Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Machine learning approaches for predicting student dropout

Tran Thanh Nam[2], Nguyen Van Linh[2], Nguyen Anh Duy[1], Ngo Ho Anh Khoi[2*]

[1]*Adhightech Ltd., Vietnam*

[2]*Faculty of Information Technology, Nam Can Tho University, Vietnam*

[*]*Corresponding author: Ngo Ho Anh Khoi (email: Ngohoanhkhoi@gmail.com)*

Received: 30/12/2024

Revised: 25/1/2025

Accepted: 10/2/2025

Keywords: *AI, classification, machine learning, predict student dropout dataset*

Từ khóa: *AI, dự đoán bộ dữ liệu học sinh bỏ học, học máy, phân loại*

ABSTRACT

The issue of student dropout during their educational journey is a growing concern with far-reaching implications. This problem not only affects students and their families directly but also poses significant challenges to higher education institutions and society at large. The ramifications for students include increased difficulties and a deficiency in soft skills and life experience, often leading them to seek part-time employment. This study focuses on developing a machine learning model through the process of analyzing, comparing, and evaluating the performance of five models: AdaBoost, DecisionTree, RandomForest, ExtraTree, and BernoulliNB. All models are implemented using the "Predict Student Dropout Dataset." Based on the results obtained after processing the data, the study will conduct an analysis based on two main criteria: evaluation by average percentage, standard deviation, and final outcomes, as well as evaluation using a time-series model of age (Balanced Accuracy Progression). From these analyses, the model with the optimal performance will be selected. By identifying the underlying causes and addressing these issues effectively, the research aims to reduce the burden on families and society, mitigate social problems, stimulate economic growth, generate job opportunities, and enhance both competitiveness and productivity. This dataset is of substantial value for researchers conducting comparative studies on student academic performance and serves as a crucial resource for training in the field of machine learning.

TÓM TẮT

Vấn đề học sinh bỏ học trong quá trình học tập đang là mối lo ngại ngày càng tăng với những hệ lụy sâu rộng. Vấn đề này không chỉ ảnh hưởng trực tiếp đến sinh viên và gia đình họ mà còn đặt ra những thách thức đáng kể cho các cơ sở giáo dục đại học và xã hội nói chung. Hậu quả đối với sinh viên bao gồm việc gia tăng khó khăn và thiếu hụt các kỹ năng mềm cũng như kinh nghiệm sống, thường khiến họ phải tìm kiếm việc làm bán thời gian. Nghiên cứu này tập trung phát triển mô hình machine learning thông qua quá trình phân tích, so sánh và đánh giá hiệu suất của 5 mô hình: AdaBoost, DecisionTree, RandomForest, ExtraTree và BernoulliNB. Tất cả các mô hình đều được triển khai bằng cách sử dụng "Bộ dữ liệu dự đoán học sinh bỏ học". Dựa trên kết quả thu được sau khi xử lý số liệu, nghiên cứu sẽ tiến hành phân tích dựa trên 2 tiêu chí chính: đánh giá theo tỷ lệ phần trăm trung bình, độ lệch chuẩn và kết quả cuối cùng cũng như đánh giá bằng mô hình chuỗi thời gian theo độ tuổi (Balanced Accuracy Progression). Từ những phân tích này, mô hình có hiệu suất tối ưu sẽ được lựa chọn. Bằng cách xác định các nguyên nhân cơ bản và giải quyết các vấn đề này một cách hiệu quả, nghiên cứu nhằm mục đích giảm bớt gánh nặng cho gia đình và xã hội, giảm thiểu các vấn đề xã hội, kích thích tăng trưởng kinh tế, tạo cơ hội việc làm và nâng cao cả khả năng cạnh tranh và năng suất. Bộ dữ liệu này có giá trị đáng kể đối với các nhà nghiên cứu thực hiện nghiên cứu so sánh về kết quả học tập của sinh viên và đóng vai trò là nguồn tài nguyên quan trọng để đào tạo trong lĩnh vực học máy.

1. INTRODUCTION

Success in higher education, particularly at the university level, is critically important for employment opportunities and societal advancement. The issue of student dropout is a global concern, with increasing rates posing significant challenges to societal development and economic growth. This problem has drawn the attention of administrators and scholars alike. According to Dr. Hamish Coates, factors contributing to student dropout include the impractical and disengaging nature of current university curricula, often marked by rigid and unappealing programs. Additionally, students from rural and remote areas frequently face financial hardships that force them to abandon their studies due to the high costs of education and living expenses in urban areas [1]. In Vietnam, this issue is particularly pronounced. In 2017, approximately 30% of students at the Ho Chi Minh City University of Technology (HCMUT) were unable to continue until their final academic year, as reported by Mr. Le Chi Thong, Head of the Training Department [5].

Similarly, in 2019, the Ho Chi Minh City University of Industry (HUIC) issued a warning to 2,252 students who had voluntarily withdrawn during the first semester of the 2019-2020 academic year. This included students from various programs such as regular university degrees, vocational college degrees, and part-time university programs [6],[7]. Additionally, statistical data from the Ministry of Education and Training (MOET) indicated that 1,163 students had to discontinue their studies due to financial constraints, with 556 enrolled in university programs and 607 in vocational college programs [8].

The dropout issue in Vietnam and globally has severe implications, including difficulties in keeping up with curricula, economic hardships, insufficient institutional support, and high unemployment rates among graduates. These challenges significantly affect educational institutions, families, and society. To address this, higher education institutions need to identify and tackle the root causes of student dropout, motivating students to return to their studies. Initiatives have been proposed to prioritize student learning and provide access to loans for continuing education.

This study aims to develop effective intervention strategies, thereby improving the quality of education, reducing student dropout rates, and optimizing resources for student support. The dataset, "Predict Students' Dropout and Academic Success," includes demographic, socio-economic, and academic achievement data of students in higher education. This data can be used to analyze predictive factors for student dropout and success.

Key Applications of the Dataset:

Predicting Student Retention Rates: The dataset can be used to create models that identify risk factors for dropout, enabling timely interventions to improve student retention.

Enhancing Academic Performance: By analyzing the dataset, institutions can gain insights into students' learning progress and identify areas for improvement, allowing for the development of more effective courses and initiatives. The demographic data can help develop initiatives that make educational resources more accessible to targeted groups, addressing disparities in access to higher education services based on region or socio-economic status.

This research initiative is distinct from typical classroom knowledge and requires significant time and effort to develop solutions that enhance educational quality and training, ultimately providing optimistic predictions to address the issue of student dropout effectively.

2. RESEARCH METHODS

During the process of data search for the research topic, numerous datasets were identified (approximately 108 datasets related to student dropout). However, a few datasets with comprehensive specifications and the highest availability were selected, including: "Predict_Student_Dropout_dataset" by Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins (licensed by MDPI, Basel, Switzerland), "Student Performance Prediction: Complete Analysis" by Victor Régis [4], "Predicting Student Dropout with KNN" by Lukás Németh [3], and "Predict students' Dropout and Academic Success" by The Devastator [2].

In the paper "Student Dropout Prediction" by Stefano Pio Zingaro, it is argued that predicting dropout rates can be addressed by harnessing machine learning techniques that have demonstrated effectiveness in the education domain for evaluating student performance. The paper "Challenges and Solutions to the Student Dropout Prediction Problem in Online Courses" by Lorenzo Madeddu applies automated policies to fully exploit the advantages of students' activities (referred to as e-activities) on digital platforms, thereby identifying at-risk students. These approaches involve machine learning and deep learning techniques to predict student dropout. Hence, coping with the changing trend of student interactions with course platforms in real-time has become critically important. This guide presents a thorough review of the literature addressing the issue of student dropout prediction. It includes mathematical representations for different definitions proposed in the field and explores both basic and advanced prediction approaches. The discussion encompasses key aspects such as defining student dropout, input modeling, the application of deep learning and traditional machine learning methods, evaluation metrics, datasets, and concerns related to data privacy [12]. In a research report titled "Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian's Secondary Schools" from Taylor & Francis Online, an AutoML model is discussed, which combines various automatic techniques to create the final ML model. These techniques encompass data preparation, feature engineering, model creation, and model evaluation. The AutoML model also addresses large-scale automated data preprocessing, feature engineering, model search, and hyperparameter optimization, to predict student dropouts in developing secondary schools by selecting the best-relevant model through a document summarization process.

In our country, Tra Vinh Province has the highest dropout rate among students. As a result, the Chairman of the People's Committee of Trà Vinh Province has issued directives to review the number of dropout students and find effective solutions. The reasons for this issue stem from the low living standards of the residents, limited infrastructure, and deficiencies in the teaching staff's qualifications. A seminar was held to address the issue of student dropouts, proposing measures such as enhancing the quality of teaching staff and management personnel, improving infrastructure, and promoting educational socialization. Given the increasing trend of impoverished students dropping out, the official government portal collaborated with the Ministry of Education and Training, the Ministry of Finance, and the Vietnam Social Policy Bank to organize a scientific seminar with the theme "Enabling Poor Students to Afford Education". During the seminar, representatives from the Ministry of Education and Training, the Ministry of Finance, and the Vietnam Social Policy Bank reached a consensus that the purpose of providing student loans is to alleviate some economic difficulties, allowing financially challenged students to continue their education. The seminar emphasized the genuine economic challenges faced by a segment of students, particularly those from rural areas with economic hardships. However, the representative from the Vietnam Social Policy Bank asserted that they only

provide partial support for the students' difficulties. The listed datasets are all related to predicting student dropout status. However, only one recent dataset can fulfill the requirements of the research topic. To be suitable for the research, the dataset must be numeric, have specific class labels, and include multiple fields to yield objective results. The "Predict Student Dropout dataset" is the only dataset that meets these criteria, making it the chosen dataset for this study.

In Vietnam, there is currently a lack of publicly available datasets for predicting student dropout status on data-sharing websites. The dataset used in this research is sourced from another country. The prediction of student dropout status can encompass various factors, including the quality of education in each country and region, with varying criteria for determination. The centralized database remains unchanged over time, trained using classical methods (a one-time training process that needs to be retrained from scratch with new data). However, in the modern context, the data environment evolves over time, necessitating continuous real-time training and model updates. As the foundational database is collected from a few countries, the scope of the research topic on student dropout is also limited. The research topic "Predicting Student Dropout Using Machine Learning Methods" employs the RandomForest algorithm during the investigation. It necessitates continuous learning in an evolving data environment, implying that the experimental approach involves continuous learning in non-stable environments.

Modern practice demands continuous real-time training to efficiently update predictive models, leading to numerous studies on continuous learning methods, also known as incremental learning or evolving learning. Continuous learning methods have been thoroughly studied and analyzed in [10]. One of the simplest methods is the sliding window technique, which has been applied to many classical algorithms to facilitate continuous learning. This method continuously updates the model at each time point 't' by using the latest training data within a predefined window size 's'. The window size 's' can be based on time or the number of data points and typically overlaps with the previous window 'w'. At each iteration, a new model is trained, reflecting an updated set of classes. The basic model of the sliding window technique is illustrated in Figure 1.

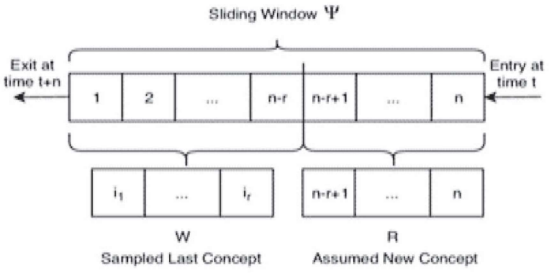

Figure 1. Operational model of the Sliding Window technique [11]

After downloading, the dataset is referred to as raw data and cannot be directly used for research due to the need for extensive preprocessing. This involves transforming various attributes and features from their stated parameters to their expressed characteristics. Machine learning methods require that attributes or features of the dataset be represented as numerical values in the input layer, and unnecessary information such as serial numbers or IDs can be eliminated. Subsequently, label transformation will be performed, where "Dropout" labels representing student dropouts

will be converted to the number "0," "Graduate" labels representing graduating students will be converted to "1," and the remaining "Enrolled" labels representing students yet to graduate will be transformed to "2".

All data processing steps are carried out entirely within the Microsoft Excel software. After transforming the data from raw format to a standardized form for system use, the dataset must be saved in. CSV file format. Only one cell is allowed for each data row, and attributes as well as labels must be separated by a comma. According to system regulations, data attributes must be placed at the front, while labels should be positioned at the end. Following all preprocessing steps to standardize the data, we will obtain a normalized dataset containing 34 features with a total of 4424 data points, including attributes in the following format in Figure 2.

```
1,8,5,2,1,1,1,13,10,6,10,1,0,0,1,1,0,20,0,0,0,0,0,0,0,0,0,0,0,0,10.8,2.2,6.74,0,
1,6,1,11,1,1,1,1,3,4,4,1,0,0,0,1,0,19,0,0,6,6,6,14,0,0,6,6,6,13.66666667,0,13.9,0.5,5.79,1,
1,1,5,5,1,1,1,22,27,10,10,1,0,0,0,1,0,19,0,0,6,0,0,0,0,0,6,0,0,0,0,10.8,2.2,6.74,0,
1,8,2,15,1,1,1,23,27,6,4,1,0,0,1,0,0,20,0,0,6,8,6,13.42857143,0,0,6,10,5,12.4,0,9.4,0,1.88,1,
2,12,1,3,0,1,1,22,28,10,10,0,0,0,1,0,0,45,0,0,6,9,5,12.33333333,0,0,6,6,6,13,0,13.9,0.5,5.79,1,
2,12,1,17,0,12,1,22,27,10,8,0,0,1,1,1,0,50,0,0,5,10,5,11.85714286,0,0,5,17,5,11.5,5,16.2,1.1,4.08,1,
1,1,1,12,1,1,1,13,28,8,11,1,0,0,1,0,1,18,0,0,7,9,7,13.3,0,0,8,8,8,14.345,0,15.5,3.6,0.94,1,
1,9,4,11,1,1,1,22,27,10,10,1,0,0,0,1,0,22,0,0,5,5,0,0,0,0,5,5,0,0,0,15.5,3.6,0.94,0,
1,1,3,10,1,1,15,1,1,10,10,0,0,0,1,0,1,21,1,0,6,8,6,13.875,0,0,6,7,6,14.14285714,0,16.2,1.1,4.08,1,
1,1,1,10,1,1,1,1,14,5,8,1,0,1,0,0,0,18,0,0,6,9,5,11.4,0,0,6,14,2,13.5,0,8.9,2.2,8.51,0,
1,1,1,14,1,1,1,23,14,6,8,1,0,0,1,0,0,18,0,0,6,6,6,12.33333333,0,0,6,7,5,14.2,0,13.9,0.5,5.79,1,
1,1,1,12,1,1,1,13,28,10,10,1,0,0,1,0,1,18,0,0,8,8,7,13.21428571,0,0,8,8,7,13.21428571,0,12.7,4.5,3.3,1,
1,1,2,16,1,1,1,13,27,5,10,1,0,0,1,0,0,19,0,0,6,6,0,0,0,0,6,6,0,0,0,12.7,4.5,3.3,0,
1,17,1,11,1,16,1,1,1,5,8,1,0,0,1,0,1,21,0,0,6,7,6,10.57142857,0,0,6,8,5,11,0,8.9,2.2,8.51,1,
1,1,1,6,1,1,1,23,27,6,6,1,0,0,1,0,1,18,0,0,5,7,4,13.25,0,0,5,5,5,12,0,10.8,2.2,6.74,1,
1,1,1,15,1,1,1,13,27,10,4,1,0,0,1,0,0,20,0,0,6,6,5,13.2,0,0,6,7,0,0,0,15.5,3.6,0.94,0,
1,9,1,10,1,1,1,13,28,6,9,1,0,0,1,0,0,18,0,0,6,10,1,12,0,0,6,14,2,11,0,10.8,2.2,6.74,2,
1,8,2,12,1,1,1,13,1,6,5,1,0,0,1,0,0,18,0,0,7,8,7,13.30625,0,0,8,8,8,14.545,0,15.5,3.6,0.94,1,
1,1,1,8,1,1,1,3,14,4,6,1,0,0,1,0,0,20,0,0,5,8,4,12.5,1,0,5,8,4,12.25,2,10.8,2.2,6.74,1,
1,1,1,16,1,1,1,13,14,8,8,1,0,0,1,0,0,18,0,0,7,7,6,11.66666667,0,0,7,8,6,13.5,0,16.2,1.1,4.08,2,
```

Figure 2. Attributes of the dataset

To effectively utilize this dataset, one needs to familiarize oneself with all the variables provided in the dataset, including categorical variables such as gender or study program; numerical variables like the number of instructional units at the beginning of the semester or age at enrollment; ordinal measurement variables like marital status; study of temporal trends such as inflation rate or GDP; frequency measurement variables like the percentage of scholarship recipients, etc. Details regarding the attributes present in the dataset are described in the Table 1.

Table 1. Description of attributes in the Predict Student Dropout dataset

No.	Name	Type	Describe
1	Marital status	Float	1- Single, 2- Marry, 3- Divorce
2	Application mode	Float	1- Formal, 2- Students over 23 years old, 3- Connected objects, 4- Second degree subject, 5- Have a postgraduate degree, 6- International student, 7- Change major, 8- Transfer school

3	Application order	Float	From 1 to 9 (default is 1)
4	Course	Float	1- Biotechnology, 2- Economy, 3- The law, 4- Agriculture, 5- Medicine, 6- Pharmacy, 7- Information Technology, 8- Aquaculture, 9- Business administration, 10- The law, 11- Tourism, 12- Nursing, 13- Other
5	Daytime evening attendance	Float	1- Daytime, 0- Evening
6	Previous qualification	Float	1- High school diploma, 2- College degree, 3- Master's degree, 4- Other
7	Nationality	Float	1- Portugal, 2- Other
8	Mother's qualification	Float	1- High school graduation, 2- Graduated from high school or college, 3- Graduate, 4- Master graduate, 5- Graduated with a doctorate, 6- Haven't graduated from college yet, 7- Haven't graduated from high school, 8- Other (10)
9	Father's qualification	Float	1- High school graduation, 2- Graduated from high school or college, 3- Graduate, 4- Master graduate, 5- Graduated with a doctorate, 6- Haven't graduated from college yet, 7- Haven't graduated from high school, 8- Other (10)
10	Mother's occupation	Float	1- Student, 2- State employees, 3- Science, 4- Intermediate technician, 5- Management staff, 6- Personal business, 7- Farmer, 8- Craftsman, 9- Worker, 10- General labor, 11- Army, 12- Other
11	Father's occupation	Float	1- Student, 2- State employees, 3- Science, 4- Intermediate technician, 5- Management staff, 6- Personal business, 7- Farmer, 8- Craftsman, 9- Worker, 10- General labor, 11- Army, 12- Other
12	Displaced	Float	1- Yes, 0- No
13	Educational special needs	Float	1- Yes, 0- No
14	Debtor	Float	1- Yes, 0- No
15	Tuition fees up to date	Float	1- Yes, 0- No
16	Gender	Float	1- Male, 0- Female

17	Scholarship holder	Float	1- Yes, 0- No
18	Age at enrollment	Float	From 17 to 70
19	International	Float	1- Yes, 0- No
20	Curricular units 1st sem (credited)	Float	From 0 to 20
21	Curricular units 1st sem (enrolled)	Float	From 0 to 26
22	Curricular units 1st sem (evaluations)	Float	From 0 to 45
23	Curricular units 1st sem (approved)	Float	From 0 to 26
24	Curricular units 1st sem (grade)	Float	From 0.000 to 18.875
25	Curricular units 1st sem (without evaluations)	Float	From 0 to 12
26	Curricular units 2nd sem (credited)	Float	From 0 to 19
27	Curricular units 2nd sem (enrolled)	Float	From 0 to 23
28	Curricular units 2nd sem (evaluations)	Float	From 0 to 33
29	Curricular units 2nd sem (approved)	Float	From 0 to 20
30	Curricular units 2nd sem (grade)	Float	From 0.
32	Unemployment rate	Float	From 7.600 to 16.200
33	Inflation_rate	Float	From -0.800 to 3.700
34	Gross Domestic Products	Float	From -4.100 to 3.500

The Table 1 above provides a structured overview of various data fields related to student information. Below, each field is explained in detail, highlighting its nature and relevance to the dataset. This explanation offers clarity on the categorical and numerical attributes used to describe and analyze the students' profiles.

Marital status: The marital status of the student. (Categorical)

Application mode: The method of application used by the student. (Categorical)

Application order: The order in which the student applied. (Numerical)

Course: The course taken by the student. (Categorical)

Daytime/evening attendance: Whether the student attends classes during the day or in the evening. (Categorical)

Previous qualification: The qualification obtained by the student before enrolling in higher education. (Categorical)

Nationality: The nationality of the student. (Categorical)

Mother's qualification: The qualification of the student's mother. (Categorical)

Father's qualification: The qualification of the student's father. (Categorical)

Mother's occupation: The occupation of the student's mother. (Categorical)

Father's occupation: The occupation of the student's father. (Categorical)

Displaced: Whether the student is a displaced person. (Categorical)

Educational special needs: Whether the student has any special educational needs. (Categorical)

Debtor: Whether the student is a debtor. (Categorical)

Tuition fees up to date: Whether the student's tuition fees are up to date. (Categorical)

Gender: The gender of the student. (Categorical)

Scholarship holder: Whether the student is a scholarship holder. (Categorical)

Age at enrollment: The age of the student at the time of enrollment. (Numerical)

International: Whether the student is an international student. (Categorical)

Curricular units 1st sem (credited): The number of curricular units credited by the student in the first semester. (Numerical)

Curricular units 1st sem (enrolled): The number of curricular units enrolled by the student in the first semester. (Numerical)

Curricular units 1st sem (evaluations): The number of curricular units evaluated by the student in the first semester. (Numerical)

Curricular units 1st sem (approved): The number of curricular units approved by the student in the first semester. (Numerical)

The dataset is transient in nature as new and diverse datasets will be added in the future, leading to potential changes in the data classification process. This indicates that the dataset exists in an unstable environment, necessitating improvements to accommodate Vietnam's context. Therefore, classical methods in static classification environments cannot be employed. Instead, an incremental machine learning approach should be selected to address the task, allowing for easy adaptation to Vietnam's standards or specific regional requirements without the need to reuse old data. This approach enables continuous updates in real-time, which is crucial given the evolving nature of the data. Consequently, traditional static machine learning algorithms are unsuitable for use with this dataset. The legacy methods do not allow for class expansion, making it necessary to employ progressive algorithms capable of handling data in dynamic environments.

The system calculates the balanced accuracy metrics for the results. Traditional accuracy measures the proportion of correctly classified cases out of the total number, and it is generally reliable. However, this metric can be misleading in cases of severe class imbalance, such as a 90:10 ratio. For example, if 100 cases are tested with 99 cases being diseased and 1 case healthy, the balanced accuracy might appear high even if no meaningful model is present. Therefore, for imbalanced datasets, Balanced Accuracy (BA) is used. The choice of evaluation metrics depends on the problem's objectives and the composition of the dataset. In situations with significant class imbalance, where one class is underrepresented, traditional accuracy becomes unreliable. Therefore, metrics like the area under the ROC curve (AUC) and BA are preferred. Metrics such as balanced accuracy, sensitivity, and specificity are less effective for imbalanced data. For concordance detection, metrics based on the true positive rate/false positive rate, such as balanced accuracy, sensitivity, and F-Score, are appropriate. In contrast, for discordance detection, metrics based on the true negative rate/false negative rate, such as specificity, are suitable, although less common in practice. Sensitivity, balanced accuracy, and F-Score are criticized for ignoring the true negative cell of the confusion matrix and being prone to prediction bias [9]. BA, which includes both the true positive rate and the true negative rate, provides a balanced evaluation, making it suitable for both concordance detection and imbalanced data situations [10]. BA is an important and simple metric for evaluating binary classifiers in the context of class imbalance, where one class is much more prevalent than the other. The formula for balanced accuracy (BA), which provides a practical and optimal evaluation, is:

Balanced Accuracy (BA)=

1⁄2 (Specificity + Sensitivity) (1)

3. RESULTS AND DISCUSSION

The software used in this study are all common tools widely used by the majority of users. The prerequisite for a dataset to be usable in this study is that it must contain numerical data, specific class labels, and multiple attributes to yield objective results. There are two methods to interact with the system: one is using the command line for developers with advanced technical knowledge, and the other is a user-friendly interface that allows everyone to access and utilize the system easily and quickly without requiring extensive technical expertise. The system model describes both the user's usage process and the developer's system installation process. The system was simultaneously developed in two directions: command line and graphical interface, a practice that is not common among many systems. This approach effectively addresses the challenge of dynamic data by developing algorithms that adapt to the model-based training, unlike other algorithms that can only be applied to static data through traditional training approaches. The "Predict Student Dropout Dataset" project employs software such as Visual Studio Code, Sublime Text, Microsoft Office, SQLite, among others. The challenge is overcome by developing algorithms that are geared towards model-based training, which sets it apart from other algorithms that can only be applied using traditional static data training methods. The models used for conducting scientific experiments have been discussed earlier, so in this section, the emphasis will be on

comparing and analyzing 5 algorithms: AdaBoost, BernoulliNB, ExtraTree, DecisionTree, and RandomForest to determine the most suitable algorithm.

Table 2. Performance metrics of Machine Learning models

Algorithms	Average (%)	Last step (%)	Standard deviation (%)
AdaBoost	66.21	69.25	3.74
DecisionTree	59.96	62.31	2.18
RandomForest	70.78	73.51	3.18
ExtraTree	57.02	58.92	2.24
BernoulliNB	56.98	57.95	0.80

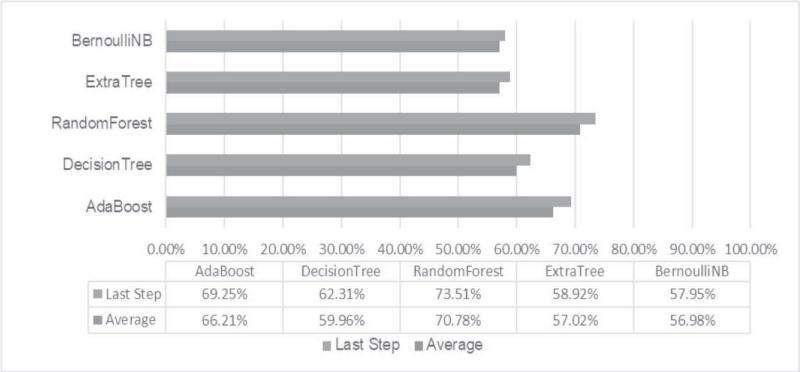

Figure 3. Chart of average percentage for experimental algorithms by age groups (Data groups)

Based on the chart data in Fig 3. and performance metrics in Table 2, when comparing the average and final results of the five algorithms, there is a noticeable relative difference between them. In particular, detailed analysis shows that the ExtraTree and BernoulliNB algorithms have lower performance compared to the others, with both achieving less than 60% for both the average and final rates. Specifically, the average rate of ExtraTree is about 57.02%, while the final rate is slightly higher at 58.91%. For BernoulliNB, both rates are even less effective, with an average rate of 56.98% and a final rate of 57.95%. Therefore, it can be concluded that these two algorithms are not suitable for solving the problem.

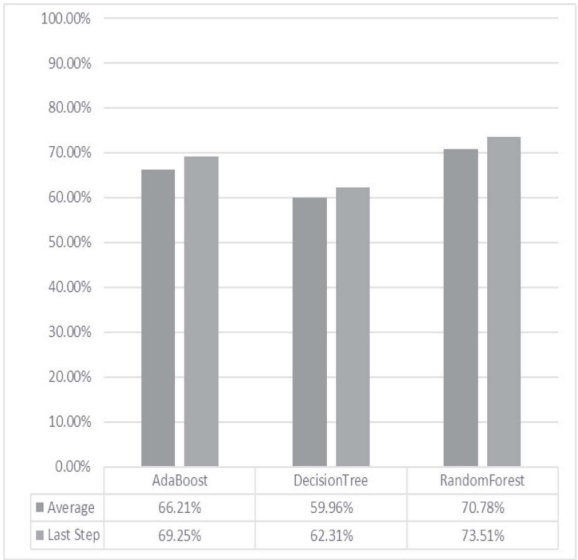

Figure 4. Comparison chart of the top 3 algorithms in average and last steps results

By eliminating the two least effective algorithms, based on observations from Figure 4, we focus on the remaining three algorithms: AdaBoost, DecisionTree, and RandomForest. A significant gap is observed between the performance of AdaBoost and DecisionTree. The

average rate difference between the two is 6.25%, and this gap increases to 6.94% when considering the final step rate. RandomForest stands out as the leading algorithm, achieving an average rate of 70.78% and a final step rate of 73.51%. From this analysis, it is clear that RandomForest is the most efficient and stable algorithm. AdaBoost is also a strong contender with high performance in the final step. Conversely, DecisionTree, ExtraTree, and BernoulliNB show lower performance, with BernoulliNB being the least effective.

Beyond simply averaging algorithm results, another approach to analyzing the effectiveness of experimental models is to examine performance by age group. This analysis provides a more comprehensive and detailed view, allowing us to understand the model's capabilities for specific age groups. It also enables a more accurate assessment and conclusion about the experimental model's performance. The improvement in balanced accuracy is illustrated in Figure 5:

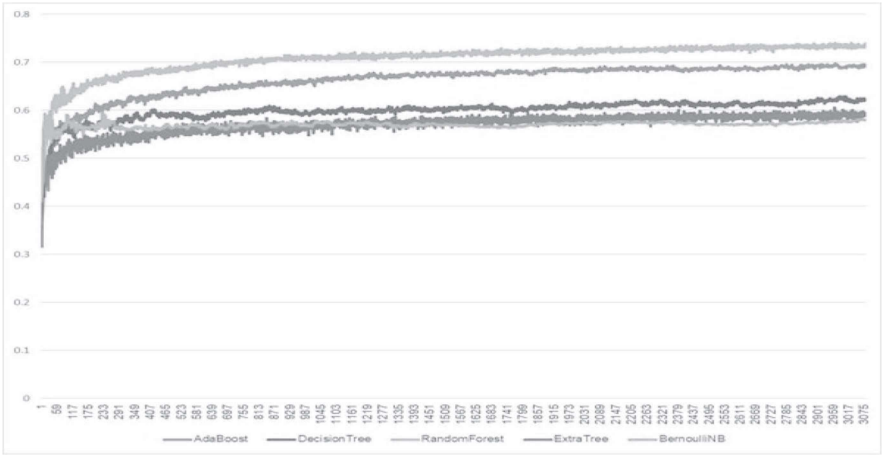

Figure 5. BA progression chart of the five algorithms

Refer to Figure 5., we see that all five algorithms start with relatively low performance in the first 10-50 steps and gradually stabilize in the subsequent steps.

First, with the AdaBoost algorithm, there is a rapid increase from 30% to around 60% in the initial steps, stabilizing around 66%. The strength of this algorithm lies in its fast growth and ability to reach stable accuracy quickly, though its final performance is not as high as RandomForest. DecisionTree increases from 30% to about 60%, but then fluctuates significantly and stabilizes only around 62%. Although it is a simple and easy-to-understand algorithm, DecisionTree has lower final performance and more fluctuations throughout most stages.

ExtraTree also rises from 30% to about 57% but shows significant fluctuations and stabilizes around 58%. The strength of ExtraTree lies in its faster execution compared to RandomForest, but the performance achieved is not as high and shows more variability in results. BernoulliNB increases rapidly from 30% to around 56% in the initial steps and stabilizes around 57%. This algorithm's strength is its simplicity, speed, and effectiveness with binary classification data, but in this problem, it shows lower performance compared to the other algorithms.

RandomForest stands out with the highest performance, quickly rising from 30% to about 70% in the initial steps and stabilizing around 73%. This is the best-performing algorithm, with high stability and minimal fluctuations. Using multiple decision trees helps RandomForest minimize overfitting and increase accuracy. Although it requires more computational resources and is more complex, its superior performance and stability make RandomForest the best choice among the compared algorithms. To select the optimal algorithm for solving the problem of "Predicting Student Dropout," a comprehensive review of both the average and final step rate comparison charts and the experimental chart by age group indicates that RandomForest is the best choice in terms of rate and stability. The strengths of the RandomForest algorithm include high and stable performance, the ability to handle large and complex datasets, flexibility in application for both classification and regression tasks, and the ability to assess the importance of features. RandomForest performs well on large datasets, can handle complex relationships among input variables, and provides insights into feature importance, helping users better understand the data and model. Therefore, RandomForest is prioritized as the best algorithm to solve the problem of "Predicting Student Dropout."

3.1 System setup

From the perspective of real-world application, creating the necessary dataset for learning and prediction in the experimental environment in Vietnam is crucial. In Vietnam, due to the lack of related databases, immediately establishing a prediction system is not feasible. The initialization of a dataset for this topic, aimed at conducting learning processes and building a prediction system, requires substantial resources, which could take years or even decades. Currently, implementing this solution remains challenging. A more cost-effective approach, which has been applied in similar cases, is to use continuous learning models. Instead of waiting for a large amount of data to proceed with the learning process and build the prediction system, we can use a small amount of data to improve the model by continuously adjusting the basic concepts and gradually shifting the initial basic concept closer to a new basic concept (based on the Vietnamese dataset). This process is called "concept drift," and the model will be continuously improved by adding more accurate new data (data on Vietnam’s land and crops). This method allows the prediction system to be used immediately and gradually improved through small errors in the model, rather than waiting a long time before the improved model can be utilized.

Based on the final results presented in the previous section, the Random Forest algorithm was selected to address the problem. The application will include features such as prediction functionality, running classical algorithms, a list of processed models, system configuration, and login. It will be installed on a web environment and will be divided into main functions: the algorithm installer (administrator or developer) and the diagnostician (user), described by the use case diagram below:

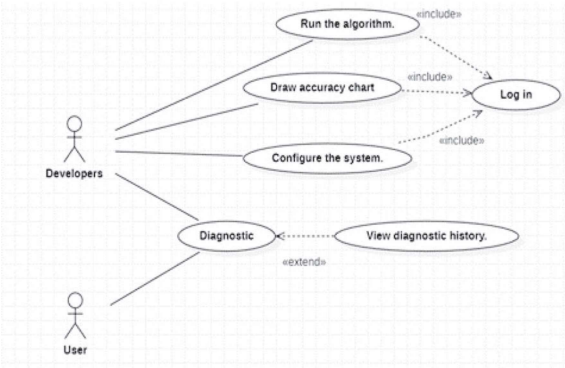

Figure 6. Use case diagram

Download the file "setup.zip," and after extracting it, you will find the following key files and folders: SETUP, DB, APP, INSTALL.bat, RunServer.bat, requirements.txt. Install the Python program by running the 'python-3.9.9-amd64.exe' file located in the SETUP folder. Install the necessary libraries to run the program by executing the CaiThuVien.bat file. Running the Remove.bat file will delete all program data. The database file is located in the 'DB' folder and is named Data.db, which can be opened using the 'DB Browser for SQLite.exe' tool located in 'DB\DB Browser for SQLite'. To change the administrator account, edit the file '\APP\static\dataUser.csv'.

To start the program, run the 'RunServer.bat' file or open the command line and run the command ‘manage.py runserver’. The default server port is 8000, which can be changed by using the command ‘manage.py runserver <port>’. When the command line displays ‘Starting development server at http://127.0.0.1:8000/’, you can access the main application page at ‘http://127.0.0.1:8000/’ (Figure 7).

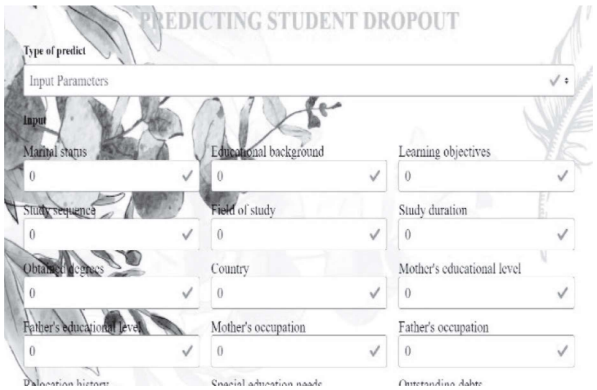

Figure 7. Website index

3.2 Key features

User functions: Predict dropout rates based on input data (see in Fig 7).

View prediction history for past analyses:

Developer Functions (requires login): Predict dropout rates with the ability to adjust settings. Configure the platform, including switching between machine learning models (see in Figure 8) Analyze datasets using selected machine learning models and export results for further study.

Convenience: The platform provides tailored functionality for both end-users (students, teachers) and developers (data analysts, educators). The ability to test and compare algorithms makes the system adaptable and insightful.

Advantages: User-focused design: Easy for non-technical users to predict dropout rates. Developers can customize machine learning models, enhancing the platform's utility for research and educational purposes.

Disadvantages: The reliance on a specific dataset may limit the generalizability of predictions across other contexts. The developer functionality requires technical expertise, which may not be accessible to all users.

Figure 8. Algorithm selection and data execution interface

4. CONCLUSION

The implementation of the classic RandomForest algorithm has been successfully completed in the project. The report is thorough, with detailed explanations of each data point, chart, and algorithm. Following a comprehensive research and evaluation process, the project delivers well-rounded results. The model training and prediction phases, along with the challenges of handling variable data and dynamic environments, have been meticulously addressed, fully satisfying the initial requirements. Future development plans for this project include the continuous updating of new data through survey methods and domain expertise to ensure an optimized and up-to-date dataset that accurately reflects real-world conditions. The project will be segmented into two distinct sections: one for regular users and another for developers. Regular users will have access to a user-friendly interface for predicting student dropout status, while developers will have more advanced pages for further system development.

REFERENCES

[1] Hien. (2020). *Thesis: Measures to overcome the dropout situation of university students.* Tailieumau. Retrieved from https://tailieumau.vn/de-tai-bien-phap-khac-phuc-tinh-trang-bo-hoc-cua-sinh-vien-dh-hot/

[2] The Devastator. (2023). *Predict students' dropout and academic success.* Kaggle. Retrieved from https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

[3] Nememth, L. (2023). *Predicting student dropout with KNN.* Kaggle. Retrieved from https://www.kaggle.com/code/luknmeth/predicting-student-dropout-with-knn

[4] Regis, V. (2021). *Student performance prediction: Complete analysis.* Kaggle. Retrieved from https://www.kaggle.com/code/devassaxd/student-performance-prediction-complete-analysis

[5] Le, H., & Thanh, T. (2017). *University students keep 'dropping out'.* Vietnamnet. Retrieved from https://vietnamnet.vn/sinh-vien-dai-hoc-lien-tuc-roi-rung-405761.html

[6] Le, H. (2019). Students dropping out midway. *People's Newspaper.* Retrieved from https://nhandan.vn/sinh-vien-bo-hoc-giua-chung-post381039.html

[7] Ha, A. (2022). *Thousands of students don't attend even after being admitted.* Youth. Retrieved from https://thanhnien.vn/hang-ngan-sinh-vien-khong-hoc-du-trung-tuyen-1851518611.htm

[8] Thao, P. (2017). Nearly 1,200 students drop out due to economic difficulties. *Sai Gon Liberation Newspaper.* Retrieved from https://www.sggp.org.vn/gan-1200-sinh-vien-bo-hoc-vi-kinh-te-kho-khan-post119317.html

[9] Powers, D. M. (2020). *Evaluation: From precision, recall and F-measure to ROC,*

informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[10] Khoi, N. H. A. (2015). *Méthodes de classifications dynamiques et incrémentales: Application à la numérisation cognitive d'images de documents* (Doctoral dissertation). Tours.

[11] Raab, C., Heusinger, M., & Schleif, F. M. (2020). Reactive soft prototype computing for concept drift streams. *Neurocomputing,* 416. https://doi.org/10.1016/j.neucom.2019.11.111

[12] Prenkaj, B., Stilo, G., & Madeddu, L. (2020). Challenges and solutions to the student dropout prediction problem in online courses. *In CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* https://doi.org/10.1145/3340531.3412172