Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07997-6

ORIGINAL ARTICLE

Academic performance warning system based on data driven for higher education

Hanh Thi-Hong Duong[1,2] · Linh Thi-My Tran[1,2] · Huy Quoc To[1,2] · Kiet Van Nguyen[1,2]

Received: 21 January 2022 / Accepted: 25 October 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
Academic probation at universities has become a matter of pressing concern in recent years, as many students face severe consequences of academic probation. We carried out research to find solutions to decrease the situation mentioned above. Our research used the power of massive data sources from the education sector and the modernity of machine learning techniques to build an academic warning system. Our system is based on academic performance that directly reflects students' academic probation status at the university. Through the research process, we provided a dataset that has been extracted and developed from raw data sources, including a wealth of information about students, subjects, and scores. We build a dataset with many features that are extremely useful in predicting students' academic warning status via feature generation techniques and feature selection strategies. Remarkably, the dataset contributed is flexible and scalable because we provided detailed calculation formulas that its materials are found in any university or college in Vietnam. That allows any university to reuse or reconstruct another similar dataset based on their raw academic database. Moreover, we variously combined data, unbalanced data handling techniques, model selection techniques, and research to propose suitable machine learning algorithms to build the best possible warning system. As a result, a two-stage academic performance warning system for higher education was proposed, with the F2-score measure of more than 74% at the beginning of the semester using the algorithm Support Vector Machine and more than 92% before the final examination using the algorithm LightGBM.

Keywords Academic performance · Warning system · Machine learning · Data driven · Feature selection · Feature generation · Imbalanced data

1 Introduction

Faced with an ever-increasing influx of big data, the remarkable advancement of techniques from machine learning, artificial intelligence, neural networks, database systems, and other related fields has enabled people to extract valuable values from data. It results in a more modern and all-encompassing way of life in all aspects through extracted information and values. In this paper, we are especially curious about opportunities and social values that can be derived from data in education, notably higher education.

Today, many student data are stored in educational databases, giving universities a solid basis for promoting and changing development. With a competitive operating environment, a data-driven approach is widely adopted in universities worldwide, which helps them understand their

✉ Kiet Van Nguyen
kietnv@uit.edu.vn

Hanh Thi-Hong Duong
18520711@gm.uit.edu.vn

Linh Thi-My Tran
18520999@gm.uit.edu.vn

Huy Quoc To
huytq@uit.edu.vn

[1] Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

Published online: 07 November 2022

students better and bring support on many fronts. However, this problem is not common in Vietnam. Based on that, we conducted this research intending to be able to take advantage of data sources from education to solve a persistent problem in Vietnamese universities: academic probation.

Academic probation is the most general term used in the undergraduate context to signify that a student is not making the academic standard required by the institution for graduation. Furthermore, reviewing this matter is a periodic activity specified in the academic regulations to ensure each university's quality of teaching and learning. Each student's academic probation status will be determined at the end of each semester based on several disciplinary and academic factors, with academic performance being one of the most important.

In recent years, the academic probation of students has been becoming alarmingly common in Vietnam. According to university statistics, the number of students on academic probation reaches hundreds or even thousands each year. Furthermore, this dramatically affects the quality of training, the school's output standards, and students' academic activities. In particular, it can have serious academic consequences for students, such as limiting the number of credits enrolled, losing the opportunity to pursue their preferred major, or even being dismissed.

Based on the above, we decided to investigate and research this topic to assist universities in predicting the student's academic probation status, initially based on their academic performance. This research was developed to give students sound and timely warnings about their academic probation possibility in the new semester. Receiving warnings will make students more concerned about their current learning level to adjust their learning attitudes more appropriately, strive to improve their grades, and avoid academic probation. On the other hand, the academic advisor also partially understands the student's academic probation risk, allowing them to provide more appropriate and timely assistance to students. The warning system will be heavily dependent on the environment of learning and academics accordingly because it primarily uses students' academic results as inputs. In fact, input features can be easily collected in both online and offline learning programs, allowing the warning system to be used in almost all situations, even in the COVID-19 case—an emerging situation.

In this paper, we conduct research with data provided by a university in Vietnam with policies on academic handling specified in the academic regulations of this university. Specifically, academic probation in Vietnam is divided into three main types of status: academic suspension, academic warning, and dismissal. However, with the goal of building a warning system based only on students' academic performance, the models we trained will only support the classification of two states, including Academic Warning and Dismissal, because Academic Suspension is not affected by academic performance criteria. Rules related to academic performance-based warning are detailed in Sect. 3.1.2.

Our entire research process will be detailed in the sections below. Section 1 focuses on introducing the topic's purpose. Section 2 contains information about related research. In Sect. 3, we show how to create the dataset and information about the dataset used, and in Sect. 4, we give a thorough explanation of the tasks posed. Section 5 of the paper describes our approaches to the problem. In succession, the details of our experiments, results, and evaluation will be presented in Sect. 6. Finally, Sect. 7 contains the conclusions and future development directions we discovered during our research on this topic.

2 Related work

Academic warnings for students enable the university to respond to students' learning problems as soon as possible. Advisors can adopt various guiding measures to assist students in improving their academic performance or prevent delayed graduation. As a result, more and more academics recognize the immense societal potential of educational data and conduct research in this area.

Firstly, we cannot fail to mention the research work of Migueis et al. [1] in 2018. The study collected data from 2459 students to build models to predict students' academic performance. They conducted the research with Machine Learning algorithms such as Support Vector Machines, Decision Trees, Random Forest, and Naive Bayes. Finally, the model proposed by the author is Random Forest, with an accuracy of over 95% at an early stage of the student's academic path. This study demonstrates that previous semester grades are heavily weighted while building the model.

In 2021, Mingyu et al. [2] used groupings of features concerning students' studies, living, internet activities, and basic information to take the above problem. Prediction models are built from many machine learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting Decision Tree, XGBoost, LightGBM, etc. Furthermore, the best-obtained result is the Catboost–SHAP method with R-Squared (R2), Mean Squared Error (MSE), and Mean Absolute Error (MAE) of 80.3%, 24.976 and 3.551, respectively, with tenfold cross-validation. Besides providing a method to assist in detecting students with academic performance problems, this study also provides the ability to analyze further the influence of dependent features on students' academic performance. In addition to experimenting with various methods for predicting final students' grades in first-semester courses, Bujang et al. [3] resolve imbalanced data challenges to better-improved performance. The Synthetic Minority Oversampling Technique (SMOTE) was used to handle this problem. The results were surprising, with Random Forest producing the highest F-measure of 99.5%. The research has demonstrated the possibility of using classification algorithms to resolve the academic warning problem, with Random Forest and Support Vector Machine showing their advantages in this domain.

In 2022, Hamim et al. [4] classified students' learning outcomes based on many different aspects: personal behavior, social behavior, study habits, families, majors, and grades. Research focuses on experiments and comparing many boosting algorithms such as Adaptive Boosting, Gradient Boosting, Extreme Gradient Boosting, LightGBM, and CatBoost to find the most effective algorithm for this problem. In addition, they also applied feature selection techniques to improve model performance. Experimental results demonstrate that LightGBM is the most powerful algorithm with an accuracy of 89.26%, with 14 features selected through the feature selection technique; those features are mainly related to students' learning behavior and grades.

In addition to applying modern models and techniques in predicting academic performance, many studies related to the raw set approaches are also widely used. In 2014, the work of Namdeo et al. [5], Rough Set Theory (RST), was used in feature selection. RST was used to reduce the number of attributes from 12 to 3. In 2021, Madeira et al. [6] also applied RST in predicting student performance problems. After applying RST, the training cost, as well as the prediction efficiency, increased significantly, and the number of layers of the neural network decreased from 5 to 3, the number of features from 10 to 5, and the results also improved better.

We can see from the mentioned studies that the research field of academic warning for students is not uncommon worldwide. To the best of our knowledge, there are few published articles about student academic warnings have been issued in Vietnam. Realizing the potential and social values that the study on this topic brings, we have carried out this study with the data collected in Vietnam with the desire to improve somewhat the number of students being disciplined in the present.

3 Dataset

3.1 Data transformation

We look at how to take advantage of students' learning grades at university. Learning grade is the most intuitive and explicit reflection of the student's learning situation and is also a critical factor in the school's academic warning decision-making process. However, because the raw data about student grades contain many features, we decided to transform existing features with similar characteristics into new ones. Section 3.1.1 describes the formula for calculating new features, and our system only uses new features to predict.

To build a system that can easily apply in Vietnam's universities, we based on the current education system in Vietnam. Usually, a learning year is divided into two semesters. Students will register for the semester's subjects before the semester begins. With each subject, we need to be interested in four types of grades, and we call them component grades: process grade (student's attendance), midterm grade (midterm exam), practice grade (practice exam), and final grade (final exam). Each component grade will be assigned a weight, and the sum of the weights of the four types of grades is equal to 1. The GPA for each subject is computed as the total of the component grades multiplied by their weights. Within component grades, because midterm and final grades hold most of the weight, they have the role of determining the majority of the high or low subject's GPA. Finally, the semester's GPA will be computed based on the GPA of all the subjects students complete in the semester, with the formula:

$$semGPA = \frac{\sum_{1}^{n} subGPA_i.subCredit_i}{\sum_{1}^{i} subCredit_i} \tag{1}$$

where,

- $semGPA$: the semester's GPA.
- n: the number of subjects in the semester.
- $subGPA_i$: the GPA of i-th subject in the semester.
- $subCredit_i$: the number of credits of i-th subject in the semester.

The GPA in this paper is GPA out of 10, specifically as follows: 9.0–10.0 (A+), 8.0–9.0 (A), 7.0–8.0 (B+), 6.0–7.0 (B), 5.0–6.0 (C), 4.0–5.0 (D+), 3.0–4.0 (D), <3 (F). And component grades also use this scale. It is possible to see how to convert a score out of 10 to another grading system in Table 9 in the Appendix.

We must ensure data quality while adhering to academic warning regulations because data is crucial in building our systems. The dataset was created in three stages: calculating the necessary features, annotating the dataset, and evaluating and reviewing the dataset.

3.1.1 Calculating the required features

We used a raw database provided by a prestigious Vietnamese university to build a new dataset to aid in the research process. The raw database contains information about student learning results as well as detailed subject information (e.g., subject name, subject code, number of credits, component coefficients, previous subjects, prerequisite subjects) and information about student learning results (e.g., student identification code, subject name, component grades). The new features were calculated using the following formula:

Feature pre_avg: Some subjects encourage students to have fundamental or related knowledge before enrolling in the training program. With previous subject data (including the identification of the subject and previous subjects of the subject) we can compute the average grade of all subjects considered the subject's previous subject in which the student enrolled in the current semester, which we call pre_avg. For example, student A registered two Course 334, in which Course 206 and Course 334 have a previous subject are Course 005 and Course 101. Suppose student A completed previous subjects. The pre_avg is the average GPA of Course 005 and Course 101.

Group 1: The list of features contains the GPA of each semester that the student completed (e.g., student A has completed six semesters, then A has six features in Group 1 is s1, s2, s3, s4, s5,and s6)

$$s(m) = \frac{\sum_1^n score_{m,i}.credit_{m,i}}{\sum_1^n credit_{m,i}} \tag{2}$$

Group 2: Each subject has corresponding component grades. Before the final exam, students know their process grades, practice grades, and midterm grades. Group 2 contains the average grade of process grades (avg1), the average grade of practice grades (avg2), and the average grade of midterm grades (avg3) of all subjects in the current semester.

$$avg(j) = \frac{\sum_1^n score(j)_{p,i}.coef(j)_{p,i}.credit_{p,i}}{\sum_1^n credit_{p,i}} \tag{3}$$

Group 3: Each component grade has a corresponding weight. Group 3 contains the average weight of process grades (coef1), the average weight of practice grades (coef2), and the average weight of midterm grades (coef3) of all subjects in the current semester.

$$coef(j) = \frac{\sum_1^n coef(j)_{p,i}.credit_{p,i}}{\sum_1^n credit_{p,i}} \tag{4}$$

where,

- p: is representative of the student's current semester (p-th semester).
- m: is representative of the m-th semester studied previously, $0<m<p$.
- n: represents the number of subjects that students enrolled in the semester under consideration.
- $score_{m,i}$: is the GPA of subject i in the m-th semester.
- $credit_{m,i}$: is the number of credits of subject i in semester m-th.
- $score(j)_{p,i}$: is the j-th component grade for subject i in current semester.

  (j=1: process grade, j=2: practice grade, j=3: midterm grade).
- $credit_{p,i}$: is the number of credits of subject i in current semester.
- $coef(j)_{p,i}$: is the corresponding weight to $score(j)_{p,i}$.

3.1.2 Data annotation, evaluating and reviewing the dataset

Academic warnings based on previous semester performance were regulated in Vietnamese university policies. However, some universities have different standards for each warning status than others. As a result, we used the popular standard to annotate the dataset we used in our experiment. After that, the dataset was reviewed, and null values were removed to ensure high-quality data. Table 1 shows the labeling information, whereas Sect. 3.2 shows the dataset specifics.

3.2 Data generation

Following Sect. 3.1, we transformed data from 4383 students' information. Because each student has different completed semesters, we grouped students who have the same semesters into the small dataset, and models are separately trained in each small dataset. Moreover, to increase the model performance and obtain more data for the training procedure, we generated the data for each small dataset according to the approach shown in Fig. 1. We have a list of GPAs for completed semesters for individual students. The list of GPAs was divided into groups of GPAs that are adjacent. Each group was added to the corresponding small dataset.

Table 1 Academic probation rules base on academic performance

Condition	Warning status	Label
GPA in 2 continuous semesters <4.0 or GPA in the considered semester <3.0	Warning	1
GPA in 2 continuous semesters $=0.0$	Dismissal	2
Others	Normal	0

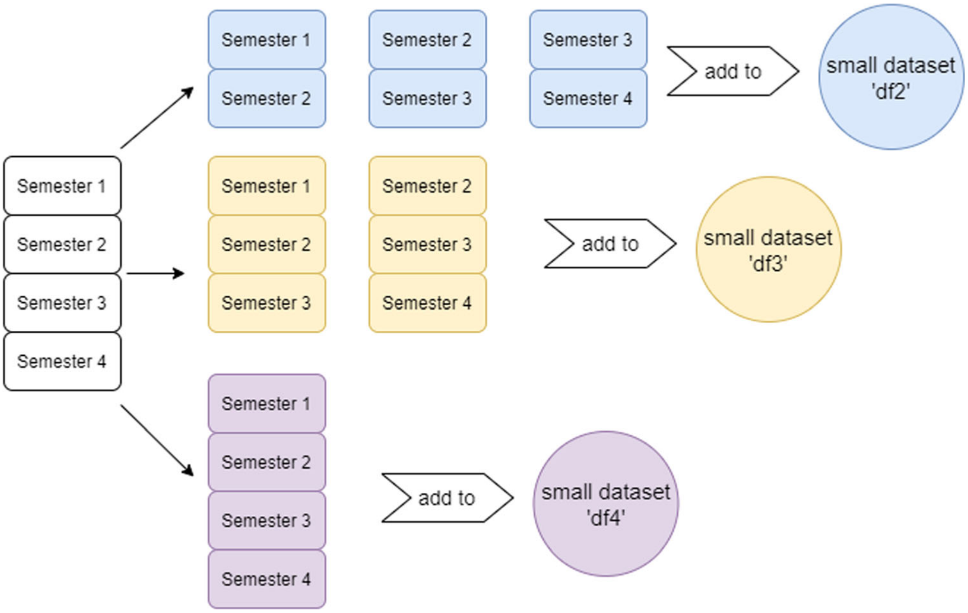


Fig. 1 Example for data generation of individual students

After the generation and transformation process, a group dataset was created. We call it as "Academic Performance Warning Dataset" (APWD). It consists of 9 individual small dataset named in the form df*i* for $i \in \{x \in \mathbb{N}; 2 \leq x \leq 10\}$. The index i is representative of the order of the current semester that the student is learning. Each small dataset is used to train and predict the warning status of the student with a different number of input features (e.g., df2 is used to train and predict the status of students who have completed the first semester and are in the second semester, whereas df3 is used to train and predict the status of students who have completed the second semester and are in the third semester). APWD is statistically and detailed description in Tables 2 and 3.

4 Task definition

The study's objective is to create a two-stage warning system that includes:

The first warning: the warning is issued at the beginning of the semester.

The second warning: the warning is issued before the final examination.

The first warning helps students understand their current academic status early on, allowing them to adjust their learning attitude accordingly right from the start of a new semester. Simultaneously, outcomes from the system also allow universities have appropriate, timely educational plans and strategies for students. With support from the warning system, both students and universities can be proactive in limiting academic probation early on.

Besides, we implemented one more warning before the final exam of each semester when further data on student performance during the semester under examination. That is expected to yield more accurate predictions than the warning at the beginning of the semester. Furthermore, the second warning helps universities identify more students whose academic performance has begun to decline during the semester being considered and provides them with a final wake-up call before they complete the semester.

4.1 Task 1: the first warning

Task 1 is described as follows:

Input: List GPA of previous semesters and feature pre_avg.

Output: A warning status (normal, warning, dismissal).

Table 4 shows three data points from three separate small datasets that represent three different sorts of warning statuses. The first data point is taken from df7 with six GPA values (s1, s2, s3, s4, s5, s6) and one feature pre_avg to be used as input to the warning system. Similarly, the following two data points, df4 and df6, have GPA values of 3 and 5, respectively.

4.2 Task 2: the second warning

Task 2 is described as follows:

Input: List GPA of previous semesters, list of values and weight of the component grade in the current semester.

Output: A warning status (normal, warning, dismissal).

In the second warning, students completed the midterm examination and the practice examination, and the teacher evaluated the student's process grades. For this reason, the

Table 2 Statistics on APWD

		Label 0	Label 1	Label 2	Total
df2	Train	23,588	1981	22	25,591
	Val	2948	248	3	3199
	Test	2949	248	2	3199
df3	Train	18,389	1475	21	19,885
	Val	2299	184	3	2486
	Test	2299	185	3	2487
df4	Train	14,329	1189	18	15,536
	Val	1791	149	2	1942
	Test	1792	149	2	1943
df5	Train	10,452	976	14	11,442
	Val	1306	122	2	1430
	Test	1307	122	2	1431
df6	Train	7654	754	12	8420
	Val	956	94	2	1052
	Test	957	95	1	1053
df7	Train	4965	554	11	5530
	Val	620	69	2	691
	Test	621	70	1	692
df8	Train	3021	417	9	3447
	Val	377	53	1	431
	Test	378	52	1	431
df9	Train	1172	258	6	1436
	Val	147	32	1	180
	Test	147	33	1	181
df10	Train	328	158	5	491
	Val	41	20	1	62
	Test	42	20	1	63

group features relevant to component grades (Group 2 and Group 3 in Sect. 3.1.1) was added to the data for the training model. Table 5 shows three data points from three separate small datasets. All three small datasets have the same number of features regarding component grades (avg1, avg2, avg3, coef1, coef2, coef3) and feature pre_avg but are different in the number of completed semesters (GPA of the previous semester).

5 Our approach

After data transformation and generation, we focus on researching preprocessing methods, algorithms, and evaluation techniques to propose a suitable approach for our system. We initially proposed techniques based on the tremendous potential of data characteristics, such as applying feature generation, feature selection techniques, handling imbalanced data, and dividing the model into smaller models. Furthermore, to evaluate the appropriate model and techniques, the selection of evaluation measures and visualization and analysis of results are also thoroughly explored. Figure 2 provides an overview of the approach we propose to the academic warning problem.

5.1 Preprocessing: handling missing data

The pre_avg and avg10 feature was detected as missing data with the number of missing records shown in Table 6.

We effort to preserve as much of the data for the model's training process by using the value-filling method to address the missing data issue. On a scale of 10, we have

Table 3 Feature description for APWD

Group feature	Feature	Description
Group 1	s1	The GPA of the 1st semester that the student completed
	s2	The GPA of the 2nd semester that the student completed
	s3	The GPA of the 3rd semester that the student completed
	s4	The GPA of the 4th semester that the student completed
	s(x)	The GPA of the (x)th semester that the student completed (df2: x2, df9: x = 9)
Group 2	avg1	The average grade of process grades of all subjects in the current semester
	avg2	The average grade of practice grades of all subjects in the current semester
	avg3	The average grade of midterm grades of all subjects in the current semester
Group 3	coef1	The average weight of process grades of all subjects in the current semester
	coef2	The average weight of practice grades of all subjects in the current semester
	coef3	The average weight of midterm grades of all subjects in the current semester
Other	pre_avg	The average grade of the subject's previous subject that the student enrolled in the current semester
	label	The warning status (0: normal, 1: warning, 2: dismissal)

Table 4 Several examples of features used in Task 1 for the first warning

Small dataset	s1	s2	s3	s4	s5	s6	pre_avg	Status
df7	7.54	8.11	7.55	7.91	8.64	8.62	7.77	Normal
df4	3.56	0.55	3.83	–	–	–	–	Warning
df6	5.92	5.95	4.71	2.10	0.00	–	4.93	Dismissal

Table 5 Several examples of features used for the second warning

Small dataset	s1	s2	pre_avg	avg1	avg2	avg3	coef1	coef2	coef3	Status
df3	7.54	8.11	–	0.99	0.75	1.71	0.11	0.09	0.25	Normal
df2	5.02	–	5.77	0.29	0.16	0.74	0.05	0.16	0.28	Warning
df3	5.35	1.10	0.00	0.00	–	–	0.01	0.00	0.17	Dismissal

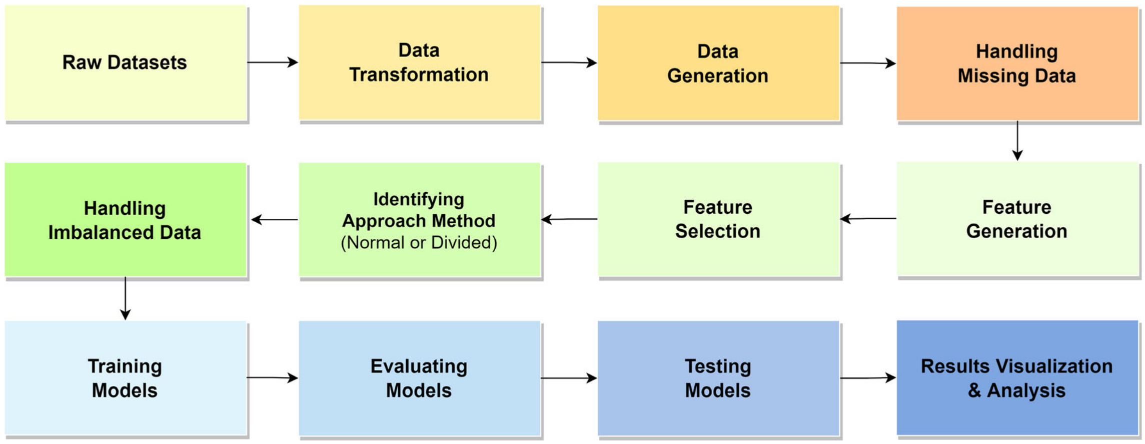


Fig. 2 The proposal approach process

chosen 5 as the value to use in that solution. That value was decided because we found the mean of the scale to be the most reasonable because selecting a value left or right based on the overall scale can cause undesired bias.

Table 6 Emptied data statistics table of features pre_avg and avg10

df	pre_avg			avg10		
	Train	Val	Test	Train	Val	Test
2	1928	251	248	644	85	87
3	1771	214	228	659	74	74
4	1066	130	129	606	70	72
5	1023	125	123	599	71	72
6	873	119	126	542	75	74
7	817	101	104	538	72	76
8	710	87	96	532	66	68
9	435	53	57	350	39	45
10	111	15	19	79	11	13

5.2 Proposal algorithms

5.2.1 LightGBM (LGBM)

Traditional machine learning methods may not be as efficient as they should be because they do not fully exploit the underlying characteristics and structure of the data, despite their considerable success in knowledge discovery. The ensemble model was born in this context and has been widely used with various variations, achieving promising knowledge discovery and predictive performance [7–10]. In addition to the basic approach of combining models trained on different algorithms for the final result, it is impossible not to mention ensemble models based on decision trees and boosting ensemble techniques [7, 9, 11]. A plethora of related studies confirms the excellent power of these models. LightGBM [12] is one of the algorithms based on an ensemble model, and it has proven its effectiveness in machine learning. It is implemented based on the Gradient Boosting Decision Tree [13]. It begins by constructing a tree to attempt to fit the data, and subsequent trees are constructed to minimize residuals using a loss function influenced by the Gradient Descent. It accomplishes this by concentrating on areas underperforming existing learners by applying the Boosting. However, it uses Gradient-based One Side Sampling and Exclusive Feature Bundling to significantly speeds up the computing process.

5.2.2 Support vector machine (SVM)

Besides the proposed modern algorithm LGBM, we also propose another algorithm to tackle the classification problem: the algorithm Support Vector Machine (SVM) [14]. SVM is a robust and commonly used algorithm in many fields. The power of this algorithm is, in many cases, comparable to the architectures of the deep learning approach. Specifically, in the study by Hasan et al. [15], SVM with kernel = 'rbf' has outperformed the CNN solution with an accuracy of 98.84%. The algorithm's fundamental idea is to execute a hyperplane discovery that splits the classes based on margin. For its excellent generalizability and powerful optimization ability in hyperplane search, the algorithm has piqued the interest of the scientific research community.

Moreover, SVM was also introduced as a golden technique in classification and regression tasks. Because of the above, we decided to propose this algorithm to solve our three-label classification problem. Specifically, we use a specialized algorithm for the classifier problem provided by SVM, which is C-Support Vector Classification (svm.SVC).

5.3 Feature generation (FG)

With the continuous development of research and automation, many complex problems appear. If only algorithms apply, many data characteristics are not fully exploited. In that context, many data techniques were born to improve performance and attract much attention from researchers. In feature engineering, feature generation is a remarkable technique and is applied in many machine learning, and deep learning problems [16–18]. On that basis, we have applied feature generation in the experiment to achieve good performance for the system. At the data preprocessing stage, the information gained from the data analysis helps us devise strategies to improve the performance of models. Based on feature characteristics and academic probation rules (Sect. 3.1.2), we created several new features and added them to input data in our training process.

avg_final sum average of component grades (group features 2) in the current semester.

coef_final sum average of component weights (group features 3) in the current semester.

avg10 is the converted component average grade calculated using avg_final divided by coef_final.

history is the number of semesters the student has been placed on probation in the past.

5.4 Feature selection (FS)

Through data transformation and feature generation, we obtained a large set of features for the training process. However, to find out the features that are useful in predicting the academic alert status of students, we have incorporated feature selection in our experiments. Many study areas dealing with machine learning problems have been interested in feature selection because [19–21] it allows classifiers to be faster, more cost-effective, and more accurate [22, 23]. In our approach, we calculate the correlation between the input features using the Pearson correlation coefficient and remove those highly correlated with the remaining features. Pearson correlation coefficient is defined as follows:

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}} \tag{5}$$

where,

- r: pearson correlation coefficient
- N: the number of pairs of features
- $\sum xy$: the sum of the products of paired features
- $\sum x$: the sum of x features
- $\sum y$: the sum of y features
- $\sum x^2$: the sum of squared x features
- $\sum y^2$: the sum of squared y features

The correlation between two features is stronger when the correlation between them is closer to -1 or 1.

5.5 Handling imbalanced data

The majority of machine learning algorithms perform their best when there is an equilibrium in the number of samples between classes. That is because most algorithms are designed to maximize accuracy to minimize errors. When imbalanced data is used for training a model, the model will be more biased toward the target class when making a prediction. As a result, the handling of imbalanced data is an issue of concern in many studies in the field of academic particular [2, 24, 25] and general fields [26, 27]. We apply one technique to handle the imbalanced data to deal with that issue: adjusting class weights.

When creating an algorithm instance, we adjust the class weights by assigning the value 'balanced' to the class_weight parameter. Class-weight is a built-in parameter that assists in optimizing scoring for the minority class. The model automatically takes class weights that are inversely proportional to their occurrence frequencies in the dataset. The weights of the classes will be determined according to the following formula (6):

$$w_j = \frac{n\_samples}{(n\_classes * n\_samples_j)} \tag{6}$$

where,

- w_j represents the weight of each class, and j denotes the class.
- $n\_samples$ is overall the number of samples that the dataset has.
- $n\_classes$ is the total number of class types in the dataset.
- $n\_samples_j$ is the number of samples of class j.

5.6 New approach: divided approach

Our academic warning system is supposed to give each student one of three warning statuses. Basically, the way we normally build a single 3-label model for prediction and we call the Normal approach. However, the normal status covers a broader range of situations than warning and dismissal status, and the disparity in quantity is an inconvenience for the model's learning. Therefore, we approached the problem with a new method to improve the predictability of the system. That new approach is named the Divided approach. In that approach, to complete the warning purpose in 3 statuses, we combined warning and dismissal into a new status called academic probation and built two 2-label classification models. Model 1 classifies normal status and academic probation. Model 2 classifies students placed on academic probation in model 1 into a warning state or dismissal state. Our divided approach is depicted in Fig. 3.

5.7 Performance evaluation

5.7.1 Measure

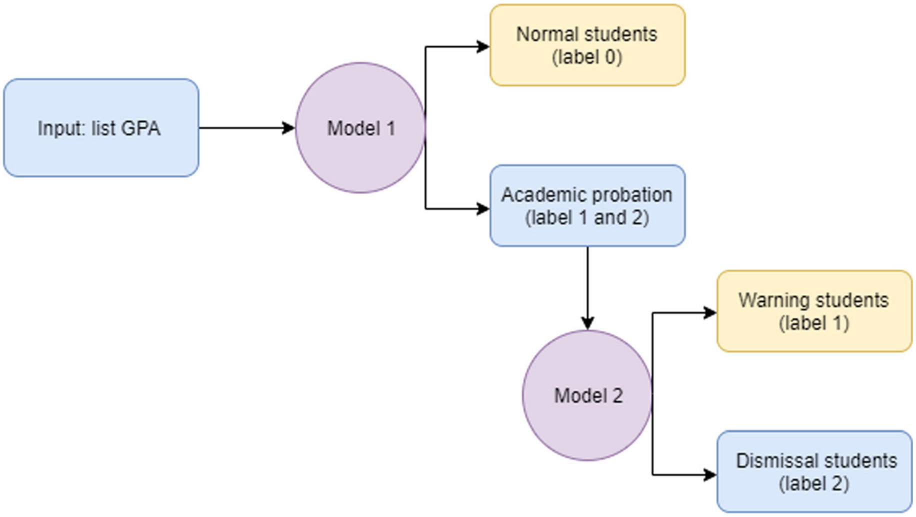
We used the F2-score as a critical measure in the evaluation phase. The F-score is an approach to combining the precision and recall of the model. It is typical for the task of evaluating many types of machine learning models. This measure is particularly effective when applied to unevenly distributed data or where the costs of false positives and false negatives differ, as in the case of predicting disease in medicine.

In machine learning evaluation tasks, the F1-score is a widely used metric, whereas the F2-score is less popular. Instead of providing a balanced measure of precision and recall, as the F1-score does, the F2-score focuses more on recall and less on precision. Specifically, it puts more emphasis on recall to minimize false negatives. The F1 score and the F2 score are mathematically defined by the formula of F_β (7) with β equal to the values 1 and 2, respectively.

$$F_\beta = (1+\beta^2)\frac{recall.precision}{recall + \beta^2.precision} \tag{7}$$

In our problem, false negatives represent students whose classification is academic probation (class 1, 2) predicted as normal students (class 0), and false positives represent normal students (class 0) predicted as students whose classification is academic probation (class 1, 2). Using the F2-score allows our model to increase the importance of cases of omission of students who should be warned while

Fig. 3 The divided approach


Input: list GPA
Model 1
Normal students (label 0)
Academic probation (label 1 and 2)
Model 2
Warning students (label 1)
Dismissal students (label 2)

decreasing the volume of false warnings. From there, models chosen based on this metric will be able to reduce the number of missed warnings significantly.

However, the F2-score is only used for binary classification evaluation. Therefore, before evaluating the three-label classification tasks, a transformation is required. Specifically, we considered labels 1 and 2 to be in the same class. To accomplish that, We converted the predicted results labeled 2 to label 1 to start the evaluation.

5.7.2 Strategy for evaluation

As discussed in Sect. 5.6, we approach this problem in two main directions, including the normal approach and the divided approach. After converting the data, we use the F2-score in the normal approach. With the divided approach, model 1 is assessed with an F2 score, and model 2 uses the F1 score because we consider the importance of labels 1 and 2 to be the same. Finally, perform a combination of these two binary classification models and evaluate the final result in the same manner as the normal approach. Furthermore, both the F1 and F2 scores are combined with the macro-avg method to accurately evaluate the imbalanced dataset because this method is not affected by class weights.

6 Experiments and results

6.1 Experimental produce

In this section, we have an overview of our research from the beginning until the results are suitable for the warning system in Fig. 4. From the raw data about the subjects and students' learning results, we have transformed them into necessary features in the training model process, and the terms and formulas related to the transformation are presented in detail in Sect. 3.1. Then, in order to get more data, as well as improve the performance of the problem, we performed data generation in Sect. 3.2. Through data processing, we initially obtained a group of datasets corresponding to each semester, with the number of samples and feature descriptions listed in Tables 2 and 3. Next, the problems of missing data and imbalanced data were applied with appropriate techniques before applying the prediction algorithms. We apply feature approaches to both the normal approach and the divided approach for each proposed algorithm as well as the baseline in turn to determine the best combination for improved performance.

Regarding the evaluation and training on algorithms and techniques, since our group of datasets consists of many small datasets, we will evaluate based on the average performance. The algorithm's performance will be

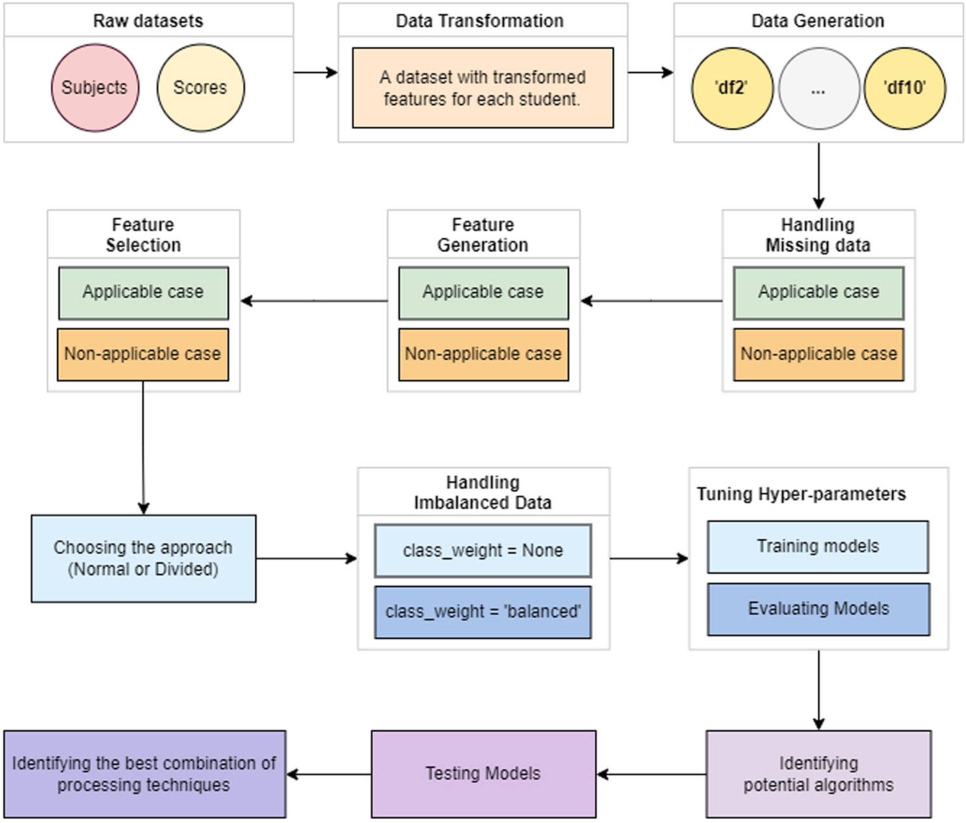


Fig. 4 Our experimental procedure

evaluated based on its average results over the entire set of small datasets.

The search for the optimal model is evaluated by the method K-folds cross-validator approach with K=5, which is a technique that helps ensure that the model performs well with imbalanced data and gets away from overfitting. Furthermore, the parameter search to build the best model is done manually and automatically using the HalvingGridSearchCV technique. A new model selection strategy significantly reduces the time required to operate a model selection procedure. Our average evaluation and experiment process is shown in detail in Fig. 5.

Many machine learning and deep learning models have proven their effectiveness in many studies related to the field of prediction of students' academic performance [28–30]. Therefore, besides the algorithms proposed in Sect. 5.2, other models are also built with baseline algorithms to see the effectiveness of the proposed algorithms. We selected a variety of popular traditional and modern machine learning algorithms to build baseline models, including Decision Tree, Random Forest, Extra Trees, Gradient Boosting Decision Tree, Logistic Regression, and Stochastic Gradient Descent Classifier.

Decision Tree (DT)[31] is one of the most widely used machine learning algorithms. Both classification and regression issues can for its applied. In principle, a decision tree is a tree in which every node stands for a feature, every branch for a rule, and every leaf for a conclusion (specific value or a continuation branch). The decision tree is built based on many different types of algorithms, such as ID3, C4.5, CART, CHAID, and others.

Random Forest (RF)[32] works based on an ensemble of many decision trees, but each decision tree is unique (with a random factor). It allows for the resolution of overfitting and data interference issues. Each decision tree assigns a classification to each object; the Random Forest's ultimate class is determined by the class that receives the most votes.

Extra Trees (ET)[33] is similar to Random Forest, except instead of bootstrapping observations, nodes are split depending on random splits among a random subset of the features chosen at each node.

Gradient Boosting Decision Tree (GB)[13] builds a tree initially to try to fit the data and then builds more trees using a loss function influenced by the Gradient Descent to minimize residuals. It accomplishes this by concentrating on areas where existing learners are underperforming by applying the Boosting.

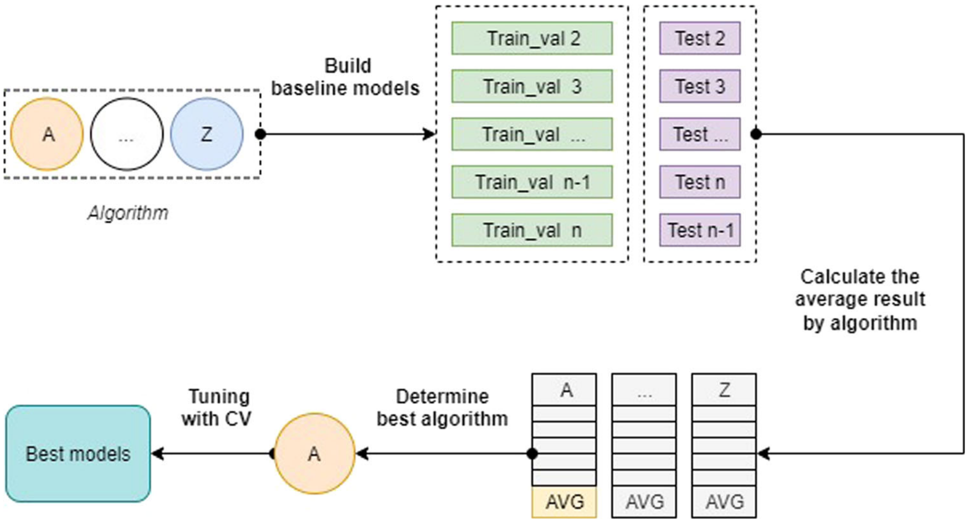
Logistic Regression (LR) [34] is one of the machine learning algorithms that is most frequently applied to binary classification issues. It makes use of the logistic sigmoid function to produce probabilistic estimates for prediction purposes.

Stochastic Gradient Descent Classifier (SGD) [35] can be seen as a linear classifier (in this paper, we use Logistic Regression) optimized by Stochastic Gradient Descent.

Neural Support Vector Machine (N-SVM) [36] is a hybrid learning technique that combines Support Vector Machines (SVM) and neural networks (SVM). SVM takes the output of NN as an input to predict final outcomes. The architecture of the N-SVM is depicted in Fig. 6.

The obtained prediction models are a harmonious combination between algorithms and the proposed strategies and techniques. However, a point that needs noticing is that algorithm GB, one of our baseline algorithms, does not support using class weight to handle data imbalance. That algorithm handles class imbalance by generating successive training sets from incorrectly classified

Fig. 5 Our average evaluation and experiment process



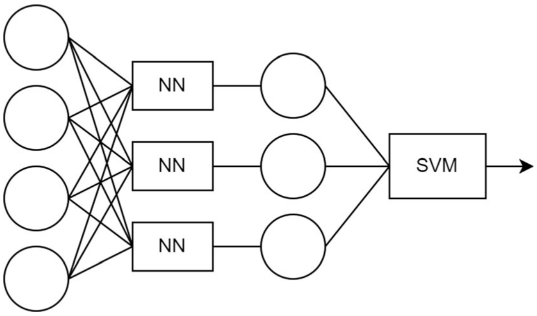


Fig. 6 The architecture of N-SVM

examples. As a result, we did not use any imbalanced processing steps when creating the model with it.

6.2 Results

6.2.1 The first warning: beginning of each semester

Table 7 represents our experimental results on Task 1. Based on the results table, the proposed algorithms svm.SVC and LGBM give superior results in most of the experimental cases. The combination of algorithms svm.SVC and LGBM with the use of the ’class_weight = balanced’ parameter is a potential solution to improving the learning ability of the model. Besides, that solution also shows its strong influence on some other algorithms, such as N-SVM, SGD, and LR. We also averaged the results obtained over the tests applying class_weight adjustment and the case without using it. As a result, that solution improved the model performance by 15% and 11%, respectively, with svm.SVC and LGBM. Regarding using the Divided approach, in the first task, this solution is unsuitable in most algorithms and causes a negative effect on the experimental results. We compared the results obtained when experimenting with feature generation and feature selection techniques, which also improved. Specifically, in our proposed algorithms, the successive application of creating new features and removing redundant features increases the model’s predictive power by nearly 1% on LGBM and more than 5% on svm.SVC. Based on the above results, in the two proposed algorithms, svm.SVC is the most suitable for the first task. The best result obtained on test data is F2 = 74.37%. Combinations of svm.SVC, class_weight = ’balanced,’ and feature processing methods such as feature generation and selection produced that result. Besides, we were awarded that the algorithm N-SVM is also a potential algorithm to solve the first task. This algorithm gives good results second only to svm.SVC in most experimental cases. The power of neural computing seems to be reduced because the amount of data is not large enough. We are in great hope of being able to experiment with this algorithm on a larger dataset to see the performance of neural computing clearly.

6.2.2 The second warning: before the final examination

Table 8 shows the experimental results before the final examination. The academic warning system predicts outcomes with greater accuracy when applying feature techniques and using a new approach—the divided approach. In the two proposed algorithms, the LGBM model is more suitable than svm.SVC for the second warning, LGBM outperforms the others when mostly achieving higher results in this predicted stage. With SVM, we have conducted more experiments on the Neural network and SVM

Table 7 The experimental results of the first warning

Approach	Without FG, FS				With FG				With FG and FS			
	Normal		Divided		Normal		Divided		Normal		Divided	
Handling imbalanced data	None	cw=’B’	None	cw=’B’	None	cw=’B’	None	cw=’B’	None	cw=’B’	None	cw=’B’
Baseline												
DT	0.6471	0.6421	0.5130	0.5572	0.6397	0.6482	0.4866	0.5364	0.6334	0.6220	0.4771	0.5467
ET	0.6597	0.6551	0.4481	0.4707	0.6567	0.6502	0.4641	0.4509	0.6553	0.6405	0.4335	0.4756
GB	0.6709	–	0.5811	–	0.6803	–	0.5575	–	0.6638	–	0.4895	–
LGBM	0.6681	0.7161	0.4948	0.5822	0.6739	0.7129	0.4702	0.5811	0.6715	0.7164	0.4915	0.6183
LR	0.6408	0.6996	0.4819	0.6116	0.6459	0.7000	0.4840	0.6092	0.6428	0.7129	0.5394	0.6192
RF	0.6672	0.6494	0.5295	0.4927	0.6684	0.6375	0.4558	0.4649	0.6631	0.6423	0.4611	0.4972
SGD	0.6550	0.6995	0.4976	0.5042	0.6556	0.6565	0.4686	0.5312	0.6491	0.7129	0.4542	0.4995
NSVM	0.6287	0.7048	0.4272	0.5880	0.6338	0.7011	0.4159	0.6014	0.6332	0.7175	0.4183	0.6136
Proposal												
svm.SVC	0.6485	0.7351	0.4467	0.5197	0.6470	0.7249	0.4467	0.7317	0.6380	0.7437	0.4467	0.7397

Table 8 The experimental results before the final exam

Approach	Without FG, FS				With FG				With FG and FS			
	Normal		Divided		Normal		Divided		Normal		Divided	
Handling imbalanced data	None	cw='B'	None	cw='B'	None	cw='B'	None	cw='B'	None	cw='B'	None	cw='B'
Baseline												
RF	0.7403	0.7840	0.8296	0.8636	0.8454	0.8795	0.8819	0.8809	0.8280	0.8473	0.8828	0.8623
SGD	0.5515	0.6464	0.6333	0.7270	0.6039	0.6807	0.6806	0.7859	0.5762	0.6784	0.8584	0.8092
LR	0.6178	0.7808	0.6741	0.7350	0.7282	0.8316	0.7290	0.7706	0.7695	0.8718	0.7512	0.8117
DT	0.8411	0.8106	0.8513	0.8495	0.8957	0.8836	0.8956	0.8494	0.8793	0.8914	0.8780	0.8745
svm.SVC	0.5284	0.7425	0.5285	0.6931	0.5747	0.7738	0.5748	0.7402	0.5284	0.6420	0.5768	0.7881
N-SVM	0.6608	0.7987	0.5616	0.7031	0.6692	0.8709	0.5760	0.7406	0.4135	0.6533	0.5708	0.7849
ET	0.6613	0.7311	0.6961	0.7311	0.6419	0.7684	0.6766	0.7487	0.7133	0.7909	0.7437	0.7284
GB	0.7963	–	0.8792	–	0.9034	–	0.9012	–	0.8845	–	0.9041	–
Proposal												
LGBM	0.7268	0.9081	0.8615	0.9116	0.8160	**0.9271**	0.8837	0.8861	0.8359	0.9224	0.8833	**0.9262**

(N-SVM) combination. The results obtained are better than traditional SVM (87%) but still not higher than LGBM. The divided approach demonstrates how effective it is in our training. In the algorithms, the divided approach experimental results by an average of 5.5% compared to the normal approach. When it comes to improving most experiments, the feature generation technique is no less competitive, with up to 6% improvements in some cases. Furthermore, the feature selection technique aids in the removal of unnecessary features while maintaining model accuracy. In conclusion, our proposed approach and algorithm delivered excellent results, with F2-score of 92.71% when applying all features generation combined with handling imbalanced data and with F2-score delivered of 92.62% for the divided approach and feature selections combined with handling imbalanced data.

6.3 Result analysis and visualization

6.3.1 The first warning: beginning each semester

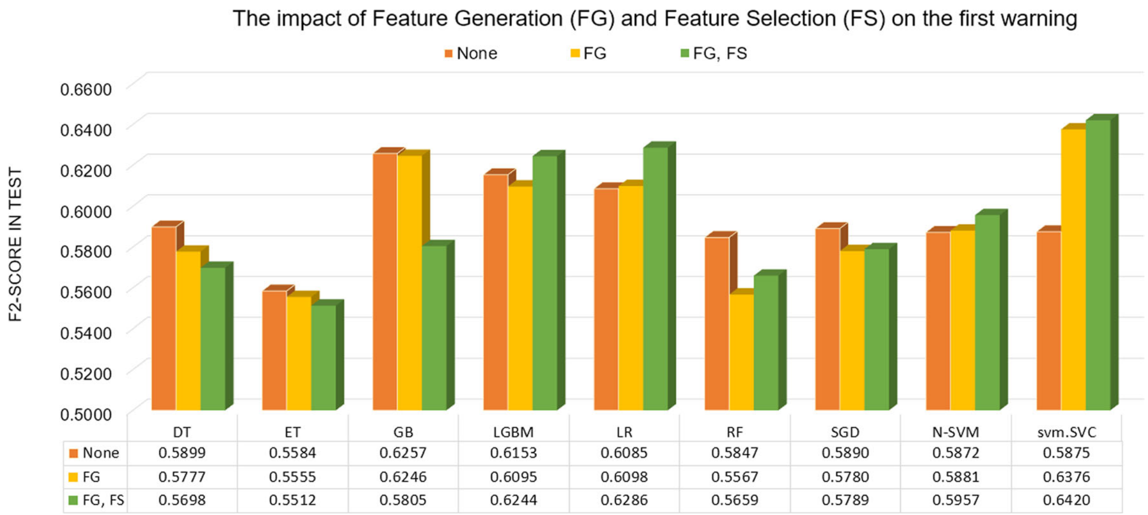
The bar chart in Fig. 7 is built based on the results of Table 7. Where the values used are the average of experiments grouped by feature processing method and algorithm. We have a clearer view of the effectiveness of the process of applying feature generation and selection solutions. The suitability between those solutions and the algorithms svm.SVC and LR are shown visually through the ascending of the value columns. Particularly for LGBM, creating more new features causes a negative effect, but when choosing the really important features, the performance is immediately improved. Moreover, the remaining baseline algorithms are not suitable for this data as well as this processing method. Based on the above comments, it can be affirmed that the harmony between the algorithm and the processing methods is extremely important.

The mismatch of the Divided approach is clearly shown in Fig. 8. At all algorithms, that approach causes a significant reduction in model performance, ranging from 8% to 20%.

Figure 9 vividly demonstrates the greatness of using class_weight to handle data imbalances through the columns chart. With all algorithms, when grouping by results is obtained by case class_weight is used and not used, it is easy to see that using this solution is completely wise. The most significant improvement is on the models of algorithm svm—more than 13% of performance is improved. For RF alone, there is a slight 1% drop in performance, which is hard to explain with a black box algorithm like RF.

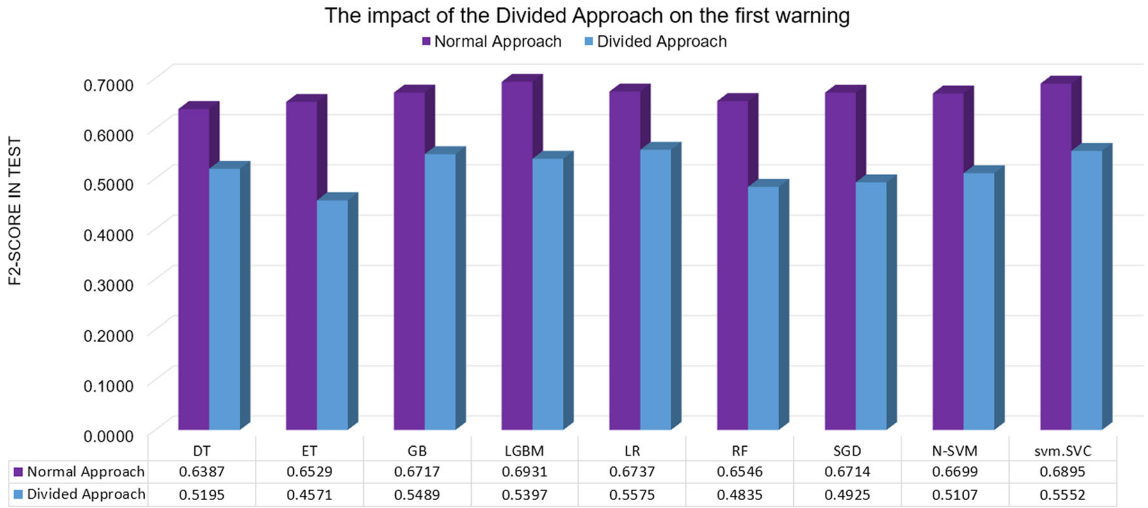
6.3.2 The second warning: before the final examination

The chart in Figs. 10, 11, 12, and 13 are built based on the results of Table 8—the experimental results before the final examination. For each combination of techniques and approaches, we calculate the average and visualize it to assess each method's performance by each algorithm easily. It can be seen that our model (LGBM) and proposed techniques yield higher average results than most models. Figures 10, 11, and 12 show that dealing with imbalanced data, applying feature generation techniques, and feature selection works well on mostly baseline and our proposal models. The results show the importance of pre-processing the data as well as finding new and suitable features in the



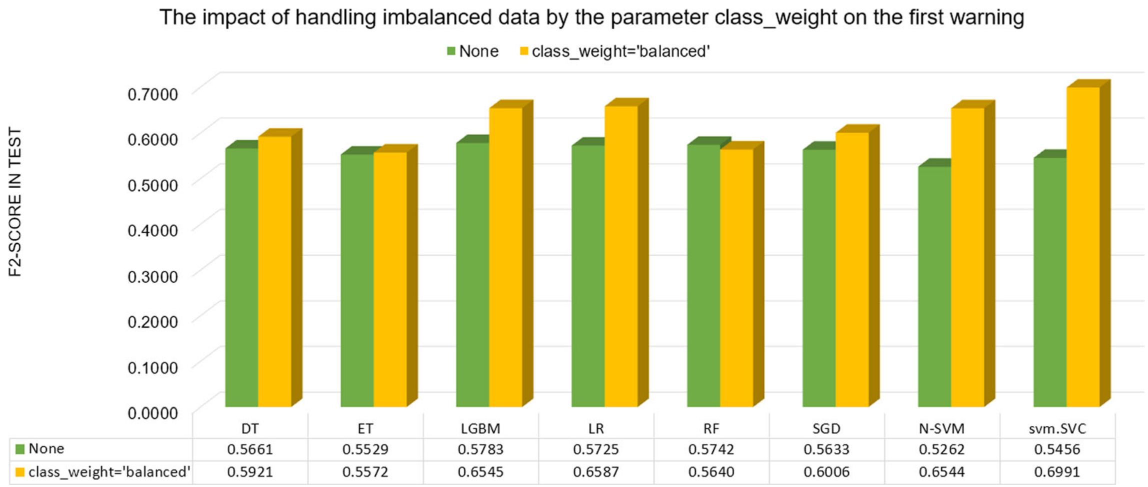
	DT	ET	GB	LGBM	LR	RF	SGD	N-SVM	svm.SVC
■ None	0.5899	0.5584	0.6257	0.6153	0.6085	0.5847	0.5890	0.5872	0.5875
■ FG	0.5777	0.5555	0.6246	0.6095	0.6098	0.5567	0.5780	0.5881	0.6376
■ FG, FS	0.5698	0.5512	0.5805	0.6244	0.6286	0.5659	0.5789	0.5957	0.6420

Fig. 7 The impact of feature generation and selection techniques on the first warning



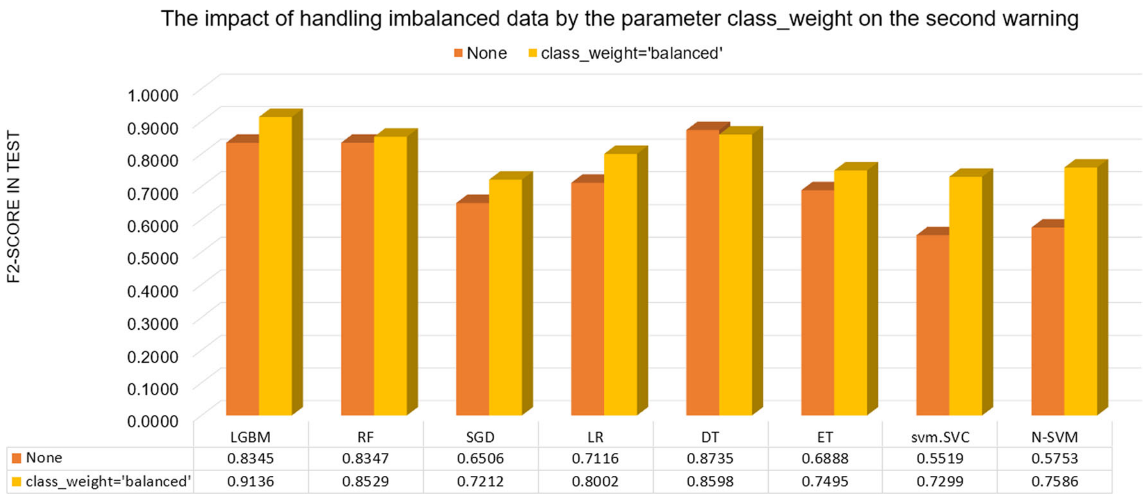
	DT	ET	GB	LGBM	LR	RF	SGD	N-SVM	svm.SVC
■ Normal Approach	0.6387	0.6529	0.6717	0.6931	0.6737	0.6546	0.6714	0.6699	0.6895
■ Divided Approach	0.5195	0.4571	0.5489	0.5397	0.5575	0.4835	0.4925	0.5107	0.5552

Fig. 8 The impact of divided approach on the first warning



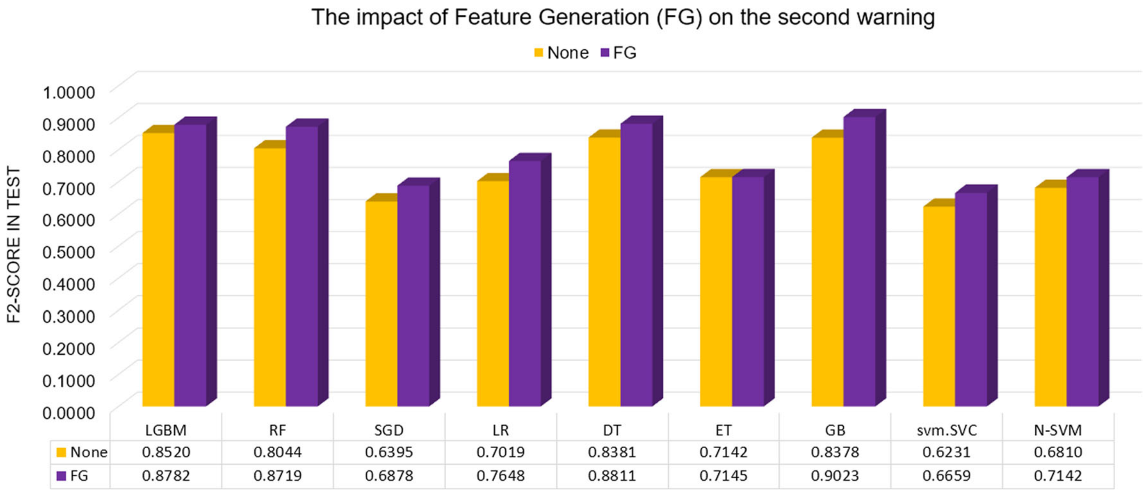
	DT	ET	LGBM	LR	RF	SGD	N-SVM	svm.SVC
■ None	0.5661	0.5529	0.5783	0.5725	0.5742	0.5633	0.5262	0.5456
■ class_weight='balanced'	0.5921	0.5572	0.6545	0.6587	0.5640	0.6006	0.6544	0.6991

Fig. 9 The impact of handling imbalanced data on the first warning



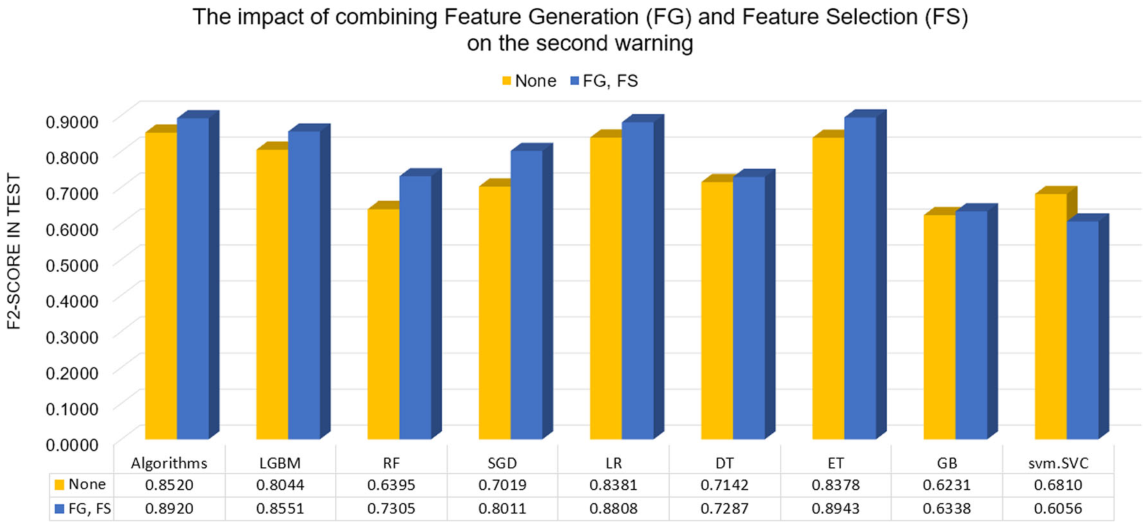
	LGBM	RF	SGD	LR	DT	ET	svm.SVC	N-SVM
None	0.8345	0.8347	0.6506	0.7116	0.8735	0.6888	0.5519	0.5753
class_weight='balanced'	0.9136	0.8529	0.7212	0.8002	0.8598	0.7495	0.7299	0.7586

Fig. 10 The comparison before and after applying handling imbalanced data



	LGBM	RF	SGD	LR	DT	ET	GB	svm.SVC	N-SVM
None	0.8520	0.8044	0.6395	0.7019	0.8381	0.7142	0.8378	0.6231	0.6810
FG	0.8782	0.8719	0.6878	0.7648	0.8811	0.7145	0.9023	0.6659	0.7142

Fig. 11 The comparison before and after applying feature generation



	Algorithms	LGBM	RF	SGD	LR	DT	ET	GB	svm.SVC
None	0.8520	0.8044	0.6395	0.7019	0.8381	0.7142	0.8378	0.6231	0.6810
FG, FS	0.8920	0.8551	0.7305	0.8011	0.8808	0.7287	0.8943	0.6338	0.6056

Fig. 12 The comparison before and after applying feature generation and feature selection

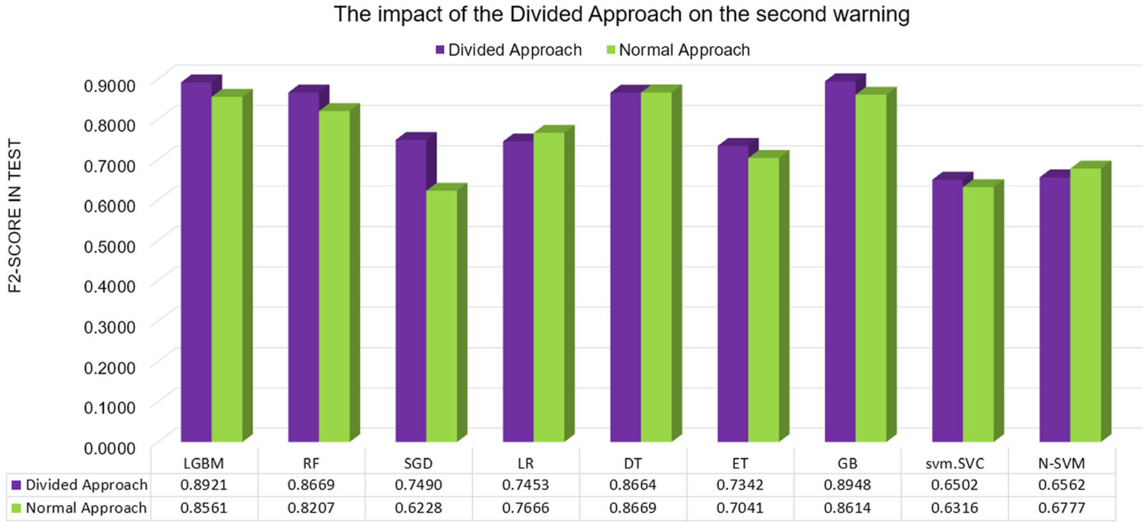


Fig. 13 The comparison before and after applying the divided approach

training process. The new approach—the divided approach, gives promising results when it works well and significantly improves performance compared to the normal approach, in some cases up to almost 13% improvement (Fig. 13). Therefore, we conclude that all the proposed new techniques and approaches are important factors in detecting the risk of academic warnings in our system and topics related to student classification in academics in general.

6.4 Feature selection

Because the small datasets have a different number of features belonging to Group 1 (the average grade of the previous semesters completed by students), we used another feature, s_all, which is calculated as the average of the features in Group 1, to calculate the correlation with the remaining features. Figure 14 is a heatmap showing the average Pearson correlation of the features. Based on this heatmap, we selected highly correlated features with other features, then experimented with removing them from the training process. Through the process of testing and selection, at each academic warning stage, we have found suitable features.

We examined the correlation between features s_all, pre_avg, and history for the first warning. According to the theory referenced in Bruce Ratner's article [37] about the correlation coefficient and the heatmap in Fig. 14, feature history, generated by feature generation, is thought to be correlated with feature pre_avg. Additionally, both pre_avg and history are highly correlated with s_all (0.74 and −0.68). Therefore, we decided to drop the feature pre_avg because it has a higher correlation with s_all than the feature history. When applying the feature selection technique, the performance of our proposed model (svm.SVC) has improved slightly, about 1.5% average across experiments on the test set. Although the change is not too significant, this technique has contributed to reducing the complexity of the learning process of the model.

For the before final exam warning model, we remove avg1, avg2, avg3, pre_avg, coef_final, avg_final, and history. As a result, in Table 8 this removal yielded the unexpected result of improving the results of the proposed model from 88.61% to 92.62% with a divided approach. In the normal approach, although the application of feature selection has not shown any significant effect in improving performance, it has succeeded in fewer features that are not necessary, decreasing the system's training and prediction but still keeping model performance.

7 Conclusion and future work

Academic warnings for students in universities help counselors can detect abnormal student problems as early as possible. Academic performance is the direct factor reflecting the student's situation in learning, providing comprehensible analysis and data for decision-making assistance for the university. However, academic performance is shown in many aspects, and it is not easy to choose affecting aspects in decision-making. Our research suggested possible techniques and formulas extract and generate valuable academic performance features. Many algorithms and combined inputs are experimented with to find best practices for the academic warning system. We used the result of the feature generation, feature selection, and handling of the imbalanced data to build a system to

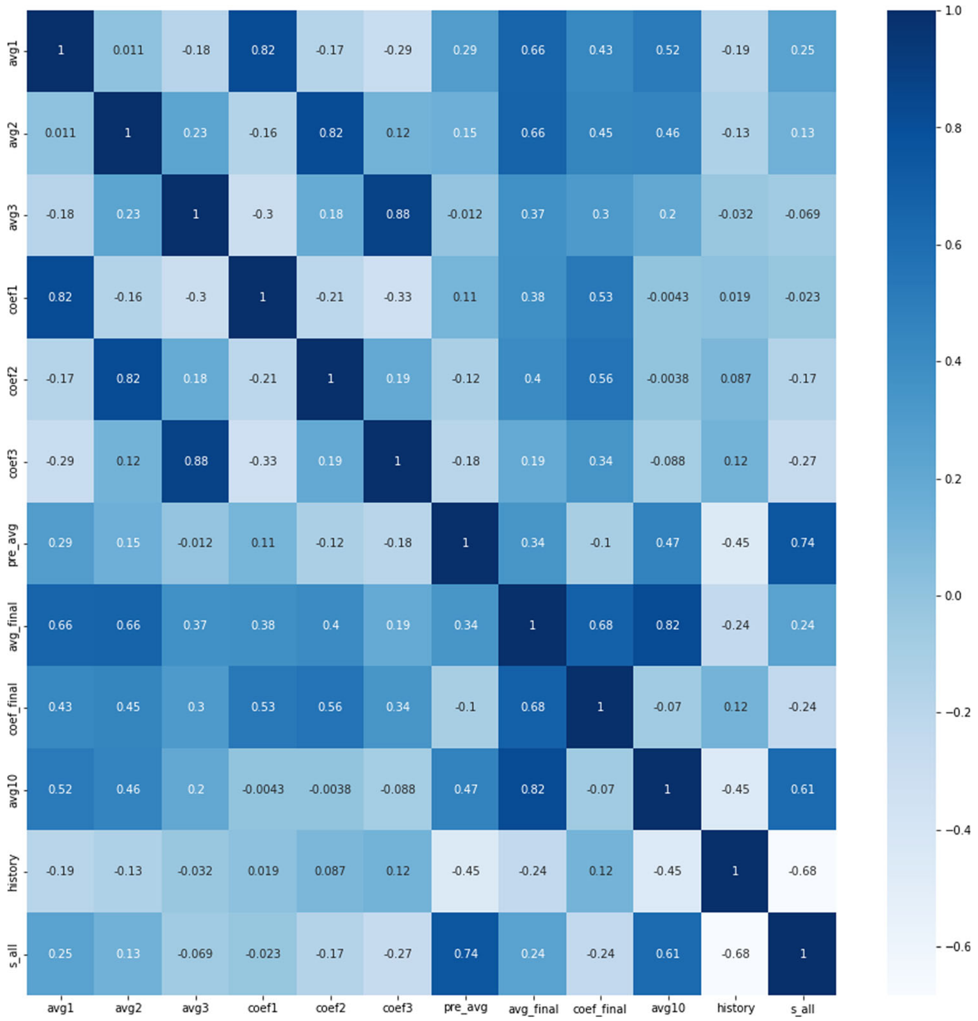


Fig. 14 Pearson correlation coefficient between the input features

provide a prediction of academic performance in the new semester. We develop a group dataset (APWD) with valuable features from a huge educational raw database. In addition to the semester GPA features, the emergence of generation features such as “history” and “avg10” and handling imbalanced data techniques are outstanding results discovered from our research processes. APWD can be reused or reconstructed to apply to other universities in their academic warning system, and the proposed techniques can be considered to improve their system. The proof is that the performance of the warning model is considerably improved as a result of these techniques. In general, we initially achieved fairly good results. A two-stage academic performance warning system for higher education is suggested with the F2-score measure of over 74% at the beginning of the semester and over 92% before the final exam. In light of the findings, we believe that predictive models based on our suggested techniques can help mitigate the rise in academic probation warnings at universities, hence establishing a positive learning environment for Vietnamese students in particular and students worldwide in general.

In the future, we will continue to look for additional features in the raw database to boost the predictability of the beginning of the semester model and collect new and timely data to improve model performance and check the model quality when put into reality. In addition, not stopping with warnings only based on academic performance, we always want to be able to exploit and explore more new aspects that affect students’ learning status even more to build a perfect warning system. The elements that are intended to be explored further are subject types, semester characteristics, significant influences, and so on. Furthermore, to simplify the problem, we plan to discover a method to keep only one input format for all semesters while still providing good performance.

Appendix

Grading System

See Table 9.

Table 9 Grade conversion table

Category	10 Scale	100 Scale	4 Scale	US Grade	Grade meaning
Pass	9,0 – 10,0	90 – 100	4,0	A+	Outstanding
	8,0 – 9,0	80 – 90	3,5	A	Excellent
	7,0 – 8,0	70 – 80	3,0	B+	Good
	6,0 – 7,0	60 – 70	2,5	B	Fairy Good
	5,0 – 6,0	50 – 60	2,0	C	Average
Fail	4,0 – 5,0	40 – 50	1,5	D+	Conditional Fail
	3,0 – 4,0	30 – 40	1,0	D	Fail
	< 3,0	<30	0,0	F	

Acknowledgements This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Miguéis VL, Freitas A, Garcia PJV, Silva A (2018) Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. 115:36–51
2. Mingyu Z, Sutong W, Yanzhang W, Dujuan W (2021) An interpretable prediction method for university student academic crisis warning. Complex Intell Syst 8(1):323–336
3. Bujang SDA, Selamat A, Ibrahim R, Krejcar O, Herrera-Viedma E, Fujita H, Ghani NAM (2021) Multiclass prediction model for student grade prediction using machine learning. IEEE Access 9:95608–95621
4. Hamim T, Benabbou F, Sael N (2022) Student profile modeling using boosting algorithms. Int J Web-Based Learn Teach Technol (IJWLTT) 17(5):1–13
5. Namdeo J, Jayakumar N (2014) Predicting students performance using data mining technique with rough set theory concepts. Int J Adv Res Comput Sci Manag Stud 2:367–373
6. Madeira B, Tasci T, Çelebi N (2021) Prediction of student performance using rough set theory and backpropagation neural networks. Eur Sci J. https://doi.org/10.19044/esj.2021.v17n7p1
7. Pham H-D, Le TD, Nguyen VT (2018) Static PE malware detection using gradient boosting decision trees algorithm. In: FDSE
8. Corchs S, Fersini E, Gasparini F (2019) Ensemble learning on visual and textual data for social image emotion classification. Int J Mach Learn Cybern 10(8):2057–2070
9. Yunan Z, Huang Q, Ma X, Yang Z, Jiang J (2016) Using multi-features and ensemble learning method for imbalanced malware classification, pp. 965–973. https://doi.org/10.1109/TrustCom.2016.0163
10. Possebon IP, Silva AS, Granville LZ, Schaeffer-Filho A, Marnerides A (2019) Improved network traffic classification using ensemble learning. In: 2019 IEEE symposium on computers and communications (ISCC), pp. 1–6. https://doi.org/10.1109/ISCC47284.2019.8969637
11. Carrasco R, Sicilia-Urban M-A (2020) Evaluation of deep neural networks for reduction of credit card fraud alerts. IEEE Access 8:186421–186432. https://doi.org/10.1109/ACCESS.2020.3026222
12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: a highly efficient gradient boosting decision tree. In: NIPS
13. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
14. Hearst MA (1998) Trends and controversies: support vector machines. IEEE Intell. Syst. 13:18–28
15. Hasan H, Shafri H, Al-Habshi M (2019) A comparison between support vector machine (SVM) and convolutional neural network (CNN) models for hyperspectral image classification. IOP Conf Ser Earth Environ Sci 357:012035. https://doi.org/10.1088/1755-1315/357/1/012035
16. Gabrilovich E, Markovitch S (2005) Feature generation for text categorization using world knowledge. In: IJCAI
17. Li L, Yang H, Jiao Y, Lin K-Y (2020) Feature generation based on knowledge graph. IFAC-PapersOnLine 53(5):774–779. https://doi.org/10.1016/j.ifacol.2021.04.172 (**3rd IFAC Workshop on Cyber-Physical & Human Systems CPHS 2020**)
18. Shi H, Li H, Zhang D, Cheng C, Cao X (2018) An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. Comput. Netw. 132:81–98. https://doi.org/10.1016/j.comnet.2018.01.007
19. Nahar N, Ara F, Neloy MAI, Biswas A, Hossain MS, Andersson K (2021) Feature selection based machine learning to improve prediction of Parkinson disease. In: Mahmud M, Kaiser MS,

Vassanelli S, Dai Q, Zhong N (eds) Brain informatics. Springer, Cham, pp 496–508

20. Rahman L, Setiawan NA, Permanasari AE (2017) Feature selection methods in improving accuracy of classifying students' academic performance. In: 2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE), pp. 267–271. https://doi.org/10.1109/ICITISEE.2017.8285509
21. Chen R-C, Dewi C, Huang S, Caraka R (2020) Selecting critical features for data classification based on machine learning methods. J Big Data 7:26. https://doi.org/10.1186/s40537-020-00327-4
22. Chumerin N, Van Hulle MM (2006) Comparison of two feature extraction methods based on maximization of mutual information. In: 2006 16th IEEE signal processing society workshop on machine learning for signal processing, pp. 343–348. https://doi.org/10.1109/MLSP.2006.275572
23. Guyon I (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
24. Barros T, SouzaNeto P, Silva I, Guedes LA (2019) Predictive models for imbalanced data: a school dropout perspective. Educ. Sci. 9:275. https://doi.org/10.3390/educsci9040275
25. Rachburee N, Punlumjeak W (2021) Oversampling technique in student performance classification from engineering course. Int J Electr Comput Eng 11:3567
26. More A (2016) Survey of resampling techniques for improving classification performance in unbalanced datasets
27. Kumar P, Bhatnagar R, Gaur K, Bhatnagar A (2021) Classification of imbalanced data: review of methods and applications. IOP Conf Ser Mater Sci Eng 1099(1):012077. https://doi.org/10.1088/1757-899x/1099/1/012077
28. Rovira S, Puertas E, Igual L (2017) Data-driven system to predict academic grades and dropout. PLoS ONE 12(2):0171207
29. Huynh-Ly T-N, Le H-T, Thai-Nghe N (2021) Integrating deep learning architecture into matrix factorization for student performance prediction. In: Dang TK, Küng J, Chung TM, Takizawa M (eds) Future data and security engineering. Springer, Cham, pp 408–423
30. Yağcı M (2022) Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn Environ 9(1):11. https://doi.org/10.1186/s40561-022-00192-z
31. Quinlan JR (2004) Induction of decision trees. Mach Learn 1:81–106
32. Breiman L (2004) Random forests. Mach Learn 45:5–32
33. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42
34. Niu L (2020) A review of the application of logistic regression in educational research: common issues, implications, and suggestions. Educ Rev 72(1):41–67. https://doi.org/10.1080/00131911.2018.1483892
35. Robbins HE (2007) A stochastic approximation method. Ann Math Stat 22:400–407
36. Wiering M, Ree M, Embrechts M, Stollenga M, Meijster A, Nolte A, Schomaker L (2013) The neural support vector machine
37. Ratner B (2009) The correlation coefficient: its values range between $+1/-1$, or do they? J Target Meas Anal Market. https://doi.org/10.1057/jt.2009.5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.