, and admissions status. Correspondingly, Backham et al. (2023) constructed a high predictor model using three techniques (DT, MLP, and RF) based on 15 features that covered different demographical dimensions. The result demonstrated that MLP with a constructed 12-input NN achieved the greatest performance with an RMSE of 4.32. To discover important factors, they employed the Pearson correlation to compute values. The results revealed that factors like age, mother's education, and past course failures dominantly influence student grades. In another investigation, Assami et al. (2022) developed four prediction models applying LOR, NB, RF, and SVM to identify the best technique to predict student motivation and select the right course of study using 238 Canvas network online courses. Notably, RF acts as a good classifier based on accuracy at 95.24%. Furthermore, they used RF to identify feature importance techniques among factors associated with learner motivation. Their results indicated that MOOC feature (e.g., course end dates, course requirements, and grade) has a more significant predictive relationship, while other learner features (e.g., education level and primary reason) are important attributes but have a low positive correlation.

Research Gap

According to the literature, although many studies utilize a student's past academic record such as national admission tests and school GPA to determine student performance, while the results might be satisfactory, we noticed that there are some limiting factors with such work. For example, collection from a small dataset of only one course or one department, considering just a few algorithms, a lack of tuning-appropriate parameters, or even determining only generalized input variables namely: GPA, TOEFL, science, mathematics, and attitudes test.

Therefore, this study collected a large set of 5,919 students from 2011-2021. This score dataset consisted of a national test, school GPA, and school subject scores consisting of eight subject areas based on the Thai education system, such as

(1) Art,

(2) Foreign languages,

(3) Health and physical education,

(4) Mathematics,

(5) Science,

(6) Social studies, religion, and culture,

(7) Technology and careers, and

(8) Thai Language.

To determine which techniques were most suitable for UBU's performance prediction dataset analysis, six algorithms were analyzed and reviewed with a high success rate, as shown in (Ko & Leu, 2021; Raschka, 2015; Sathe & Adamuthe, 2021). These were used for creating and comparing the predictive model by fine-tuning parameters throughout and were considerably more accurate than the default settings. In addition, we also sum up and rank factors that affect new student performance in each program at the Faculty of Science at UBU.

CONCEPT

The techniques that we used in this study are explained in this section, as follows.

Decision Tree

DT is a flowchart-like tree structure that consists of a set of decision nodes, connected by branching nodes, extending from the root node until connecting at leaf nodes. Each branch represents an outcome of the test, while the leaf node stores the class label. DT can handle multidimensional data because the representation of knowledge in the tree is easily and intuitively assimilated by humans. DT is simple, fast, and has great accuracy, which is used for classification and regression tasks in many areas.

Linear Regression

LR is a statistical test applied to a data set to define and quantify the relationship between the considered variables and is the simplest regression model. LR can measure the association between two variables, showing the relationship between a dependent variable, y, and an independent variable, x. A regression model is used for descriptions that help analyze the strength of the association between the outcome (dependent variable) and predictor variables, which can be applied to a wide variety of phenomena, cross-sectional surveys, marketing and economic research, and experimental designs and predictions. LR is simple to implement and interpret.

Multi-Layer Perceptron

MLP is the most widely known and most frequently used type of NN, which uses back-propagation to adjust the weights for training. This network consists of three layers: the input layer, hidden layer, and output layer, with each layer having one or more neurons. Each layer comprises linear or nonlinear neurons and each neuron sums its weighted inputs and yields an output through a nonlinear activation function with a bias threshold. In

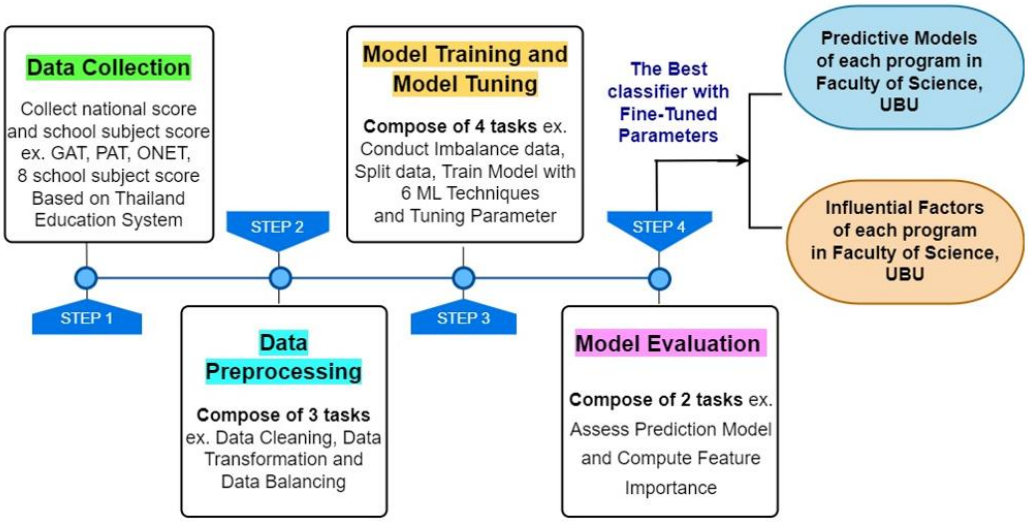

Figure 1. Research methodology (Source: Authors' own elaboration)

MLP, the learning process involves updating parameters including weights and biases. Training can be divided into three main steps: forward propagation, error/loss calculation, and backpropagation. This classifier has been used extensively in classification and regression. MLP can be applied to complex non-linear problems and is efficient for large data sets.

Naïve Bayes

NB is a simple machine learning supervised method and is a powerful classifier for independent attributes and was implemented using the Bayes theorem as shown in Eq. (1). Given class *C* is the class used for prediction, where a child has the parameter (*A1*, *A2*, …, *An*) then the value of *C* that maximizes the *P*(*C*|[*A1*, *A2*, ..., *An*]) must be identified.

$$(C|A) = \frac{P(A|C)P(C)}{P(A)}. \quad (1)$$

Eq. (1) describes the relationship of the conditional probability of statistical quantities. This classifier has been used for classification, other decision-support applications, and predictions in various domains.

Random Forest

RF is a supervised machine learning approach and a computationally efficient technique used for classification and regression problems. It is operated by building multiple DTs during the training period and producing average forecasting of all the DTs involved. RF is composed of a DT of given training data and matching the test data, which are used to rank the importance of the variable in the problem. This algorithm can be imported from sklearn as was the linear model and can be applied for measuring variable importance, weighing class, visualization, missing value imputation, and prediction. RF can generate highly accurate classifier and their predictive models are robust to over-fitting.

Support Vector Machine

SVM is an algorithm for the classification of both linear and nonlinear data. It used nonlinear mapping to transform the original training data to a higher level, which provided a set of training examples, each marked as belonging to one of the many categories. Also, the SVM training algorithm creates a model that predicts the category of new examples. It has a greater ability to generalize problems, which is the goal of statistical learning. However, the statistical learning theory provides an outline for studying the problem of gaining knowledge, making predictions, and making decisions from a set of data (Raschka, 2015).

METHODOLOGY

This study sought to predict the student performance of new students at the Faculty of Science at UBU, and also identifies factors of admission score that dominantly affect their performance grouped by science and technology fields. In this section, we provide details about the dataset, discuss data preprocessing, perform model training and tuning, evaluate the models, and summarize this study. The methodological structure of this study is illustrated in **Figure 1**.

Data Collection

To predict the first-year status of new students, we have used a total of 5,919 student records at the Faculty of Science at UBU from the academic year 2011-2021, across 10 programs such as BIO, CHEM, DSSI, ENV, ICT, MATH, MICRO-BIO, OCC-HS, PHY, and RUBBER. This dataset was provided by REG UBU database containing student information, as shown in **Figure 2**.

After categorization, we used score data solely for considering and monitoring student performance. Based on Thailand's core curriculum and the university's recruitment system, 19 features consisting of national

SeqID	Program	Entry Year	Status	Class	entry type	University GPA	School-GPA	SC-Thai	SC-Math	SC-Sci	SC-Soc	SC-Het	SC-Art	SC-Tech	SC-Eng	ON-Thai	ON-Soc	ON-Eng	ON-Math	ON-Sci	GAT 1	GAT 2	PAT 1	PAT 2
2455	Mathemati	2019	Graduated	1	1	2.76	2.99	2.5	2.63	2.66	3.09	3.85	3.75	3.41	1.87	52	33	17.5	15	34	67	75	85	77
2456	Chemistry	2019	Transfered	0	2	3.48	3.55	4	2.52	3.27	3.82	4	3.75	3.75	3.7	66.5	43	17.5	25	37.9	112	102	102	98
2457	Occupatio	2019	Graduated	1	2	2.87	3.01	2.92	2.26	2.86	3.86	3.8	3.7	4	0	64.5	32	17.5	20	36.3	101	115	114	87
2458	ICT	2019	Withdrawal	0	3	1.88	2.21	1.41	1.35	1.8	2.21	3.37	2.33	2.7	1.73	44	32	22.5	22.5	27.75	55	48	59	52
2459	ICT	2019	Graduated	1	1	3.37	2.43	3.3	3.1	3.12	3.65	4	0	0	3.71	54.5	36	42.5	10	27.05	89	90	110	92
2460	Micro biol	2019	Graduated	1	1	3.36	3.38	3.66	2.72	2.91	3.91	3.9	4	3.92	3.36	66.5	42	25	12.5	27.05	80	85	101	95
2461	Occupatio	2019	Graduated	1	1	3.21	3.38	3.58	2.76	2.91	3.5	3.85	3	3.79	3.5	60.5	44	26.25	25	29	69	97	102	77
2462	Occupatio	2019	Graduated	1	2	3.06	3.11	2.7	2.84	2.77	3.44	3.8	3.9	3.5	3.1	51.5	31	13.75	22.5	30.8	75	95	0	0
2463	ICT	2019	Graduated	1	3	2.87	0.00	0	0	0	0	0	0	0	0	55	46	23.75	27.5	34.7	0	0	45	54
2464	Chemistry	2019	Graduated	1	2	3.43	3.47	3.25	3.08	3.55	3.87	3.95	3.83	3.83	3.07	46.5	39	22.5	30	40.05	101	89	105	109
2465	Occupatio	2019	Studying	0	2	1.97	2.00	1.21	0	1.5	2	2	2.5	2.5	1	55	46	23.75	27.5	34.7	0	0	45	54
2466	DSSI	2019	Graduated	1	2	3.01	3.10	3.5	2.84	2.94	3.46	3.6	2.3	3.5	2.35	47.5	34	18.75	22.5	26.35	90	105	99	101
2467	Biology	2019	Retire	0	1	1.54	2.00	1.25	1.5	0	2.12	0	0	0	1	38.5	34	45	22.5	15.15	15	0	0	0
2468	Occupatio	2019	Studying	0	2	1.74	2.50	2.7	1.98	1.78	2.5	3.3	3.3	3.1	2.76	28	17	18.75	15	23.5	71	61	38	45

Figure 2. Screenshot of student data (Source: Authors' own elaboration)

Table 1. Student admission data descriptions used for prediction

Type	ID	Attribute	Information	Possible value
School subject data	1	SC-GPA	Cumulative GPA in high school	0-4.00
	2	SC-Thai	High school Thai language grade	
	3	SC-Math	High school mathematics grade	
	4	SC-Sci	High school science grade	
	5	SC-Soc	High school social studies grade	
	6	SC-Het	High school health and physical education grade	
	7	SC-Art	High school art grade	
	8	SC-Tech	High school career and technology grade	
	9	SC-Eng	High school English language grade	
National test data	10	ON-Thai	Thailand's ONET of English subject	0-100
	11	ON-Soc	Thailand's ONET social studies score	
	12	ON-Eng	Thailand's ONET English score	
	13	ON-Math	Thailand's ONET mathematics score	
	14	ON-Sci	Thailand's ONET science score	
	15	GAT1	GAT in critical & logical thinking score	0-300
	16	GAT2	GAT in analytical thinking skill in English communication score	
	17	PAT1	PAT in mathematics score	
	18	PAT2	PAT in science score	
University data	19	Entry type	Type of Thai university recruitment systems	1-Quota, 2-Portfolio, 3-Admission, & 4-Direct admission
Target	1	Class	First-year status for new students	1 (GPA≥2.00) 0 (GPA<2.00, dropped out, & ceasing)

Note. GAT: General aptitude test; ONET: Ordinary national educational test; & PAT: Professional aptitude test

test data (e.g., GAT, ONET, and PAT), university data (entry type and ID), school subject data (e.g., art, foreign languages, health and physical education, mathematics, religion and culture, science, social studies, technology and careers, and Thai language) were included in this analysis.

However, we found that data might be incomplete and still contain missing values, so cleaning and correcting all of the data by removing redundancy, noise, outliers, and inconsistencies was required. This process will be explained in the subsequent sections.

Data Preprocessing

To provide a suitable dataset, we utilized Python version 3 with Jupyter Notebook, which is an effective tool for managing the data. **Table 1** shows the student admission data descriptions used for prediction. In this study, the data preparation is done in three steps.

Data cleaning

At this stage, we screen and review the original data to correct it and reduce the error rate in the model. To operate the data, the original number of records 5,919 was identified and cleaned to include 3,407 records. Using these values, we first removed the student ID, prefix, student name, sex, school name, and province, and trained strictly within the scope of the 19 attributes. Next, the null values, missing values, or even N/A values were checked and corrected. Also, incorrect values and outlier data such as student data without any scores will be dropped.

```
#read ONET and GATPAT score from the csv files
onet = pd.read_excel('onet.xlsx')
gatpat = pd.read_excel('GAT-PAT.xlsx')


#add column of target to the dataframe
onet['target'] = onet['status'].apply(lambda x: 1 if x in ['graduated', 'studying'] else 0)

#drop some rows that do not contain ONET scores
onet = onet.dropna()

#merge ONET and GATPAT to a single dataframe
onetgatpat = pd.merge(onet, gatpat, how='inner', on='studentcode')

#initial the missing values with zero
onetgatpat = onetgatpat.fillna(0)

#rename each column to a lowercase
newname = dict()
for e in onetgatpat.columns:
    newname[e] = e.lower()
onetgatpat.rename(columns=newname, inplace='True')
```

Figure 3. Screenshot of data preprocessing with Jupyter Notebook (Source: Authors' own elaboration)

Data transformation

After cleaning, we took raw data from the previous process by actions such as changing the format, structure, and values and converting them to a suitable format that is ready to process. To reduce the time-consuming task of manually converting attributes, we first loaded the original dataset into Jupyter Notebook, which is a web-based platform for coding and configuring tasks for data mining (Santosa et al., 2021).

In **Figure 3**, we depicted sample code for data preprocessing. First, GATPAT and ONET data files were read with read_excel(). Next, to manage the dataset, some rows were removed, combining data into a single dataset and some missing values were replaced with dropna(), merge(), and fillna().

Data balancing

Since the dataset includes 3,407 students of which 2,634 (77.31%) are records that have "pass" status and 773 (22.69%) others, we obtained a ratio between the number of passing and failing students. This causes an imbalance of the dataset and risks model overfitting. Thus, the process needs to discover a balanced dataset with a weighted-adjust technique and oversampling (Raschka, 2015; Roslan & Chen, 2020).

However, to overcome imbalance issues and determine a model that predicts student performance in the first academic year, is described in the following sections.

Model Training & Tuning

After data preprocessing, we build the model and train it and then we provide the model with a training dataset using six different machine learning algorithms. To create the model, Python language using NumPy, Pandas, and Scikit-Learn libraries was implemented. The imbalanced data, training models, and handle-tuning workflow are described, as follows.

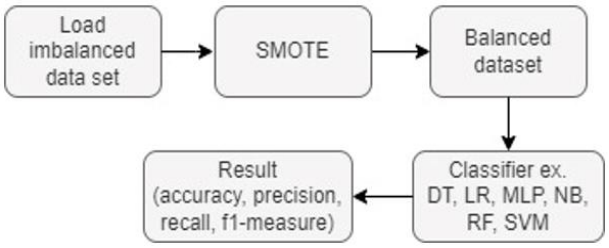

Figure 4. Process flowchart of SMOTE (Source: Authors' own elaboration)

Conduct imbalance dataset

In our dataset, we have found that the minority class has a very low number of instances and leads to imbalanced data problems, which is a general condition encountered in data modeling processes. So, we applied the well-known approach synthetic minority over-sampling technique (SMOTE) and found it could increase and rebalance the number of instances in datasets (Raschka, 2015). To implement the code, the imbalanced-learn Python library in "SMOTE()" class and "imblearn.over_sampling.SMOTE()" were used across all the nodes before combining the results. The process is described in **Figure 4**.

Splitting the dataset

In this step, the dataset will be split into two categories: training and testing dataset by using function the "train_test_split()". Regarding this, the dataset will be divided into a ratio of 70:30, which means use 70% of the training set trains the model, and the remaining set is used to test the purpose and find the accuracy of the model, respectively.

Applying classifier algorithm

To build the model, we applied six algorithms (e.g., DT, LR, MLP, NB, RF, and SVM) to create and train the model differently. That is, each machine learning model

used training datasets labeled as input and output for the model, while a sklearn.ensemble package and Scikit-Learn were imported and implemented in this step.

Tuning the parameters

To optimize results, the tuning parameter process is one task that is used for increasing accuracy (Ali et al., 2019, 2023). In this study, we not only compared predictive model-building from different algorithms but also changed and defined suitable hyperparameter settings, A few examples of tuned parameters of RF algorithms are shown in **Figure 5**.

To address this, 138 different algorithm configurations were run, with the best trials chosen for accuracy. As a process algorithm, the step involves:

- choosing classifier algorithms,
- defining possible hyperparameters for those algorithms,
- fitting them to the model, and
- evaluate the model.

Fit the model

To build and fit the model, we described the steps with a classification algorithm as presented in **Figure 6**.

```
Model = RandomForestClassifier(),
        RandomForestClassifier(n_estimators=3),
        RandomForestClassifier(n_estimators=10),
        RandomForestClassifier(n_estimators=30),
        RandomForestClassifier(n_estimators=50),
        RandomForestClassifier(n_estimators=150),
        RandomForestClassifier(n_estimators=300),
        RandomForestClassifier(n_estimators=500),
        RandomForestClassifier(n_estimators=1000),
        RandomForestClassifier(n_estimators=3, max_depth=3),
        RandomForestClassifier(n_estimators=10, max_depth=3),
        RandomForestClassifier(n_estimators=30, max_depth=3),
        RandomForestClassifier(n_estimators=50, max_depth=3),
        RandomForestClassifier(n_estimators=150, max_depth=3),
        RandomForestClassifier(n_estimators=300, max_depth=3),
        RandomForestClassifier(n_estimators=500, max_depth=3),
        RandomForestClassifier(n_estimators=1000, max_depth=3),
        RandomForestClassifier(n_estimators=3, max_depth=5),
        RandomForestClassifier(n_estimators=10, max_depth=5),
        RandomForestClassifier(n_estimators=30, max_depth=5),
        RandomForestClassifier(n_estimators=50, max_depth=5),
```

Figure 5. An example of RF classifier & corresponding hyperparameters (Source: Authors' own elaboration)

Thus, when a classifier model's object is created and the parameters are determined, all sampling objects expose a function provided by Scikit-Learn like "model.fit()" that takes a dataset to fit the data. Then, the model will be predicted to test model prediction by using the "model.predict()" method. Next, the process of system evaluation was measured, explained in the next sections.

```
for model in Model:
    model_result[model]={'max_fold':[],
              'max_f1':0,
              'max_evaluate':pd.DataFrame(columns =['accuracy','precision','recall','f1'])
}

student_id = onetgatpat_prog['studentcode'].tolist()

random.shuffle(student_id)
Fold = []
fold_size = int(len(student_id)/fold)
for i in range(fold):
    Fold.append(student_id[i*fold_size:(i+1)*fold_size])

for model in Model:
    evaluation = pd.DataFrame(columns =['accuracy','precision','recall','f1'])

    for i in range(fold):
        test_set = onetgatpat_prog[onetgatpat_prog['studentcode'].isin(Fold[i])]
        train_set = onetgatpat_prog[onetgatpat_prog['studentcode'].isin(Fold[i])==False]

        sm = imblearn.over_sampling.SMOTE()
        x_sm, y_sm = sm.fit_resample(train_set[input_list], train_set['target'] )

        model.fit(x sm, y sm)

        pred = model.predict(test_set[input_list])

        cf_matrix = confusion_matrix(test_set['target'],pred)

        accuracy = accuracy_score(test_set['target'], pred)
        precision = precision_score(test_set['target'], pred)
        recall = recall_score(test_set['target'],pred)
        f1 = f1_score(test_set['target'],pred)

        evaluation = evaluation.append(pd.DataFrame([[accuracy,precision,recall,f1]],
                                                    columns=evaluation.columns))
        evaluation = evaluation.mean()
        model_result[model] = evaluation
        model_result[model]['feature_importance'] = model_result[model].feature_importances_
```

Figure 6. Source code for building model (Source: Authors' own elaboration)

Evaluation

Based on the research question, the tasks of evaluation can be classified into two parts: First, the model's performance was assessed and compared. Second, feature values were identified that contained the information most relevant to student performance. This measurement is assessed, as follows:

Assess the prediction model

To measure the efficiency of the model, we employed confusion matrix, which is a widely used approach to assess the accuracy of a classification model, and also applied 10-fold cross-validation with several tests to avoid over-fitting problems (Raschka, 2015). That is, the dataset was divided into 10 equal-sized sets, while nine were used for training and one was used for testing. This process was repeated recursively, and new tests were also taken after each iteration. Hence, the accuracy, precision, recall, and F1-measure of six classification algorithms were reported and compared, with a short detail of each metric given in Eq. (2)-Eq. (5):

- **Accuracy** is the ratio between the number of samples that are correctly classified against the total number of samples:

$$Accuracy = \frac{True\ positive + True\ negative}{Total\ examples}. \quad (2)$$

- **Precision** denotes the proportion of predicted positive cases that are correctly real positives:

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}. \quad (3)$$

- **Recall** is the proportion of real positive cases that are correctly predicted positive:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}. \quad (4)$$

- **F1-score** also known as the F score is the harmonic mean of the precision and recall of a model. A way to combine the precision and recall of a model is given, as follows:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (5)$$

After that a comparison of measurement classifiers was computed to produce model performance and discover its accuracy level by using the best classifier technique for discovering the relation between features that influence the success of students. This result can be obtained by a plot of feature importance, which will be described in the next step.

Computing feature importance

Here we calculate feature importance values in terms of the best classifier using the coefficients feature importance of Scikit-Learn provided by the attribute "feature_importance". In this context, high scores in feature importance analysis for predicting student performance were considered important for predicting the output. To compute the feature importance of any model, the calculation method provides evaluation insight through Entropy scores in Eq. (6), where the index value is calculated for each class by deducting the summation of squared probability *(P)*.

$$Entropy = 1 - \sum_{i=1}^{n}[P(c_i) \times log_2(P[c_i])]. \quad (6)$$

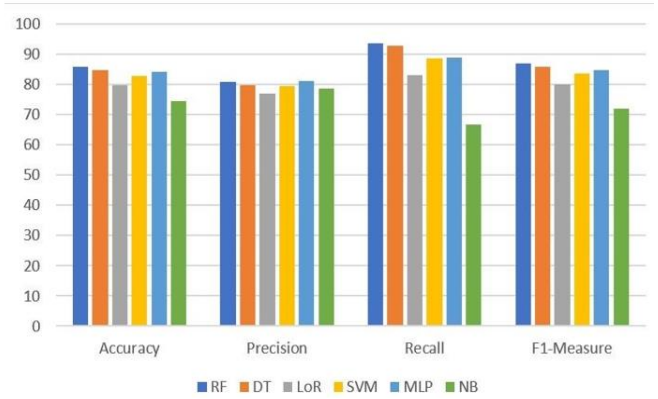

Figure 7. Performance comparison of six machine learning algorithms (Source: Authors' own elaboration)

EXPERIMENTAL RESULTS & DISCUSSION

The structure of experimental results and discussion to address each of three research questions follow below.

RQ1. What Are the Differences Among the Six Machine Learning Techniques (DT, LR, MLP, NB, RF, and SVM) With Fine-Tuned Parameters Used for Predicting First-Year Student Performance?

In this experiment, we collected a large dataset from the Faculty of Science at UBU for 10 academic years from 2011-2021. Six different data classification algorithms were used on the dataset to compare and find the best classifier to predict student performance. Furthermore, to obtain more statistical results, we also tested and compared all possible combinations of hyperparameter configurations in 138 prevailing patterns. Then, the final model with the best performance was selected for highly accurate predictions. This issue is highly related to (Ali et al., 2019, 2023; Assami et al., 2022) because they support the idea that the performance of modern machine learning algorithms depends on their parameter setting.

Results in **Figure 7** show the implemented model and demonstrated that RF is the most accurate with a value of 85.84%, while DT is the second best and has the closest values at 85.72%, whereas NB showed the worst performance at 74.42%. This is highly related to the results in Assami et al. (2022) and Sathe and Adamuthe (2021) that RF could predict student performance with the highest accuracy.

Based on the result, it can be noted that the top four classifiers such as DT, MLP, RF, and SVM improved

Table 2. Performance of optimized predictive model from 10-fold cross-validation

Algorithm	Accuracy	Precision	Recall	F1-measure
RF	**85.84**	**80.94**	**93.76**	**86.87**
DT	84.61	79.74	92.74	85.72
MLP	84.09	81.19	88.90	84.71
SVM	82.92	79.45	88.65	83.75
LR	79.74	77.07	83.17	79.96
NB	74.42	78.75	66.77	72.08

Table 3. Optimized hyperparameter values for each machine learning algorithm

C	Best optimal parameter
RF	**Split function=entropy, number of trees=50, max depth=none, max features='sqrt', & n_estimators=300**
DT	Criterion=entropy & max_depth=3
LR	L2_penalty, maximum likelihood function, & p=0.1
SVM	Radial basis function kernel, tolerance=1e-3, regularization=1, & degree=3
MLP	Hidden layer=(10, 10, 10, 10), learning rate=constant, activation=logistic, & max_iter=100
NB	Gaussian NB & var_smoothing=1e-9

Note. C: Classifier

their accuracy by more than 80%, while LR and NB are less than satisfactory, but LR is much closer to 80%. This proved that student performance could be predicted with acceptable accuracy by four algorithms applied to the admission dataset.

Another interesting observation of the result is that RF is the best model in terms of accuracy, recall, and F1-measure, but it is not the best in precision because MLP model could perform better at 81.19% by comparison. Using these values, DT performs similarly to RF, while outperforming MLP and SVM, which recorded an accuracy of 84.09% and 82.92%, respectively. The details of each algorithm are compared in **Table 2**.

Details of suitable parameter values for testing the parameters that can produce the highest accuracy from a series of tests have been concluded, as described in **Table 3**. In **Table 3**, among these variables, it can be seen that the criterion parameter of the RF model was most accurate, when set to "entropy" and consisted of 50 DTs, while the efficiency of n_estimators was found to be 300 with a maximum depth equal to none. Interestingly, this was found to be significantly related to Ali et al. (2023), Jayaprakash et al. (2020), and Sathe and Adamuthe (2021) implying that setting up the number of trees or n_estimators in this range can reduce the error rate in RF.

In the next section, we used fine-tuned RF, which is an effective algorithm to construct a predictive model for 10 programs in the Faculty of Science at UBU and to identify the factors affecting student performance at the program level by feature importance analysis.

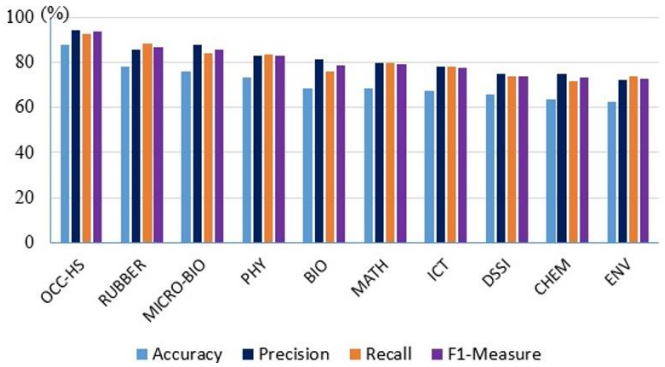

Figure 8. Program accuracy results using RF with parameter tuning (Source: Authors' own elaboration)

Table 4. Performance of fine-tuned RF for 10 programs in the Faculty of Science at UBU

FTPM	Accuracy	Precision	Recall	F1-measure
OCC-HS	88.05	94.40	92.70	92.70
RUBBER	77.89	85.83	88.40	88.40
MICRO-BIO	76.11	87.60	84.19	84.19
PHY	73.18	83.12	83.70	83.70
BIO	68.48	81.50	76.18	76.18
MATH	68.40	79.70	79.58	79.58
ICT	67.12	78.01	78.00	78.00
DSSI	65.56	75.09	73.85	73.85
CHEM	63.80	74.92	71.90	71.90
ENV	62.73	72.20	74.00	74.00

Note. FTPM: Fine-tuned predictive model

RQ2. How Was the Best Classifier With Fine-Tuned Parameters Used to Enhance and Predict Student Performance for Each Program at the Faculty of Science at UBU?

To answer this question, the same experiment as mentioned above was followed here, we next considered and constructed 10 predictive models based on the program in the Faculty of Science at UBU using fine-tuned RF models that produce the highest rates. In **Figure 8**, the result showed that values for each model were able to predict student performance with F1-measure ranging from 72% to 93%. That is, the predictive model of OCC-HS provided the highest F1-measure with a level of 92.70%, while the lowest value was achieved by the ENV model, which was 71.90%.

Remarkably, it has been observed that four prediction models (e.g., MICRO-BIO, OCC-HS, PHY, and RUBBER) achieved improved F1-measure values marked above 80%, and two prediction models like ICT and MATH have a similar F1-measure that obtained a value of 78-79%, which may not be sufficient but improved closer to 80%. While the remaining model explains low-value marks from 71% to 76%.

However, more detailed results also show in **Table 4** that the accuracy score of some programs was less than 75%, such as CHEM, DSSI, and ENV. This is due to a large number of datasets with missing values and

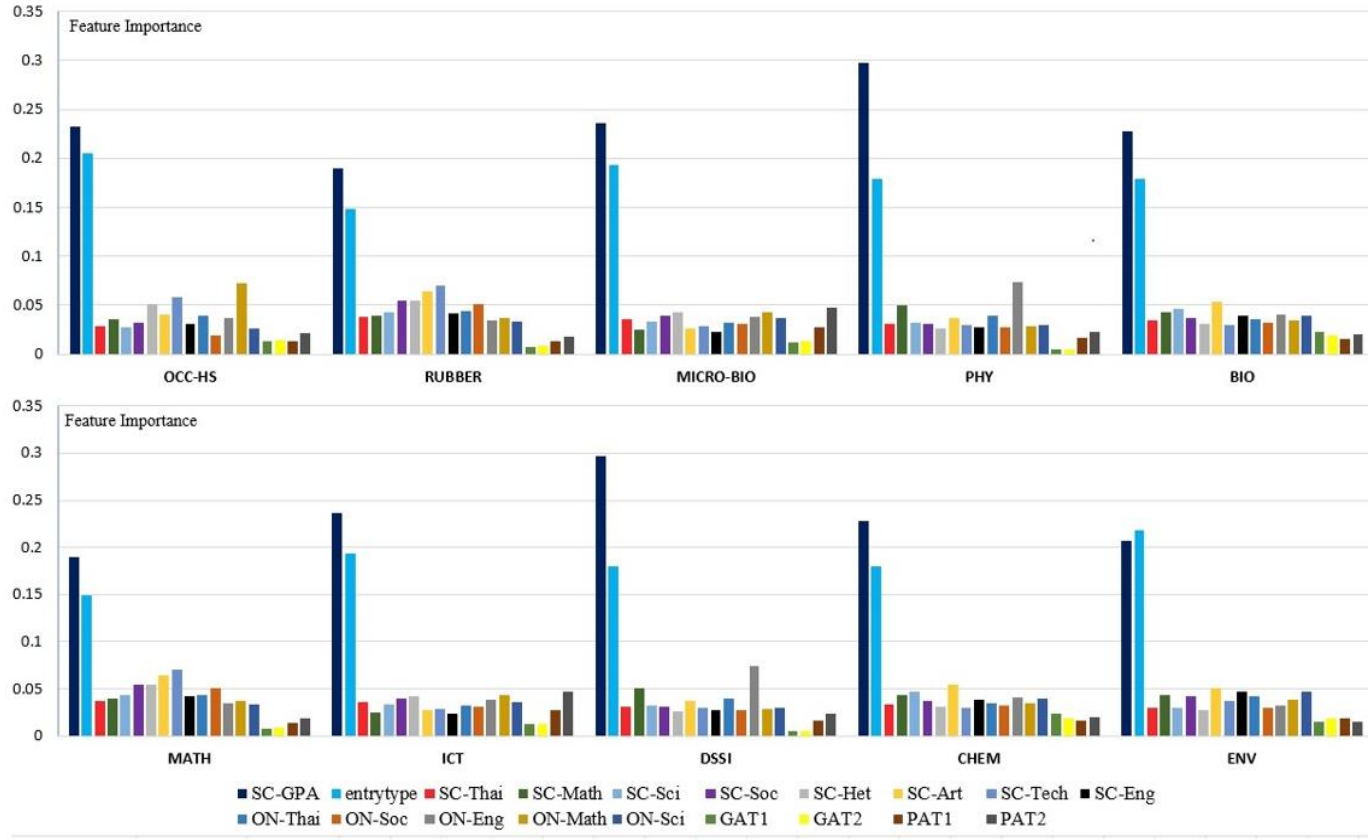

Figure 9. Feature importance of 10 programs (Source: Authors' own elaboration)

outliers that led to many dropping rows of data that could impact data distribution. On the other hand, the predictive model for MICRO-BIO, PHY, and RUBBER works better with an accuracy of more than 80%, especially for OCC-HS, which reaches 92.70% because there are fewer missing values and outliers to drop, and the data keeps the original distribution as well.

RQ3. How Can the Best Classifier With Fine-Tuned Parameters Be Used to Identify the Important Attributes That Affect Student Performance in Each Program at the Faculty of Science at UBU?

We used the fine-tuned RF algorithm to perform different kinds of program experiments with the same set of attributes to analyze and demonstrate the impact on key variables of each program for predicting student success. The features of each program are shown in **Figure 9**.

Figure 9 reveals that school grades are the most influential feature in predicting the future and performance of students of all programs except ENV, but GPA is still close to the entry type. While national test scores (e.g., ONET-English, ONET-Math, ONET-Science, ONET-Social studies, ONET-Thai, and PAT2) and grades of school subjects (e.g., art, English, health, math, science, and technology) are also important keys. This result related to Cui et al. (2021), Jayaprakash et al. (2020), Gutierrez et al. (2022), and Holladay et al. (2020), which showed that high school GPA, entrance exam scores, and school subjects in science and technology were the most important attributes for predicting student performance.

According to the results, we looked for those attributes in the next step and found correlations among programs, fields of science, and the most important attribute, depicted in **Table 5**. Through this analysis, the results can be summarized, as follows:

- **Pure Science:** A student who studies in this group and has a high school GPA, and good grades in interesting school subjects like art, health, math, science, and technology is likely to have high academic success. Other national test elements affect student performance such as ONET-English, ONET-Math, ONET-Science, and PAT2.
- **Applied Science:** A student who studies in this group and has good high school grades, good national scores for science, and good school subject scores for English and technology will likely pass the exam. Especially for social subjects, if students had good scores on both school and national tests, they would likely pass. Interestingly, in the case of the computer field (e.g., DSSI and ICT), it is obvious that the most important features are quite the same.
- **Health Science:** A student who studies in this group and has high school grades and good math scores from national and school tests, clearly

Table 5. Most important factors in each program using fine-tuned RF

Field	Program	Most important attributes & their feature importance				
		Rank#1	Rank#2	Rank#3	Rank#4	Rank#5
Pure Science	MATH	entrytype	SC-GPA	SC-Art	ON-Sci	SC-Eng
		0.21848	0.20657	0.05035	0.04742	0.04683
	MICRO-BIO	SC-GPA	entrytype	PAT2	ON-Math	SC-Health
		0.23624	0.19332	0.04701	0.04311	0.04219
	BIO	SC-GPA	entrytype	SC-Art	SC-Sci	SC-Math
		0.22772	0.17925	0.05388	0.04663	0.04274
	PHY	SC-GPA	entrytype	ON-Eng	SC-Math	ON-Thai
		0.29722	0.17946	0.07357	0.05038	0.03923
	CHEM	SC-GPA	entrytype	SC-Math	SC-Art	SC-Sci
		0.25357	0.15465	0.04839	0.04685	0.03963
Applied Science	ICT	SC-GPA	entrytype	ON-Sci	SC-Soc	SC-Eng
		0.31408	0.20992	0.04063	0.03769	0.03587
	DSSI	SC-GPA	entrytype	ON-Soc	ON-Sci	SC-Eng
		0.19593	0.19058	0.05819	0.04613	0.04521
	RUBBER	SC-GPA	entrytype	SC-Tech	SC-Art	SC-Health
		0.19000	0.14874	0.07025	0.06441	0.05438
Health Science	ENV	SC-GPA	entrytype	ON-Math	ON-Thai	SC-Math
		0.25947	0.15712	0.05193	0.04854	0.04549
	OCC-HS	SC-GPA	entrytype	ON-Math	SC-Tech	SC-Health
		0.23227	0.20482	0.07257	0.05761	0.05094

reveals that they will succeed at a university. Also, other features such as school subjects for health, technology, and Thai could impact success.

As a result, it is the fact that school GPA and national test scores are not the only important factors in assessing student performance, but also the grades for school subjects are other deciding factors for identifying student success. We found that the subject areas featured with an impact on prediction, were English, math, and science at the school and national test levels. In addition, after narrowing this down, we often see students, who are from the same field of science performing well in a similar subject, which leads to their proficient performance in the examination.

In summary, these findings show that it not only can be beneficial in predicting and improving student performance using RF with finely-tuned parameters, but also offers insight into the impact of a feature on student success for each program and field of science. As a result, we found that scores from both national tests and school subjects like GPA, entry type, math, and science, are important for the desired student performance.

Moreover, entry type and other features (e.g., art, health, PAT2, and Thai) also play a significant role. Hence, our proposed methodology and its results can be designed and implemented for application in other relevant education institutes. However, the effectiveness of such a system is possibly limited due to data sparsity, missing data, outliers, and some unobservable features. These might have a certain different influence on the results.

CONCLUSIONS & FUTURE WORK

This study aimed to predict new student performance using machine learning techniques based on large student admission data of 5,919 records collected from 2011-2021 at the Faculty of Science at UBU, Thailand. In our dataset, we emphasized considering the score data of students consisting of 19 predictor variables, which are based on Thailand's core curriculum and the university's recruitment system.

In this study, we carried out a two-part important task to address the research questions. First, six classification models were focused on exploring the performance that is most accurate for predicting student success. Especially, with greater accuracy, the hyperparameter tuning process also performed well for 138 possible patterns based on their models to obtain a higher result.

Second, the task of identifying significant predictor data that may affect student success at different programs in our faculty was also explored. The overall tasks discussed above make our study to be different from previous studies. The methodology was conducted using Python 3, Jupyter Notebook, and Scikit-Learn to build the prediction models. To evaluate the system, experimental data were divided into two parts:

First, we assess and compare six predictive models by using confusion matrix with 10-fold cross-validation. The results showed that fine-tuned RF had the highest accuracy value at 86.87%. As a consequence, when we next implemented that fine-tuned RF model at the program level, it was accurate to about 72% to 93%, while the four predictive models for MICRO-BIO, OCC-HS, PHY, and RUBBER performed well, where the

greatest accuracy ranged from 84% to 93%. However, the remaining model would probably receive lower values due to missing data and outliers.

Second, to determine features that play an important role in identifying student performance, the feature importance value of each attribute made by the best model was computed. As a result, it is suggested that entry type and national test scores (e.g., ONET-English, ONET-Math, ONET-Science, ONET-Social studies, ONET-Thai, and PAT2,) are not the only important factors, but also the school GPA and school subject grades (e.g., art, English, health, math, science, and technology) are other deciding factors for identifying the student's success. Interestingly, it was observed that many predictors for each program in the same field have mostly similar values. For example, for the program DSSI and ICT in applied science, it was found that English, science, and social studies were identical predictors of student success.

To that end, this study contributes to the analysis and prediction of student performance and also identifies significant features that make for student success in the first year of studies. Additionally, the experimental results are useful for other relevant educational institutions in that they could effectively foster student retention, reduce early dropouts and failing rate problems, plan teaching strategies, determine criteria for future admissions, and prepare appropriate readiness courses.

In future work, we plan to apply Deep Learning and additional feature extraction techniques for the classification process to enhance performance results. Additional features present in student demography, that may lead to student success, will be considered.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: This study was supported by Ubon Ratchathani University, THAILAND.

Ethical statement: The authors stated that the study was reviewed and approved by the Human Research Ethics Committee of Ubon Ratchathani University, which was valid from 8 December 2022 to 1 December 2023. Written informed consents were obtained from the participants.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Adekitan, A. I., & Noma-Osaghae, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technology, 24*, 1527–1543. https://doi.org/10.1007/s10639-018-9839-7

Ali, H., Mohd Salleh, M. N. B., Saedudin, R. R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal Of Electrical Engineering and Computer Science, 14*(3), 1560-1571. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Ali, Y. A., Awwad, E. M., Al-Razgan, M., & Maarouf, A. (2023). Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes, 11*(2), 349. https://doi.org/10.3390/pr11020349

Assami, S., Daoudi, N., & Ajhoun, R. (2022). Implementation of a machine learning-based MOOC recommender system using learner motivation prediction. *International Journal of Engineering Pedagogy, 12*(5), 68-85. https://doi.org/10.3991/ijep.v12i5.30523

Backham, N. B., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2023). Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science, 216*, 597-603. https://doi.org/10.1016/j.procs.2022.12.174

Bengesai, A. V., & Pocock, J. (2021). Patterns of persistence among engineering students at a south African university: A decision tree analysis. *South African Journal of Science, 117*(3/4). https://doi.org/10.17159/sajs.2021/7712

Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports, 12*, 12508. https://doi.org/10.1038/s41598-022-15880-6

Chang, T.-C., & Wang, H. (2016). A multi criteria group decision-making model for teacher evaluation in higher education based on cloud model and decision tree. *EURASIA Journal of Mathematics, Science and Technology Education, 12*(5), 1243-1262. https://doi.org/10.12973/eurasia.2016.1510a

Cui, J., Zhang, Y., An, R., Yun, Y., Dai, H., & Shang, X. (2021). Identifying key features in student grade prediction. In *Proceedings of the International Conference on Progress in Informatics and Computing* (pp. 519-523). IEEE. https://doi.org/10.1109/PIC53636.2021.9687042

Dabaliz, A.-A., Kaadan, S., Dabbagh, M. M., Barakat, A., Shareef, M. A., Al-Tannir, M., Obeidat, A., & Mohamed, A. (2017). Predictive validity of pre-admission assessments on medical student performance. *International Journal of Medical Education, 8*, 408-413. https://doi.org/10.5116/ijme.5a10.04e1

Devi, K., & Ratnoo, S. (2022). Predicting student dropouts using random forest. *Journal of Statistics and Management Systems, 25*(7), 1579-1590. https://doi.org/10.1080/09720510.2022.2130570

Gutierrez, O. A., Taylor, D. M. H., Santos-Guevara, A., Chavarria-Garza, W. X., Martinez-Huerta, H., & Galloway, R. K. (2022). How the entry profiles and early study habits are related to first-year academic performance in engineering programs. *Sustainability, 14*(22), 15400. https://doi.org/10.3390/su142215400

Holladay, S. D., Gogal, R. M., Moore, P. C., Tuckfield, R. C., Burgess, B. A., & Brown, S. A. (2020). Predictive value of veterinary student application data for class rank at end of year 1. *Veterinary Sciences, 7*(3), 120-132. https://doi.org/10.3390/vetsci7030120

Huynh-Cam, T.-T., Chen, L.-S., & Huynh, K.-V. (2022). Learning performance of international students and students with disabilities: Early prediction and feature selection through educational data mining. *Big Data and Cognitive Computing, 6*(3), 94. https://doi.org/10.3390/bdcc6030094

Huynh-Cam, T.-T., Chen, L.-S., & Le, H. (2021). Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. *Algorithms, 14*(11), 318. https://doi.org/10.3390/a14110318

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting students academic performance using an improved random forest classifier. In *Proceedings of the International Conference on Emerging Smart Computing and Informatics* (pp. 238-243). IEEE. https://doi.org/10.1109/ESCI48226.2020.9167547

Kaensar, C., & Wongnin, W. (2023). Analysis and prediction of student performance based on Moodle log data using machine learning techniques. *International Journal of Emerging Technologies in Learning, 18*(10), 184-203. https://doi.org/10.3991/ijet.v18i10.35841

Kemda, L. E., & Murray, M. (2021). Statistical modeling of students' academic performances: A longitudinal study. *International Journal of Higher Education, 10*(6), 153-170. https://doi.org/10.5430/ijhe.v10n6p153

Ko, C.-Y., & Leu, F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education, 64*(1), 50-57. https://doi.org/10.1109/TE.2020.3004596

Kornpitack, P., & Sawmong, S. (2022). Empirical analysis of factors influencing student satisfaction with online learning systems during the COVID-19 pandemic in Thailand. *Heliyon, 8*(3), e09183. https://doi.org/10.1016/j.heliyon.2022.e09183

Maksimova, N., Pentel, A., & Dunajeva, O. (2022). Computer science students early drop-out prediction using machine learning: A case study. In M. E. Auer, A. Pester, & D. May (Eds.), *Learning with technologies and technologies in learning* (pp. 523-549). Springer. https://doi.org/10.1007/978-3-031-04286-7_25

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905

Mothial, R. K., De Laet, T., Broos, T., & Pinxten, M. (2018). Predicting first-year engineering student success: From traditional statistics to machine learning. In *Proceedings of the 46th SEFI Annual Conference*. The European Society for Engineering Education.

Nurhachita, N., & Negara, E. S. (2021). A comparison between deep learning, naïve Bayes and random forest for the application of data mining on the admission of new students. *International Journal of Artificial Intelligence, 10*(2), 324-341. https://doi.org/10.11591/ijai.v10.i2.pp324-331

Qahmash, A., Ahmad, N., & Algarni, A. (2023). Investigating students' pre-university admission requirements and their correlation with academic performance for medical students: An educational data mining approach. *Brain Sciences, 13*(3), 456-465. https://doi.org/10.3390/brainsci13030456

Rajagopal, S. K. P. (2020). Predicting student university admission using logistic regression. *European Journal of Computer Science and Information Technology, 8*(3), 46-56.

Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

Roslan, M. H. B., & Chen, C. J. (2020). Educational data mining for student performance prediction: A systematic literature review (2015-2021). *International Journal of Emerging Technologies in Learning, 17*(05), 147-179. https://doi.org/10.3991/ijet.v17i05.27685

Santosa, R. G., Lukito, Y., & Chrismanto, A. R. (2021). Classification and prediction of students' GPA using k-means clustering algorithm to assist student admission process. *Journal of Information Systems Engineering and Business Intelligence, 7*(1), 1-10. https://doi.org/10.20473/jisebi.7.1.1-10

Sathe, M., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of sudents' performance. *International Journal of Modern Education and Computer Science, 13*(1), 1-21. https://doi.org/10.5815/ijmecs.2021.01.01

Singh, M., Verma, C., Kumar, R., & Juneja, P. (2020). Towards enthusiasm prediction of Portuguese school's students towards higher education in realtime. In *Proceedings of the International Conference on Computation, Automation and Knowledge Management* (pp. 421-425). IEEE.

https://doi.org/10.1109/ICCAKM46823.2020.9051459

Ubon Ratchathani University. (2010). *REG UBU system: Office of registration*. https://reg.ubu.ac.th

Usman, M., Iqbal, M. M., Iqbal, Z., Chaudhry, M. U., Farhan, M., & Ashraf, M. (2017). E-assessment and computer-aided prediction methodology for student admission test score. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(8), 5499-5517. https://doi.org/10.12973/eurasia.2017.00939a

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, 11. https://doi.org/10.1186/s40561-022-00192-z

Yang, J., Jiang, H., Wang, J., & Luo, H. (2022). Key factors influencing blended learning outcomes in an undergraduate course: Perspectives from learning behaviors and experiences. In *Proceedings of the 4th International Conference on Computer Science and Technologies in Education* (pp. 123-127). IEEE. https://doi.org/10.1109/CSTE55932.2022.00029