Computers and Education: Artificial Intelligence 6 (2024) 100244

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Enhancing educational evaluation through predictive student assessment modeling

Pham Xuan Lam [a,*], Phan Quoc Hung Mai [b], Quang Hung Nguyen [b], Thao Pham [a], Thi Hong Hanh Nguyen [b], Thi Huyen Nguyen [c]

[a] Faculty of Information Technology, National Economics University, Viet Nam
[b] Faculty of Mathematical Economics, National Economics University, Viet Nam
[c] Faculty of Education, Hanoi University of Science and Technology, Viet Nam

ARTICLE INFO

Keywords:
Data science applications in education
Educational data mining
Improving classroom teaching
Teaching/learning strategies
Learning analytics

ABSTRACT

This study evaluates several machine learning models used in predicting student performance. The data utilized in this study was collected from 253 undergraduate students participating in five classes within one of three courses offered by VnCodelab, an interactive learning management system, to provide insights into student performance. Leveraging the data-rich environment of the interactive learning management system proposed earlier, this study focuses on training a predictive model that forecasts student grades based on the comprehensive data collected during the teaching process. The proposed model capitalizes on the data obtained from students' engagement patterns, time spent on exercises, and progress tracking across learning activities. This study compared five different base classifiers— Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), and k-nearest Neighbor (k-NN), and an ensemble learning method Stacking Classifier —utilizing a dataset comprising 13 features. The research assesses the model's accuracy, reliability, and implications, contributing to the evolution of educational evaluation by introducing predictive assessment as a transformative tool. The results indicate that the Stacking Classifier accurately predicts students' grade ranges, surpassing individual base classification models by effectively combining their predictive capabilities. Integrating data-driven forecasting into the educational ecosystem can transform teaching methodologies and foster an informed, engaged, and empowered learning environment. This approach cultivates a proactive learning community by empowering students with real-time academic progress forecasts. Educators benefit from data-informed insights that facilitate more effective and objective performance evaluation.

1. Introduction

The booming age of technological advancement, coupled with the force of the COVID-19 pandemic has pushed the appearance of various interactive learning environments and online courses (Hermanto & Srimulyani, 2021). In Vietnam, particularly in the recent COVID-19 context, online education and the utilization of online learning support software (such as Microsoft Teams, Zoom, and Learning Management System) have become increasingly essential (Hoa et al., 2020). Many studies found that these environments can enhance student engagement, collaboration, and critical thinking skills (Dennen, 2008; Hovlid et al., 2022; Means et al., 2009). However, evaluating offline classes and subsequently grading learners are straightforward while regulators of online learning environments face several challenges (Kebritchi et al., 2017; Yuniastari & Silva, 2022). Due to the shortage of institutional administrative support, instructors on online platforms consume considerable time for administrative work while they teach online courses (Chang et al., 2014). Teachers on these platforms must deal with many administrative works including course organizing and real-time changes in data such as students' attendance during online classes as well as their tracked learning patterns (Baran et al., 2011; Kearns, 2012). Additionally, for teachers with little experience in online learning systems, assessing students based on real-time data and giving the final grades may be daunting. In traditional or blended offline classes, instructors still struggle with administrative tasks and assessment complexities. Managing course logistics, tracking attendance, and

* Corresponding author.
E-mail addresses: lampx@neu.edu.vn (P. Xuan Lam), maphquochung@gmail.com (P.Q.H. Mai), quanghung20gg@gmail.com (Q.H. Nguyen), thaop@neu.edu.vn (T. Pham), nthhanh.work.02@gmail.com (T.H.H. Nguyen), huyen.nguyenthi2@hust.edu.vn (T.H. Nguyen).

https://doi.org/10.1016/j.caeai.2024.100244
Received 11 October 2023; Received in revised form 14 May 2024; Accepted 15 May 2024
Available online 15 May 2024

analyzing students' learning patterns are challenges that educators face regardless of the mode of instruction (Cornelius & Gordon, 2008; Means et al., 2009; Picciano, 2021). Students' data are not yet collected appropriately and fully organized to generate insights into their performance and their grading scheme. Learners also lack information regarding their academic progress as the system's database is private, leading to little communication between them and educators and the resultant lack of measures for grade enhancement. Previous studies have utilized diverse methods to collect and process classroom data, such as using learning management systems for online classes to track engagement and participation and implementing cameras and eye-tracking devices in offline classes to monitor concentration (Henrie et al., 2018; Sharma et al., 2022). However, these approaches have limitations, including potential privacy concerns and the inability to capture the full scope of student engagement and cognitive processes.

With the significant advancement in the use of learning management systems in education, various approaches, such as Educational Data Mining (EDM) and Data Mining (DM), that incorporate machine learning have emerged. These approaches aim to utilize data for understanding learning styles (Hawk & Shah, 2007; Zajac, 2009), measuring and predicting learning behaviors and performance (Vandamme et al., 2007; Wook et al., 2009), as well as forecasting learning outcomes (Chi et al., 2008; Harvey & Kumar, 2019; Yağcı, 2022). For instance, Chi et al. (2008) employed Bayesian networks to predict student learning outcomes, taking into account factors like attendance and performance in class tests and assignments. In a study conducted by Yağcı (2022), various machine learning models, including Random Forest (RF), Neural Network (NN), LR, Support Vector Machine (SVM), NB, and k-Nearest Neighbor (k-NN), predicted final exam results with accuracy ranging from 70% to 75%. While there is substantial potential for this approach to build a predictive system, it necessitates the consideration of new proposals, improvements, and the identification of a compatible approach for the specific context.

This research aims to implement a predictive model that presents a well-rounded assessment of students and gives suitable final grades. Drawing from earlier research results, this model seeks to contribute to teaching methodologies and introduce a more effective grading scheme for online classes. The model can transform the current educational landscape, creating a more informed environment for educators and a collaborative space between learners and teachers. Data-based performance forecasting will ensure that students are well-informed of their learning progress thus taking full control of their learning, leading to more communication with teachers to develop effective learning strategies. Educators, in turn, will be incentivized to modify their approach to teaching and make more informed grading decisions.

The goal of this research is to tackle two questions:

- Question 1 (**Q1**): Whether and how can we predict students' performance in an early stage of a course?
- Question 2 (**Q2**): How different factors would impact the result of students' performance prediction?

The first question (**Q1**) addresses the feasibility and methodology of predicting students' academic performance at the outset of a course. It seeks to explore whether it is possible to make accurate forecasts regarding how well students will perform academically early in a learning journey. The question explores the methods and techniques employed to make these predictions. To answer this question, the process involves data collection, feature selection, model selection, data preprocessing, model training, model evaluation, and result interpretation. Educator feedback is solicited to gauge the practicality and impact of data-driven predictions on their decision-making. The second question **(Q2)** focuses on comprehending the numerous factors that affect students' academic performance forecasts. It explores how elements like demographic background, prior educational experiences, study habits, learning styles, and external situations might affect the accuracy of these predictions. The goal is to analyze and assess the importance of these factors in shaping predictive models and their results. This examination aims to clarify which variables hold greater significance in forecasting student success, potentially refining the models for improved accuracy across different student groups and educational environments.

2. Literature review

2.1. Traditional student assessment methods

Many studies have examined the impact of attendance on academic performance, however, most of them were conducted using a traditional statistical method and produced inconsistent results. Evaluating student participation is time-consuming for teachers, lacking automation and resulting in delayed assessments. This process traditionally involves observing students in the classroom, reviewing their contributions to discussions, and monitoring their engagement in group activities (Cohen & Goldhaber, 2016). Moreover, in the absence of automation in student participation assessment to track and assess participation data in real-time, educators struggle to pinpoint students in need of additional support. Early warning systems for struggling students are difficult to implement, making timely intervention challenging (Davis et al., 2013; Lee & Chung, 2019).

Latif and Miles (2013), using data from undergraduates taking business courses, majoring in Business and Economics Statistics at Thompson Rivers University, found that attendance significantly positively impacts midterm grades. However, the result of the study was not convincing since the authors also indicated that besides class attendance rate, other essential factors such as learning effort, assignment performance, and student background needed to be taken into consideration. The authors also stated that with a simple mean comparison, students with high-class attendance and a high number of completed assignments tend to perform better in the exam. Other studies also came to the same conclusion. In previous works, Romer (1993), Durden and Ellis (1995), Marburger (2001), Rodgers (2002), and Stanca (2006) concluded that attendance had a significant impact on student's grades or performance of students on examinations seems to be improved with increased attendance, even though the correlations between attendance and students' performance were relatively small. Halpern (2007) experimented on 127 students in a business management module and concluded that although attendance truly impacts student academic achievements, students' performance is largely determined by other factors such as entry qualifications, background, work, and course. This pushes for more stable and quantitative evaluation methods that produce higher accuracy. Taking advantage of Machine Learning techniques, many studies have appeared to implement predictive models capable of assessing student performance based on data relating to many variables such as student attendance, degree of attention in class, and academic background (Baashar et al., 2022; Romero & Ventura, 2010).

2.2. Data-driven assessment and predictive analytics in education

Data-driven techniques have been used for different applications including fraud detection, telecommunications, and banking (Han et al., 2012). These techniques analyze datasets and extract information to transform it into understandable structures for later use. Data-driven assessment and predictive analytics in education play an essential role in aiding teachers in using data to make grading decisions and developing efficient teaching strategies (Guan et al., 2020). Data-based evaluation is based on evidence that ensures the validity of the information collected. By extracting data as a foundation for making assessments, decision-makers are better informed and can gain more insights into a problem. Because data can be continuously updated and refined as new data becomes available over time, assessment results could be more meaningful than theory alone. Predictive analytics involves using data to make predictions (Guruler & Istanbullu, 2014). The

process makes use of data analysis, Machine Learning, and statistical models to locate patterns that might forecast future behaviors (Rastrollo-Guerrero et al., 2020). Through this, students' academic results can be predicted by analyzing the impacts of aspects such as their attendance, level of attention, and learning progress. Elrahman et al. (2023), researched a predictive model forecasting student performance in classrooms using their interactions with an e-Textbook. Using classification models to predict student performance (good or bad), they get an accuracy of 91.7% with the Random Forest Classifier (RF Classifier). They concluded that data from students' interaction with the device could be used in the development of the predictive model to intervene timely and improve students' educational outcomes. All features were important to the measuring of the final grades for students, hence their inclusion in the model. Therefore, they also stated that data mining is a very promising approach for predicting student performance.

2.3. Feature engineering

In previous studies, researchers chose variables to use for forecasting depending on each case and purpose. In general, the variables mentioned are usually data related to three categories: demographic factors, variables related to educational background and grades (midterm scores, grades from previous courses, grades from high school), and interactions in the classroom or learning platforms.

Fernandes et al. (2019) used student demographics and in-term grades to predict success with a Gradient Boosting Machine model. Previous year's grades and unattendance were the strongest predictors, while demographics like neighborhood, school, and age showed potential as well. Several studies indicate that gender plays a role in academic performance, with females traditionally demonstrating higher achievement (Mensah & Kiernan, 2010; Sánchez et al., 2019). However, this trend varies depending on the specific scientific field. Waheed et al. (2020) found that demographic information and clickstream activity significantly influenced academic performance, with more active course engagement leading to higher results. Interestingly, participation in the learning environment itself did not directly impact outcomes.

Several earlier studies have used variables indicating student engagement in the classroom or online learning platforms as a key factor in predicting grades (Bernacki et al., 2020; Chango et al., 2021; Elrahman et al., 2023; Liu et al., 2020). These factors, from a certain perspective, indicate students' concentration while in class and can also evaluate their diligence. Leveraging a wealth of data from the first week of a MOOC, Jiang et al. (2014) built prediction models relying heavily on performance indicators like quizzes and peer assessments, as well as student engagement in forum discussions. Bernacki et al. (2020) investigated whether student behaviors recorded in a learning management system (LMS) alone could predict academic performance. They considered variables such as clickstream data, quiz attempts, use of monitoring tools, and time of accessing planning resources. Their model successfully found that 75% of students are at risk of repeating a course based solely on their LMS activity.

For the task of predicting student academic performance, the inclusion of grade variables is indispensable. Yağcı (2022) used a dataset that included 1854 students in 12 different faculties who have taken Turkish Language-I courses. With the data variables including midterm exam grades, Faculty, and Department that aim to predict student final grades, he concluded that midterm grade is a vital predictor to be used in predicting the final grade. There is also a remarkable number of models that use data from students' performance in an early period of the course (Chango et al., 2021; Harvey & Kumar, 2019; Lu et al., 2018), and these factors are quite stable for making early predictions. To forecast student success, Harvey and Kumar (2019) examined a predictive model that analyzed a K-12 education dataset including both demographics and exam grades data, Fernandes et al. (2019) developed a model leveraging both demographic data and in-term activity grades. Meanwhile, Mengash (2020) considers pre-admission test scores to predict the possible grades of the future student, aiming to support the university in admission decisions for their future undergraduate students. This is quite an interesting study given that there are quite a few articles on this topic that examine the correlation between entry grades and academic grades. However, a major pitfall of early warning systems is that their data, often including course grades, can be overly entangled with the very outcome they aim to predict future grades. This can lead to models merely reconfirming the obvious, suggesting struggling students will continue to struggle if no interventions are implemented (Bernacki et al., 2020).

2.4. Predicting student performance

In terms of predictive analytics, Harvey and Kumar (2019) conducted a study to weigh up the performance of different Machine Learning models in predicting grades in K-12. The study compared the efficiency of three models: Naive Bayes (NB), Logistic Regression (LR), and Decision Tree for predicting high school student Scholastic Assessment Test (SAT) math scores. The conclusion was that NB was the most effective model with 71% accuracy. Educators could rely on this model to evaluate the correlation between several variables and student test scores and accordingly make predictions. It could also be utilized to implement early interventions for students with low academic results, leading to the improvement of their grades. In an effort to predict students' final exam grades, Yağcı (2022) used the Machine Learning models, e.g., Random Forest (RF), Neutral Network (NN), LR, Support Vector Machine (SVM), NB, and k-Nearest Neighbor (k-NN) were employed to predict the classified final exam, performing from 70% to 75% accuracy, with the highest for RF and NN models. Remarkably, the accuracy rate of the RF Classification model of the recent research on the same topic is relatively high. With the same attempt to evaluate and predict students' performance and achievement, Asif et al. (2017), Bernacki et al. (2020), Mengash (2020), and Liu et al. (2020), experimented with diverse types of input variables. Asif et al. (2017) used the grades of all courses in four-year degree programs to predict the academic achievement of students and gain the highest accuracy of the NB algorithm up to 83.65%. Mengash (2020) proposes that we can predict in advance whether students will achieve a good CGPA by Artificial Neural Network (ANN) at around 80%. Bernacki et al. (2020) Use behavior-based models extracted from user activity logs in the Blackboard Learn learning management system. Using LR, NB, J-48 Classification (J-48), Decision Tree (DT), and J-Rip DT to identify whether a user may earn poor grades based only on the recorded behavior of the user on the Learning Management System (LMS), the result reveals that about 67.36% of cases are classified accurately with the LR algorithm. From the first evaluation, by only using the data related to learners' behaviors, this approach gained an accuracy rate that is relatively low compared to other studies. Liu et al. (2020) also use data from a learning management system. Still, in this study, instead of focusing on learner behavior, they consider factors that are more related to learning rate such as how much time students spend answering questions, how many points students get for each practice question, and levels of assignments students complete to assess student grades. What is different about this study is that it focuses on using deep learning applications using Recurrent Neural Networks (RNN), Bayesian Knowledge Tracing (BKT), and Long Short-Term Memory (LSTM). The accuracy of this study is quite impressive, up to 97% with MFA-DKT (Multiple Features Fusion Attention Mechanism Enhanced Deep-Knowledge-Tracing).

Several other studies were conducted to experiment with early warning systems for student learning behavior (Bernacki et al., 2020; Chen, 2022; Lee & Chung, 2019). In these systems, from the input data, researchers often figure out how to classify students who are likely to pass, fail, and drop out using Machine Learning models. Similar to the studies that have attempted to predict student grades, scholars in this sub-topic have also attempted to use a variety of input variables to classify the students. Hoffait and Schyns (2017) used demographic

factors including Gender, Nationality, and background information including Major, Prior schooling, Math level, and Scholarship to identify first-year students with a high risk of failure. Using Machine Learning algorithms (RF, LR, ANN), it is concluded that the students facing academic difficulties can be identified with a high level of confidence (90%) as a result of the RF algorithm. Waheed et al. (2020) created a model using artificial neural networks to analyze student records about their use of the LMS. The findings indicated that student performance was significantly impacted by demographic factors and clickstream behaviors. In another study, students' performance is predicted using video data from face-to-face interaction, grades, practical class participation count, attendance and grades in online classes via the Moodle platform, and final exam scores (Chango et al., 2021). J48, REPTree, Random Tree, JRIP, Non-Nested Generalization (NNGE), and PART were the algorithms employed. The NNGE and PART algorithms handle numerical and discretized data with the greatest levels of accuracy. It was discovered that participation in Moodle discussion forums, quiz scores, and degree of concentration in theory class were the factors that predicted final performance the most accurately.

The research studies on predicting student performance that we have compiled are presented in Table 1.

3. Methodology

3.1. Research workflow

Our research workflow, consisting of four steps, can be described as follows (see Fig. 1):

- **Data acquisition**: Gather data from the system
- **Data cleaning and Normalization**: Clean and normalize the dataset, addressing missing values, outliers, and inconsistencies to ensure data quality and consistency.
- **Exploratory data analysis:** Examine the dataset using statistical and visualization techniques.
- **Model application and Evaluation**: Apply Machine Learning models to predict students' grade ranges and evaluate model performance using various metrics.

3.2. Data collection

The data used for this study gives information on various aspects of students' learning progress in five different classes provided by VnCodelab website (https://vncodelab.com). VnCodelab is an online learning webpage, which provides courses on IT-related subjects through learning materials in the forms of text, code, images, and videos formatted similarly to slides. Teachers can create learning materials using only Google Docs, such as interactive slide quizzes. Each slide can be automatically or manually stored, copied, and organized by teachers for each intended class. Students can follow the lesson content on the slides while following the lecturer face-to-face. Students can join an appointed class, using slides, answering quizzes, raising their hands, and communicating with others (in the top right corner). The interface of the VnCodelab webpage is shown in Fig. 2. Every action of students and teachers on learning material will be captured and published in real-time, so both students and teachers can keep track of each other and make changes during class if necessary (the number of participants is displayed on each slide in the left-hand bar and can show individuals name by hovering on it). There is an online compiler environment that lets students run their code directly on the website. All student behavior data is stored on the Firebase database. During the COVID-19 outbreak, such benefits from VnCodelab brought great support for lectures in delivering information and managing students without any physical barriers.

The data was gathered from 253 undergraduates, specifically second and third-year students majoring in computer science, and they have almost the same demographic characteristics (regarding nationality and age). These students were chosen because they shared similar academic backgrounds, having taken roughly the same foundational classes prior. They were enrolled in five classes offered in Web Design, Web Technologies, and Java Programming courses during a semester. Two primary metrics were recorded for each student from the database: their total time dedicated to the system throughout the semester and their time spent reviewing, indicating the periods when students logged into the system beyond regular class hours for review purposes. Demographic information such as gender and age were also collected. Moreover, the dataset included variables that highlighted students' diligence (which will be discussed later) and academic performance indicators such as final exam results, midterm exam scores, average scores, previous subject averages, entrance scores, and high school GPA. This pool of undergraduates was chosen for analysis due to their shared academic pursuits, providing a coherent basis for evaluating their study habits, performance, and related factors within the realm of computer science education. For illustration, Table 2 displays a sample of the initial dataset.

3.3. Data cleaning and normalization

We have made enhancements to four categories in our dataset: **Time interval, Submit rate, Attendance score, and review score**.

Table 1
Summary of some previous works on predicting student performance.

Variables	Prediction	Dataset size	Algorithms	Highest accuracy	Reference
15 characteristics of interactions of the student with an e-Textbook	Student performance	200	RF, DT, SVM, LR, k-NN	RF (91.7%)	Elrahman et al. (2023)
Student demographics, teacher salaries, school finance, grades	SAT math scores	403	NB, DT, Linear Regression	NB (71%)	Harvey and Kumar (2019)
Midterm exam grades, Faculty, and Department	Final exam grades	1854	RF, LR, SVM, k-NN, NB, NN	RF (74.6%), NN (74.6%)	Yağcı (2022)
Grade of all courses in four-year degree programs	Student academic achievement	210	DT, 1-NN, NB, NN, RF	NB (83.65%)	Asif et al. (2017)
Log records in LMS	Student achievements	337	LR, NB, J-48, DT, J-Rip DT	LR (67.36%)	Bernacki et al. (2020)
Gender, nationality, prior schooling, math level, scholarship	Students at high risk of failure	2244	RF, LR, ANN	RF (90%)	Hoffait and Schyns (2017)
Pre-admission test scores (HSGA, SAAT, GAT)	Student CGPA	2039	ANN, DT, SVM, NB	ANN (79.22%)	Mengash (2020)
Scores and duration of practice questions, type of exercises	Student performance	4216	DKT, RNN, BKT, LSTM	MFA-DKT (97%)	Liu et al. (2020)
Students' demographics, clickstream events	Pass/Fail, Withdrawn, Distinction	32593	ANN, SVM, LR	ANN (93.23% for Pass-Fail)	Waheed et al. (2020)
Face-to-face interactions, midterm & final scores, online grades	Fail – Pass – Dropout students	57	J-48, DT, RF, JRIP, NNGE, PART	NNGE (80.46%), PART (80.46%)	Chango et al. (2021)

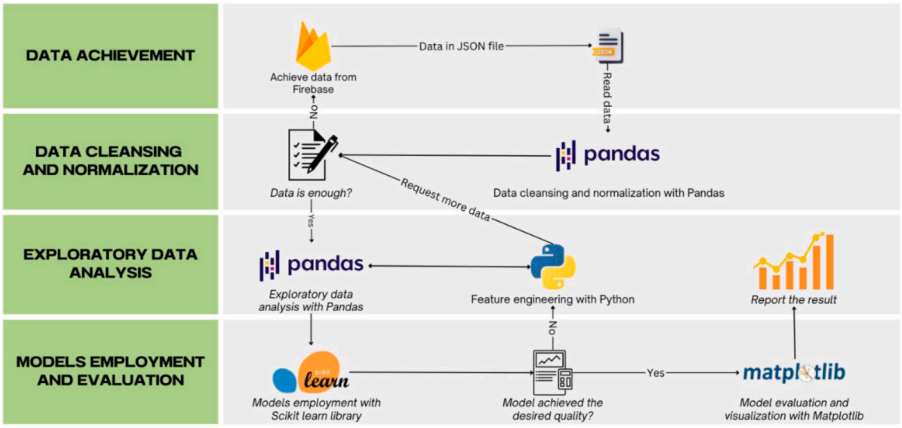

Fig. 1. Research workflow.

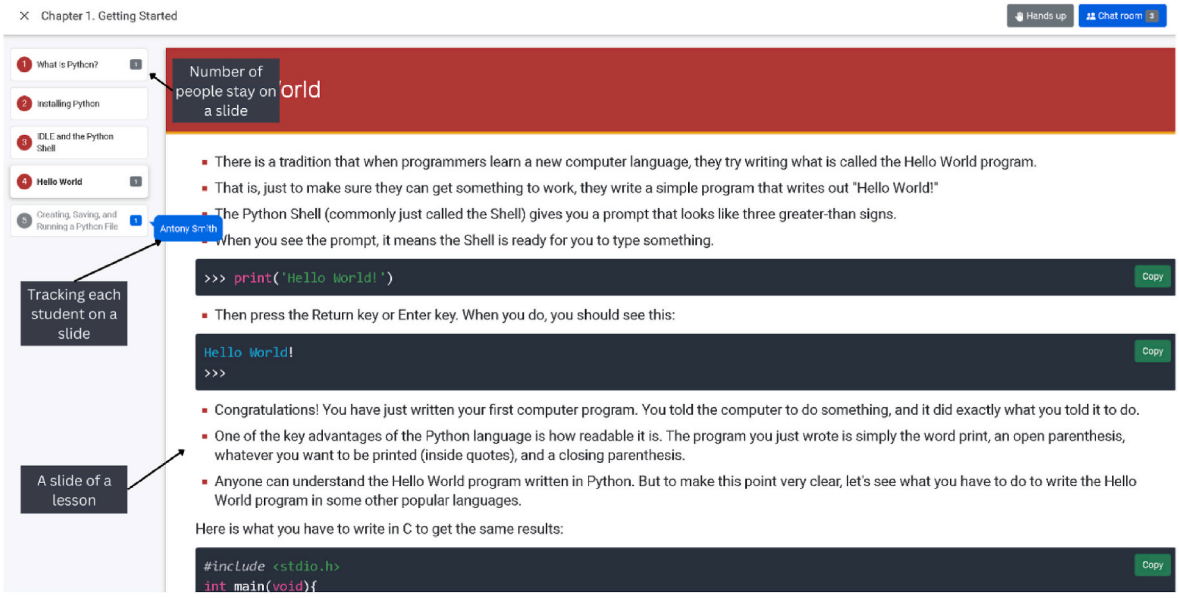

Fig. 2. Demonstration of VnCodelab interface during a lesson.

Table 2
Example of dataset.

Student id	Subject	Review	Attendance	Midterm	…	Age	Average score
1119****	Web Design	0	6	6	…	20	7
1121****	Java	14356	9	8.5	…	20	7.1
…	…	…	…	…	…	…	…
1121****	Web Design	1514	10	8.5	…	20	6.75
1121****	Web Design	8874	10	9	…	20	9.75

Time interval: The research is interested in students' level of concentration on the lecture in a class. We derived this figure by dividing the total time the lecturer spent on the whole course by the total amount of time that students were on the same slide with the teacher. However, we encountered a problem in standardizing the time interval. As the data was stored in a real-time database (the system uses Firebase to record the data), the time interval of each user was unclear and overlapped with that of others. Hence, there came a need to conceive a novel approach to refine the time logs (see Fig. 3). Firstly, time stamps with a duration lower than a certain duration were removed (in this case, we chose 90 s as the lower bound), but the overlapping was still present. The solution was to only record the times in which the users entered the slide and carefully chose the duration so that the end of the previous slide would be the start of the next slide. As a result, the data was much cleaner with few interruptions and duplicates as the original (see Fig. 3). The final perfected time intervals were able to capture up to 85% of the actual time, thus preserving the core data and affecting minor changes to the results. After refining the time intervals, we calculated the mean student focus rate by the approach that has been proposed above.

Submission rate: This is the ratio of the questions/exercises the student has answered to the total amount of questions/exercises in the lectures.

Submission rate and Attendance score: It is normal to have a difference between students in those categories. However, the average

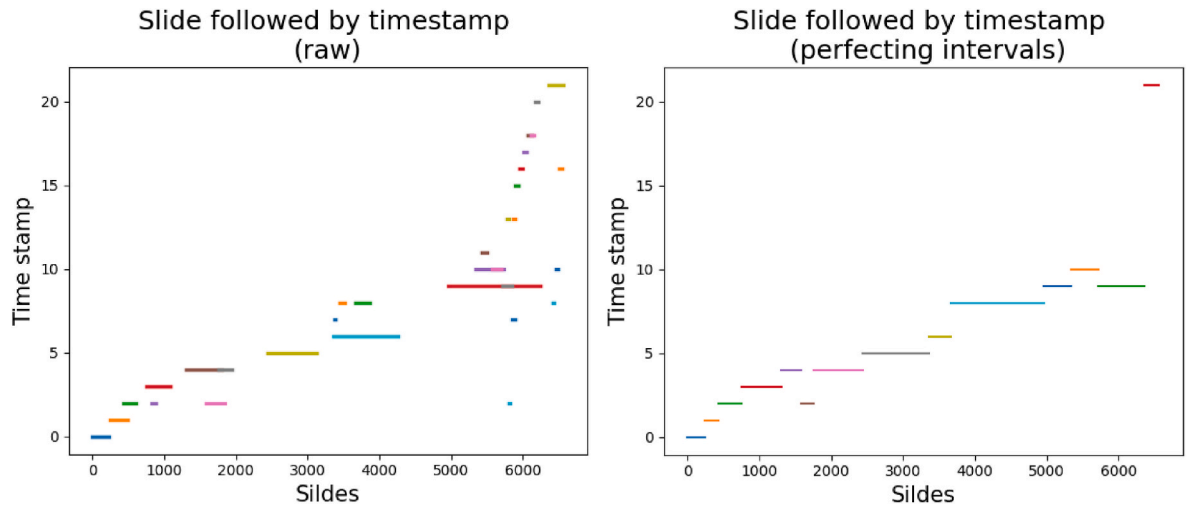

Fig. 3. Example of student time interval before and after refining.

score of each class sometimes has a significant gap (see Table 3).

It may be caused by the grading method of each lecturer and the design and necessity of each question in the materials of each course. For instance, a slide has a set of 10 questions, and students only need to answer 5 of them during class to receive a full point (1.0). This can lead to situations where students achieve the maximum score of 0.5 even though they answered all required questions. Unfortunately, we could not extract which class has the limited set of questions, so we decided that for every class that has a lower score, a certain point (we chose 0.2 but not exceed 1.0) is added to everyone in that class. Lecturers' grading methods also affect the average score the same way and we also add certain points (1.0) to each student. After the update, we have the results in Table 5.

Review time: This feature represents the total time students spend on the system outside of class hours, likely reviewing materials or doing additional study.

Review score: The original measurement of review time is the cumulative sum of every student log of each slide that is not during lecture hours. However, some students have "googol" hours on the system (see Fig. 6) but their performance only has a minimal difference in comparison to others. Therefore, we decided to use a logarithm with base e to demonstrate the students' review time (in hours). The scores are measured by the following function.

$$Review\ score = ln\left(\frac{review\ time}{3600} + 1\right)$$

After getting some basic data from students, we wanted to extract added possible data about more features of the Machine Learning model. Given the focus of the research, we have chosen the most suitable features that show students' diligence used for Machine Learning implementation. These features would play a key role in predicting students' final grades:

- **Participation rate:** measures the percentage of students attending a course during the semester.
- **Consistency rate:** measures how consistently students attend classes during specific periods of the semester. Lower consistency rates indicate less consistent attendance and attention. There are two more features to record students' rate to attend classes with the consistency rate in the first half of the semester collecting data from this period specifically and the consistency rate in the second half of the semester using only data in the corresponding period.
- **The following rate:** calculates the ratio of the total time a student spends on the same slide as the teacher to the total lesson time. It helps gauge students' engagement with the lecture material.
- **Immediately following rate:** count the total number of seconds that a student changed the slide 1 min after the teacher went to the next slide. This could prove that students still paid attention to the lecturers even if they were not viewing the current slide. The system incorporated this feature because relying solely on the "Following Rate" metric may not fully capture students' learning behaviors. Some students may prefer to preview upcoming content by skipping to later slides or review their understanding by revisiting previous slides during each lesson.
- **Compulsory:** This categorical feature encodes whether the course is relevant to each student's major. It distinguishes between compulsory, elective, and optional courses, providing information about students' choice of courses.

Table 3
The average score of five classes before normalization.

Class	Participate	...	Submission	Review	Attendance	Average
1	0.771	...	0.237	2444.860	7.820	7.621
2	0.734	...	0.366	7092.913	8.689	8.045
3	7.735	...	0.257	7200.267	9.107	8.20
4	0.870	...	0.575	2613.878	9.090	7.695
5	0.736	...	0.532	27628.50	8.586	7.735

Proper feature selection can affect the performance of models significantly. In total, we collected a dataset of 253 observations in 18 unique features, categorized into three groups: **diligence**, **demographic information**, and **academic background**. Among them, some are the cumulative sum or transformation of others, so we decided to use 13 key features as input for classification models, which are: **Consistency**, **Following**, **Review score**, **Attendance score**, **Midterm exam score**, **Compulsory**, **Average Score in other subjects**, **Entrance exam score**, **Age**, **Gender**, **Highschool GPA**, **Homework Submission rate**, and **Subject**. Each of the selected features affects the model performance differently, and how significantly a certain feature impacts models will be discussed in **4.3.**

3.4. Model selection and evaluation method

To tackle the problem of predicting students' average grades in the course, different statistical measures and machine-learning models were chosen. Since the final predictions would be students' average scores, we could use classification models. The whole dataset includes data from 253 students, and it is split into two parts: 20% for the test set and 80% for the training set. For the classification task, since the collected data

size is insufficient to classify scores from 0 to 10, we choose the approach to classify their score range, whether it is **A** (Score≥8.5), **B** (7≤Score <8.5), **C** (5≤Score <7), or **F** (Score <5). Table 4 depicts the range for each grade. Five models were implemented to compare the accuracy and using Stacking Classifier to combine five proposed base models. The base classification models are Multinomial Logistic Regressor (also called SoftMax Regressor), Random Forest Classifier, Naive Bayes, Support Vector Machine, and k-nearest Neighbors. The performance of the model was measured by confusion matrix indicators.

After the data has been compiled and cleaned, we proceed to build a Machine Learning pipeline steps to perform experiments. Figs. 4 and 5 compare different Machine Learning pipelines, according to the following steps:

- **Data preprocessing**: for every base model, categorical and numerical data is processed before fitting to the model. To be more specific, the numerical features undergo a standard scaler then use k-NN imputer to autofill into missing values, and One-hot Encoder is implemented for categorical data.
- **Machine Learning model for classification:** To classify students into different grade ranges, we employ some of the most common models for classification, such as RF, LR, SVM, NB, and K-Neighbors Classifier.
- **Stacking Classifier**: The Stacking Classifier is a popular ensemble learning technique to combine different models, with an effort to leverage the overall performance of base models. Stacking combines multiple classifiers to induce a higher-level classifier with improved performance. The Stacking Classifier works by firstly training the preprocessed data with base models, and then using predictions of base models to train the final estimator (Dietterich, 1997; Džeroski & Ženko, 2004; Wolpert, 1992). In this study, we chose the Support Vector Machine as the final estimator, which would give us the best overall performance.
- **Setting base models for Stacking Classifier**: A set of base models is selected to be used in the Stacking Classifier, the number of base models should be diverse so that it would evaluate and learn various aspects of the dataset. In the scope of this study, we chose 5 different classifiers including RF, LR, SVM, NB, and k-NN.

3.5. Research tools

The experiment phase of this study was performed with the Scikit-learn library using Python programming language. Scikit-learn is a free and open-source Machine Learning library for the Python language (Varoquaux et al., 2015). It is one of the most used libraries in terms of employing Machine Learning models for predictive problems with a wide range of supervised and unsupervised learning algorithms, as well as tools for data preprocessing, model evaluation, and visualization. Hence, Scikit-learn using Python programming language came as a convenient tool in the scope of this study.

4. Results

4.1. Data preparation and initial analysis

Some parameters of classes were normalized. The newly obtained dataset is presented in Table 5. Normalizing the dataset resulted in noticeable changes in the average submission range for Classes 1, 2, and 3. Additionally, Class 1 experienced a slight increase in both average attendance score and average grade post-normalization. Fig. 6 shows the histogram graphs of four features including **Consistency, Following rate, Submission rate,** and **Review time**. Fig. 7 shows the number of students achieved in four grade ranges. 131 students achieved B, followed by A (81 students) and C (38 students), with only 3 in F.

Table 4
Grade range criteria.

Grade range	Criteria
A	Score≥8.5
B	7≤Score <8.5
C	5≤Score <7
F	Score <5

Table 5
Average score of five classes after normalization.

Class	Participate	…	Submission	Review	Attendance	Average
1	0.771	…	0.428	0.358	8.660	7.705
2	0.734	…	0.366	0.920	8.689	8.045
3	0.859	…	0.453	0.888	9.107	8.200
4	0.870	…	0.575	0.428	9.090	7.695
5	0.736	…	0.532	1.988	8.586	7.735

By incorporating these variables into our model and subsequently analyzing their impact, we strive to address **Research Question 2** with greater depth and accuracy (further discussed in **4.3**). Firstly, we consider the variables that contribute directly to the final average grades given that we have known the midterm exam score, this would include attendance score and midterm score. Secondly, we take in the characteristics of the subject to students, including the subject title and whether it is compulsory. And thirdly are parameters showing the students' concentration and diligence in the subject, those are **Consistency**, **Following rate**, **Homework submission rate**, and **Review score**. Next, we consider the performance of students in the earlier period (**Pre-admission exam score**, **High school GPA**, and **Average score in different subjects**) and finally personal characteristics (**Age** and **Gender**). Table 6 shows the data size of each parameter, the number of null values, the percentage of missing values, the data type of the feature, and the category of that feature (with background abbreviated for academic background). The target column is the average grades converted to grade ranges (A, B, C, and F) for classification algorithms.

4.2. Evaluation of models

The dataset mentioned above includes 13 main features that are taken into consideration for predicting students' grade ranges (see Table 6). The classification models are then fitted to the data and then evaluated with different criteria: confusion matrix, classification accuracy (CA), precision score, recall score, F1 score, and ROC curves score to compare different models.

Interpreting the results and evaluations of models in Table 7, we can see that the Stacking Classifier gains a relatively high proportion of predicting the grade range of students correctly, which is 0.826 accuracy and 0.792 F1. This F1 score and accuracy rates are, as expected, the highest among the models since it is the ensemble of the base models. Regarding the confusion matrices of the Stacking Classifier in Fig. 8, the main diagonal of those matrices forms a bold line, indicating that the model gains a high accuracy level. More particular, while the confusion matrix on the test dataset gained an accuracy of 0.826 the accuracy rate on the whole dataset witnessed a higher value (see Fig. 9), which is 0.834. It is also interesting to note that the SVM base model gained a high result is 0.804 in accuracy and 0.805 in Precision score. The SVM model is also one of the most used algorithms for the classification tasks in education for the same purposes in previous studies (Rastrollo-Guerrero et al., 2020).

Other models also performed well, with the RF model having the second-best performance overall among base models. This means that in some cases the RF model may be more useful and give a more accurate classification as in the research of Yağcı (2022), and Hoffait and Schyns (2017) where RF is the best model among all basic Machine Learning methods. Compared to the results of Yağcı (2022), Mengash (2020), and Asif et al. (2017) the result of the Stacking Classification is higher than

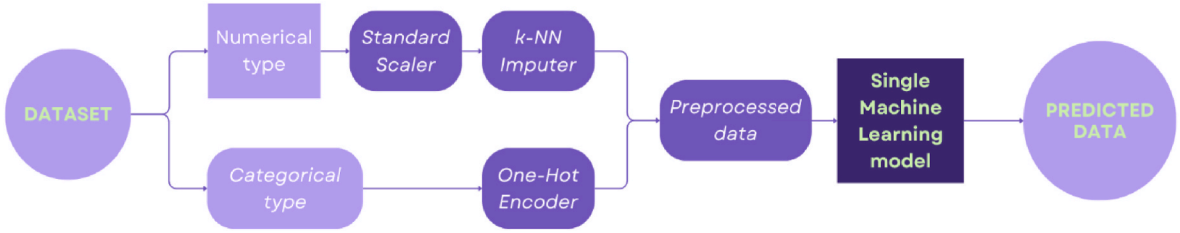

Fig. 4. Machine Learning pipelines for normal models.

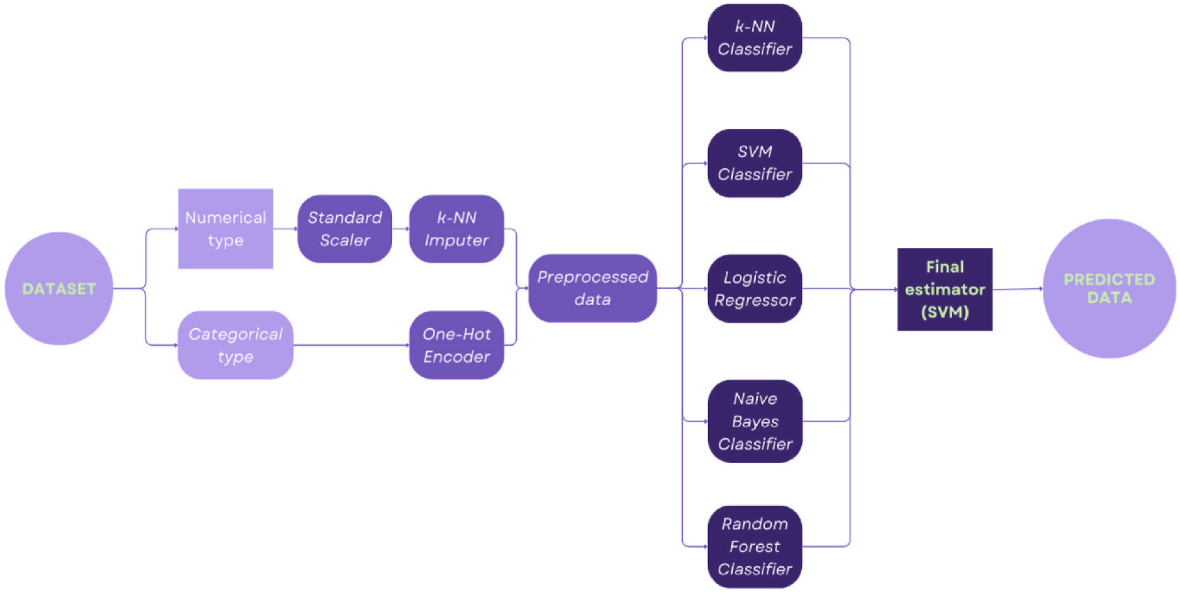

Fig. 5. Machine learning pipeline for stacking classifier.

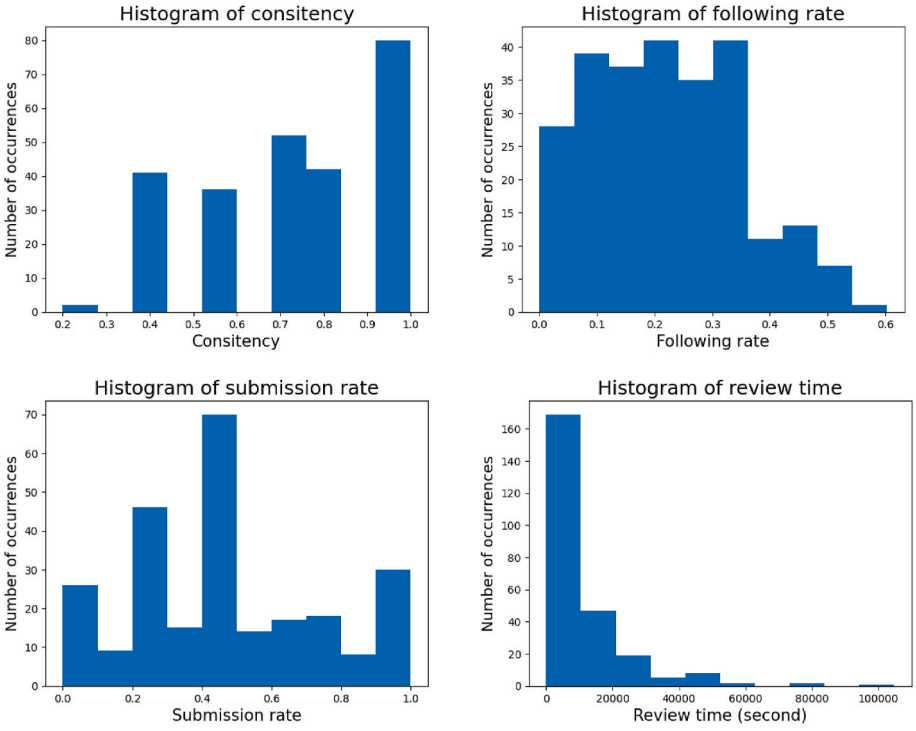

Fig. 6. Histogram graphs of some features.

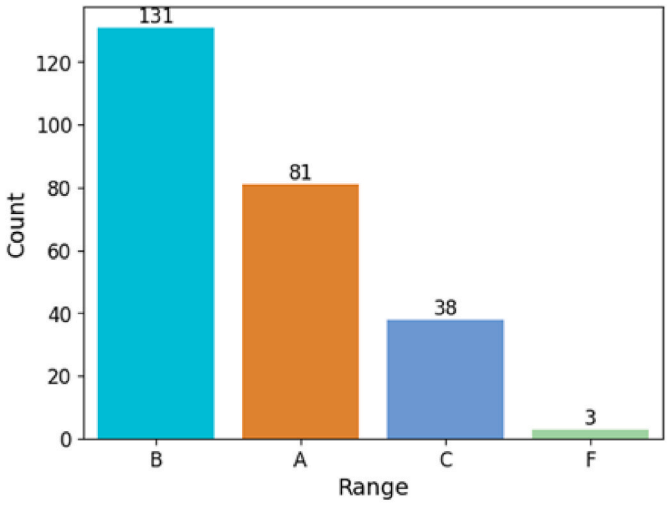

Fig. 7. The number of students in each grade range.

Table 6
Variables of model.

Feature	Null value	%Null value	Data type	Category
Consistency	0	0	Numerical	Diligence
Following rate	0	0	Numerical	Diligence
Review score	0	0	Numerical	Diligence
Attendance	0	0	Numerical	Diligence
Midterm exam	0	0	Numerical	Background
Compulsory	0	0	Numerical	Subject
Average of different subject	42	16.60%	Numerical	Background
Pre-admission exam score	142	56.34%	Numerical	Background
Age	0	0	Numerical	Demographic
Gender (Male = 1)	0	0	Binary	Demographic
Highschool GPA	144	57.14%	Numerical	Background
Homework Submission	0	0	Numerical	Diligence
Subject	0	0	Categorical	Subject

Table 7
Scores of different models.

Model	Accuracy	Precision	Recall	F1
RF	0.784	0.790	0.741	0.768
NB	0.745	0.741	0.659	0.685
SVM	0.804	**0.805**	0.765	0.783
k-NN	0.667	0.770	0.513	0.536
LR	0.784	0.797	0.705	0.739
Stacking	**0.826**	0.792	**0.825**	**0.792**

the models of this research and is more robust with the relatively same task of predictions. However, one of the limitations of earlier articles is that either they make quite simple predictions such as pass or fail, or they also try to predict students' average scores range, but the accuracy rate is relatively low. For example, Yağcı (2022) tried to predict students' final grade range taking into consideration their Faculty, Department, and Midterm exams achieved the best result is 0.764 accuracy. However, it's important to note that this accuracy is calculated on the whole dataset, not the test dataset only. Taking another example, in the research of (Hashim et al., 2020), the accuracy of LR was highest at 0.89 but for the task classifying pass or fail students.

Fig. 10 visualizes the ROC curves for 5 base models and the Stacking Classifier combined model. Whether the ROC curves are closer to the top-left corner of the graph, the better that classifier is. At first look, the Stacking Classifier has a better overall score than the rest. When comparing with the k-NN and SVM models, the AUC score of the Stacking model is much higher. It is also interesting to note that the SVM and LR models tend to perform well in predicting grades "A" and "C", with around 0.92–0.94 AUC scores. However, the Stacking Classifier is better at predicting grades "A" and "B", these two grades have a much higher proportion compared to grades "C" (see Figs. 8 and 9). Where most of the grades are in the range "B", and the Stacking Classifier has a good performance in classifying this grade range (AUC = 0.83), this result is much higher than the base models. Hence, in general, the Stacking Classifier works the best so far by combining five base classification models.

Unfortunately, interpreting Table 8 for Overfitting analysis, some base models show signs of being overfitted. Specifically, the RF model reached an accuracy of 0.985 on the training set, yet the F1 score was only 0.738. Similarly, the LR model displayed an accuracy and F1 score of 0.762 and 0.570, respectively, and the indicator for the SVM model highlighted an accuracy of 0.747 on the training set, accompanied by an F1 score of 0.560. This happens because for such small datasets like this, models like RF or SVM tend to be overfitted. However, while models like RF and SVM showed signs of overfitting, the Stacking Classifier stood out with satisfactory performance (0. accuracy and 0.823 on F1). This shows the impressive capability of stacking classifiers to surpass the limitations of their parts by combining the knowledge from its various base models, the stacking classifier effectively reduced overfitting. This combined approach acted as a defense against memorized noise and imbalanced datasets. Moreover, the stacking model avoided overfitting since we implemented a robust 3-fold cross-validation strategy. This not only contributed to superior generalization but also resulted in achieving accuracy beyond 82% on new data. Many other research papers have also adopted this approach to mitigate overfitting in their base models (Biswas et al., 2023; Shi et al., 2023; Wu et al., 2021). In future studies, to mitigate overfitting in base models and improve overall performance, we plan to augment the dataset by incorporating new data collected from more students.

4.3. *Impact of variables on the models*

Fig. 11 presents a feature correlation matrix, visualizing the pairwise relationships between variables within the dataset. The **Subject** feature is not presented in this correlation matrix since it is a one-hot encoded variable. Warmer colors (red shades) denote positive correlations, indicating direct proportional relationships between features. Conversely, cooler colors (blue shades) imply inverse relationships. The color intensity scales with the correlation strength, with darker hues representing stronger associations. Notably, certain cells appear white, denoting missing data (Nan values) in those fields.

Three variables stand out with the strongest correlations to student performance (denoted as Range in Fig. 11): attendance score (0.407), the midterm exam (0.628), and the average of different subjects (0.713). Conversely, gender and compulsory scores show the weakest correlations. This suggests, in response to **Research Question 2**, that the midterm score, attendance, and previous course scores likely have the most significant influence on the models, while gender and compulsory scores have minimal impact on student performance.

To directly compare the performance decrements when individually removing variables with varying levels of correlation, we systematically dropped each variable from the dataset and evaluated the resulting model performance using classification models. The detailed results are presented in Table 9, facilitating a clear comparison between the variables with the strongest and weakest correlations to student performance. We saw a significant drop in accuracy when excluding the average of different subjects (0.726 – declined by 0.101), followed by the midterm exam (0.765) and attendance (0.784). In contrast, removing gender and compulsory led to almost negligible changes: a 0.022 decrease in accuracy for compulsory and virtually no change for gender.

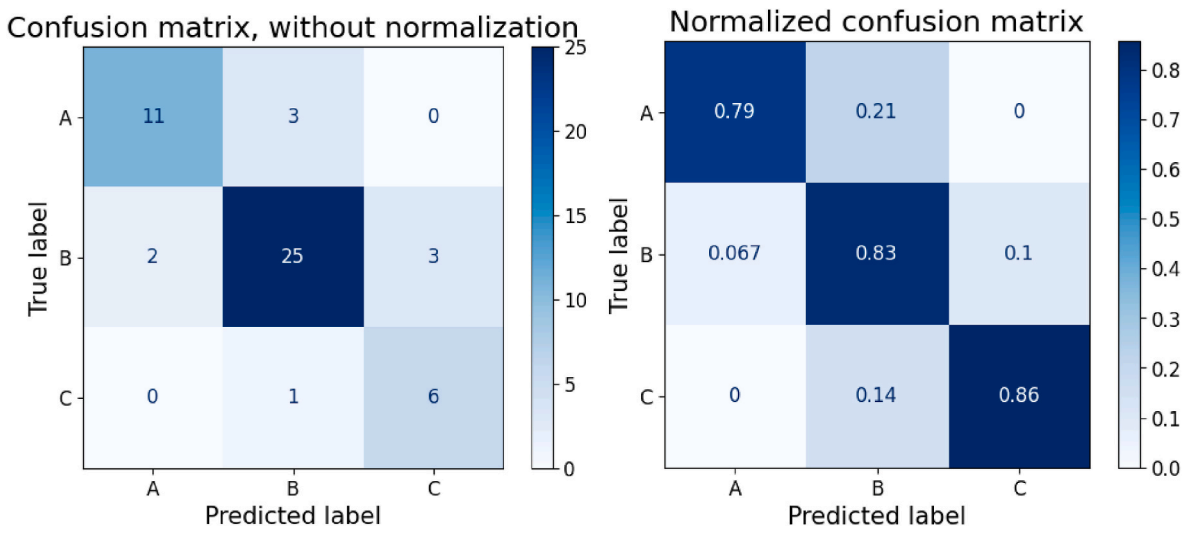

Fig. 8. Confusion matrices of Stacking Classifier on test dataset (CA = 82.63%).

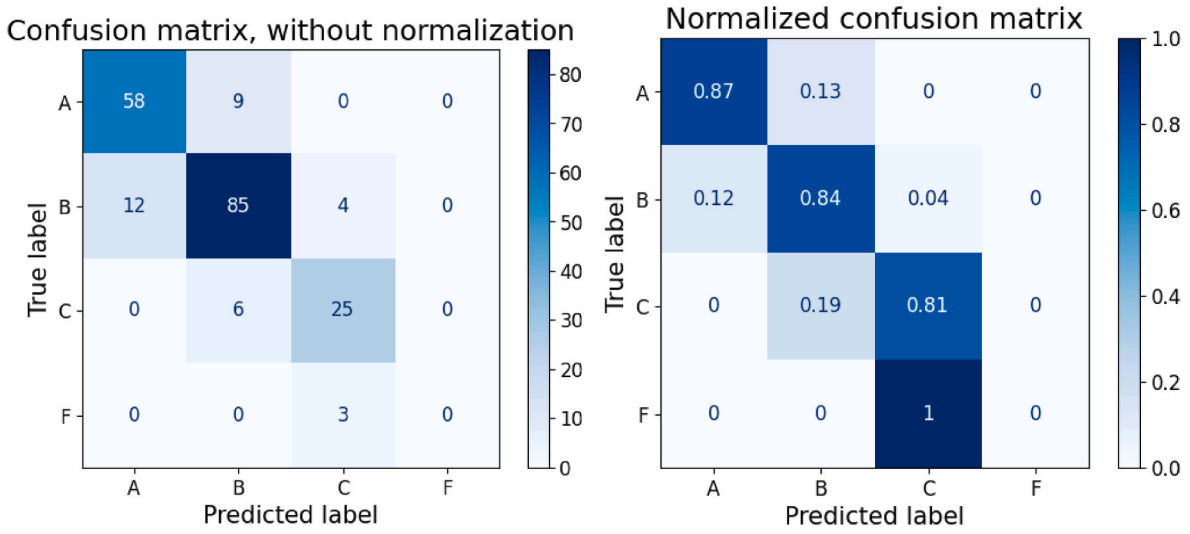

Fig. 9. Confusion matrices of Stacking Classifier on the whole dataset (CA = 83.79%).

5. Discussion

5.1. Research findings

The study has proposed a method to predict students' final course average scores from an early stage of the course, addressing **Question 1** regarding whether and how we can predict students' performance. To provide further detail, the experiment includes the use of five base classification models RF, NB, k-NN, LR, and SVM and then using the prediction of the base models to train the final estimator, which is SVM with a technique called Stacking Classifier. Notably, the Stacking Classifier emerged as the most accurate model, leveraging the strengths of multiple base models to enhance predictive performance. This approach effectively mitigated overfitting and yielded robust predictions.

Our findings align with those papers of Fernandes et al. (2019) and Xu et al. (2017), who identified prior academic performance as a key predictor of student success. Similarly, the aggregate score of various subjects appears as the strongest predictor in our study, highlighting the cumulative impact of past academic endeavors. Furthermore, the attendance score and midterm exam scores directly contribute to the predictive power of the average score, as they make up 10% and 30–40% of the average grade value, respectively. This highlights their vital role in assessing student performance within the course. Notably, earlier research reinforces the major influence of attendance, assignment, and midterm scores (Arashpour et al., 2023; Riestra-González et al., 2021; Waheed et al., 2020), mirroring our observation of the strong predictive power of the aggregate score, which incorporates these elements.

On the other hand, our findings show that gender has minimal impact on the model, which resonates with the conclusion of Alsulami et al. (2023), suggesting that individual student characteristics may hold insufficient correlation with student performance due to the relative homogeneity of the student population in terms of these characteristics. Similarly, whether the course was compulsory or not also has negligible impact, potentially showing that intrinsic interest or motivation plays a more prominent role in determining student success than external factors.

The Stacking Classifier's high accuracy and other scores indicate that this model is reliable. What makes this model robust and stable is that it considers a wide range of variables of students including diligent variables, academic background, and their characteristics. Hence, the experiment gains a higher accuracy rate compared to other models that include fewer features such as in the study of Hoffait and Schyns (2017) where the authors mainly consider demographic factors and academic background (Gender, Nationality, Prior schooling, Math level,

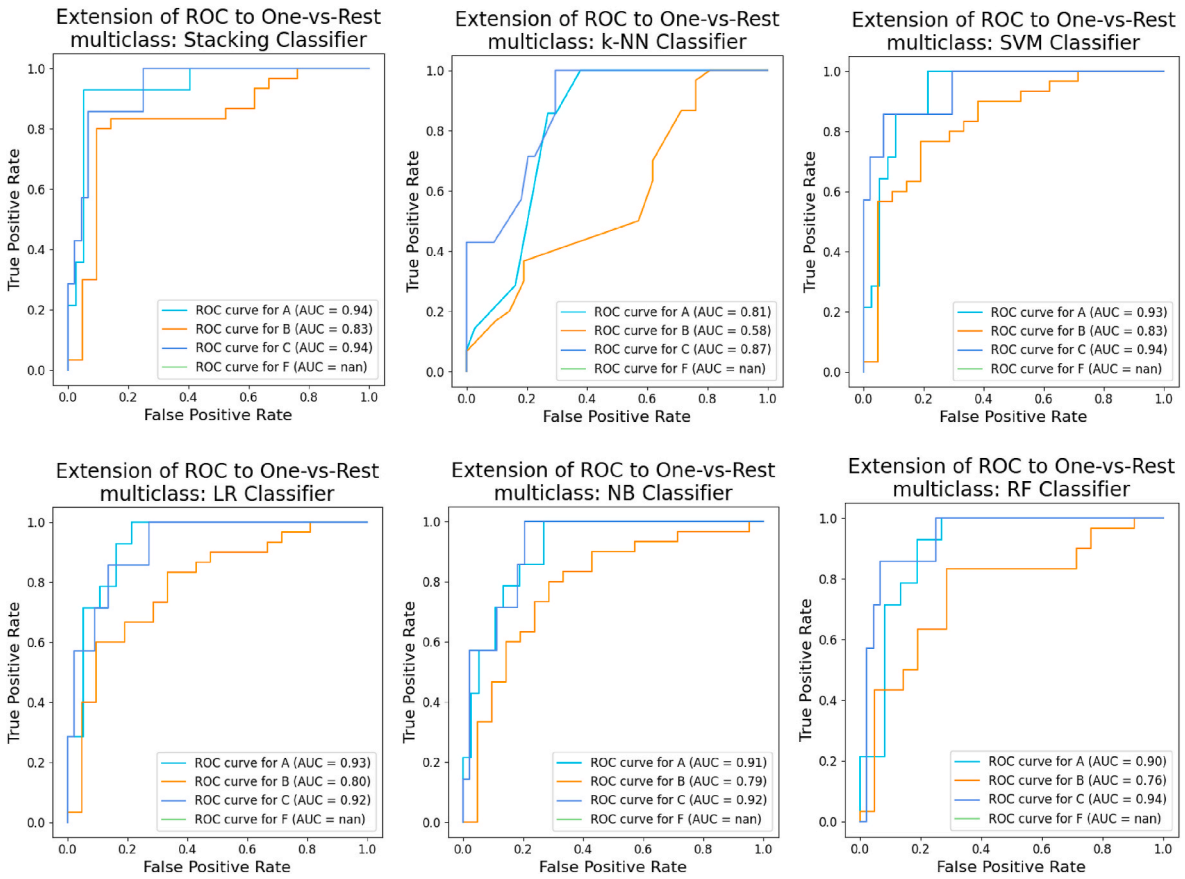

Fig. 10. ROC graphs of six models.

Table 8
Comparing Accuracy and F1 between Train set and Test set for Overfit Analysis.

Dataset	Test set		Train set	
Model	**Accuracy**	**F1**	**Accuracy**	**F1**
RF	0.784	0.768	**0.985**	0.738
NB	0.745	0.685	0.713	0.516
SVM	0.804	0.783	0.747	0.560
k-NN	0.667	0.536	0.688	0.484
LR	0.784	0.739	0.762	0.570
Stacking	**0.826**	**0.792**	0.838	**0.823**

Scholarship) to classify students with a pass or fail result. Therefore, to predict the student's grade effectively, a wide range of factors is needed, and the most crucial factor is, in fact, the academic background.

The impact of individual variables on the predictive power of the models was also investigated. Prior academic performance, as reflected in midterm exam scores and the average of different subjects, emerged as the most influential predictors of student success. In contrast, factors such as gender and compulsory course status exhibited minimal impact on model performance, suggesting that intrinsic motivation and individual characteristics may play a more significant role in determining academic outcomes.

By predicting students' results at the end of the course, students would have a chance to learn about their work and which grade they would achieve. This approach would enhance students' performance, and concentration and encourage them to be more diligent to gain higher grades. Since the midterm test is taken from class session 6 to 9 out of 16 class sessions, and more than one midterm test may be taken in the whole course, hence it is significant to note that students have more than 2 months (about 9 weeks) their midterm test to the final term exam.

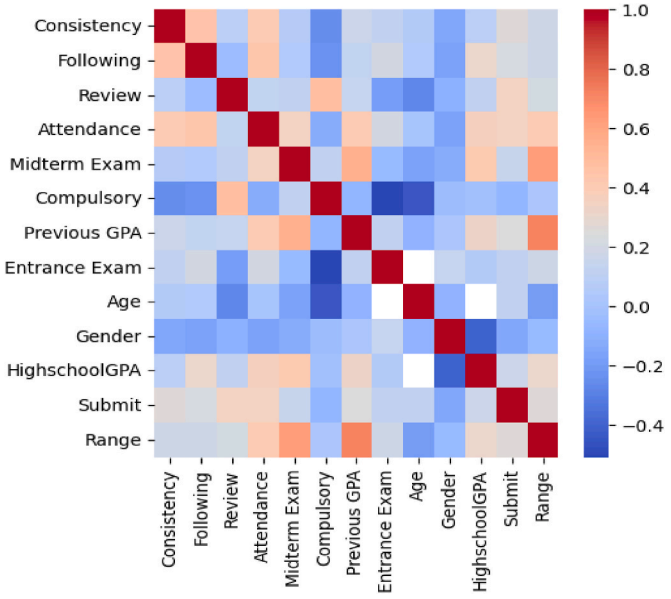

Fig. 11. Feature correlation matrix between different variables.

This would emphasize the importance of predicting student's performance in the future as students know their predicted average grade 2 months before the end of the course. Therefore, this study has tackled the task of predicting student performance, which is one of the most essential, useful, and challenging issues for educational institutions (Rastrollo-Guerrero et al., 2020).

Earlier academic performance (grades of other subjects, midterms,

Table 9
Compare of Stacking Classifier when dropping some features.

Feature dropped	Accuracy	Precision	Recall	F1	Acc. Declined
GPA Diff. Subjects	0.726	0.705	0.611	0.632	**0.101**
Midterm Exam	0.765	0.768	0.621	0.658	0.081
Attendance	0.784	0.753	0.730	0.740	0.041
Gender	0.826	0.800	0.790	0.794	0.000
Compulsory	0.804	0.778	0.778	0.778	0.022

attendance) reigns supreme in predicting student success, while individual characteristics and external factors like gender and course type hold minimal sway. As an answer to **Question 2**, our analysis revealed a stark contrast between these variable groups: those heavily impacting model accuracy (academic background) and those causing practically no change (demographic factors). This aligns with existing research (Arashpour et al., 2023; Fernandes et al., 2019; Riestra-González et al., 2021; Waheed et al., 2020; Xu et al., 2017) on cumulative achievement and suggests intrinsic motivation/past achievements might trump external factors. Our robust Stacking Classifier, encompassing a wider range of variables, achieved superior accuracy compared to models with narrower feature sets, highlighting the importance of a comprehensive approach for effective student performance prediction.

We engaged the expertise of two lecturers specializing in information technology and computer science to assess the predictive outcomes of our study. The predictive model received positive feedback from two university lecturers for its accuracy and the potential time-saving advantages it offers. Lecturer A, a leading authority in information technology from the Hanoi University of Science and Technology in Vietnam, commended the precision and accuracy of our predictive model, particularly the Stacking Classifier. He expressed that our system exhibits high accuracy and can significantly reduce the time required for instructors to evaluate student performance in their courses. Lecturer B, with experience in computer science at the National Economics University in Vietnam also expressed enthusiasm for the robustness of our approach. He believes that the system has potential and anticipates its integration into classroom settings at his university. They noted that the study addresses a crucial challenge in education by providing instructors with a valuable tool to assess and support student's academic progress. Both experts agreed that our inclusion of various student variables enhanced the model's effectiveness in predicting students' grade ranges, addressing a critical challenge in education.

5.2. Implications

The predictive outcomes of the study received positive feedback from university lecturers, commending the accuracy and potential time-saving advantages of the predictive model. The inclusion of various student variables enhanced the model's effectiveness in predicting students' grade ranges, offering instructors a valuable tool to assess and support academic progress. Integration of the predictive model into classroom settings was anticipated, addressing a crucial challenge in education by providing instructors with enhanced capabilities for student evaluation.

5.3. Discussion of future research

Although generating a high accuracy rate, the approach of this study has some limitations. One of those is that the model requires a large amount of data about students (diligence, academic background, demographic data). Although, in the scope of this study, the diligence data is quite simple to achieve, academic background and demographic data hold many null values. A temporary solution is that, in future works, when we deploy this model to the website "vnCodelab.com", the system would ask the users to input their missing values if possible. The second limitation is that the prediction of average grades still depends a lot on the Midterm exam scores, while we wanted to predict the final performance of students even earlier. This problem is worth researching more in the future to figure out novel solutions and approaches. Another future work is that, with added data from later courses, we would apply reinforcement learning and automatic continuous learning techniques in Machine Learning to sustain and enrich the model. Furthermore, we are interested in whether students' performance would be better if they knew their predicted grades. Finally, we would like to evaluate whether and how other AI applications such as large language models (ChatGPT, Bard, Llama) can assist students in their studies if they are integrated into the website. Evaluating the impact of students knowing their predicted grades and exploring the integration of other AI applications, such as large language models, into educational settings, present promising directions for future inquiry.

6. Conclusion

This study presents a predictive model for forecasting students' final course average scores, demonstrating its effectiveness through the use of a Stacking Classifier that combines multiple base models. By addressing the questions, we highlight the feasibility of early prediction of student performance, emphasizing the significant impact of prior academic performance on predictive accuracy. Our findings underscore the importance of considering factors such as midterm exam scores and the average of different subjects in predicting student outcomes, while demographic variables show minimal influence. Our study highlights the pivotal role of attendance as a predictor of student success. By incorporating attendance into our predictive model, we emphasize its significant impact on overall performance prediction accuracy, reflecting its importance in gauging student engagement and commitment. These findings underscore the value of attendance tracking in educational settings and suggest its potential as a targeted intervention area for improving student outcomes. The practical implications of our study lie in providing instructors with a valuable tool for supporting student progress and intervention. Future research directions could involve refining the predictive models and exploring additional data sources to enhance predictive accuracy and address data scarcity issues, ultimately contributing to improved student outcomes and the overall effectiveness of educational interventions.

Declaration of open data, ethics, and conflict of interest

The study was approved by an ethical committee with ID: 2022ET001. Informed consent was obtained from all participants, and their privacy rights were strictly observed. The participants were protected by hiding their personal information during the research process. They knew that the participation was voluntary and they could withdraw from the study at any time. There is no potential conflict of interest in this study. The data can be obtained by sending request e-mails to the corresponding author.

CRediT authorship contribution statement

Pham Xuan Lam: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Formal analysis, Conceptualization. **Phan Quoc Hung Mai:** Writing – review & editing, Writing – original draft, Formal analysis. **Quang Hung Nguyen:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Thao Pham:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Thi Hong Hanh Nguyen:** Writing – original draft. **Thi Huyen Nguyen:** Writing – review & editing, Writing – original draft.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and Grammarly to check spelling and grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Xuan Lam PHAM reports financial support, administrative support, and equipment, drugs, or supplies were provided by National Economics University.

References

Alsulami, A. A., AL-Ghamdi, A. S. A.-M., & Ragab, M. (2023). Enhancement of E-learning student's performance based on ensemble techniques. *Electronics, 12*(6), 1508. https://doi.org/10.3390/electronics12061508

Arashpour, M., Golafshani, E. M., Parthiban, R., Lamborn, J., Kashani, A., Li, H., & Farzanehfar, P. (2023). Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education, 31*(1), 83–99. https://doi.org/10.1002/cae.22572

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., Alsariera, Y. A., Ali, A. Q., Hashim, W., & Tiong, S. K. (2022). Toward predicting student's academic performance using artificial neural networks (ANNs). *Applied Sciences, 12*(3), 1289. https://doi.org/10.3390/app12031289

Baran, E., Correia, A.-P., & Thompson, A. (2011). Transforming online teaching practice: Critical analysis of the literature on the roles and competencies of online teachers. *Distance Education, 32*(3), 421–439. https://doi.org/10.1080/01587919.2011.610293

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education, 158*, Article 103999. https://doi.org/10.1016/j.compedu.2020.103999

Biswas, S. K., Nath Boruah, A., Saha, R., Raj, R. S., Chakraborty, M., & Bordoloi, M. (2023). Early detection of Parkinson disease using stacking ensemble method. *Computer Methods in Biomechanics and Biomedical Engineering, 26*(5), 527–539. https://doi.org/10.1080/10255842.2022.2072683

Chang, C., Shen, H.-Y., & Liu, E. Z.-F. (2014). University faculty's perspectives on the roles of e-instructors and their online instruction practice. *International Review of Research in Open and Distance Learning, 15*(3), 72–92. https://doi.org/10.19173/irrodl.v15i3.1654

Chango, W., Cerezo, R., & Romero, C. (2021). Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers & Electrical Engineering, 89*, Article 106908. https://doi.org/10.1016/j.compeleceng.2020.106908

Chen, S. (2022). Improved fuzzy algorithm for college students' academic early warning. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2022/5764800

Chi, C.-C., Kuo, C.-H., Lu, M.-Y., & Tsao, N.-L. (2008). Concept-based pages recommendation by using cluster algorithm. *Eighth IEEE international conference on advanced learning Technologies* (pp. 298–300). https://doi.org/10.1109/ICALT.2008.214

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Cornelius, S., & Gordon, C. (2008). Providing a flexible, learner-centred programme: Challenges for educators. *The Internet and Higher Education, 11*(1), 33–41. https://doi.org/10.1016/j.iheduc.2007.11.003

Davis, M., Herzog, L., & Legters, N. (2013). Organizing schools to address early warning indicators (EWIs): Common practices and challenges. *Journal of Education for Students Placed at Risk, 18*(1), 84–100. https://doi.org/10.1080/10824669.2013.745210

Dennen, V. P. (2008). Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior, 24*(4), 1624–1633. https://doi.org/10.1016/j.chb.2007.06.003

Dietterich, G. T. (1997). Machine learning research: Four current directions. *Artificial Intelligence, Magazine, 18*(4), 97–136. https://doi.org/10.1609/aimag.v18i4.1324

Durden, G. C., & Ellis, L. V. (1995). The effects of attendance on student learning in principles of economics. *The American Economic Review, 85*(2), 343–346. http://www.jstor.org/stable/2117945.

Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning, 54*, 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

Elrahman, A. A., Soliman, T. H. A., Taloba, A. I., & Farghally, M. F. (2023). A predictive model for student performance in classrooms using student interactions with an eTextbook. *Information Sciences Letters, 12*(1), 9–22. https://doi.org/10.18576/isl/120102

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies, 4*(4), 134–147. https://doi.org/10.1016/j.ijis.2020.09.001

Guruler, H., & Istanbullu, A. (2014). Modeling student performance in higher education using data mining. In A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends* (pp. 105–124). Springer International Publishing. https://doi.org/10.1007/978-3-319-02738-8_4.

Halpern, N. (2007). The impact of attendance and student characteristics on academic achievement: Findings from an undergraduate business management module. *Journal of Further and Higher Education, 31*(4), 335–349. https://doi.org/10.1080/03098770701626017

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Harvey, J. L., & Kumar, S. A. P. (2019). *A practical Model for Educators to predict student Performance in K-12 Education using machine learning IEEE symposium series on computational intelligence (SSCI)*. https://doi.org/10.1109/SSCI44817.2019.9003147. Xiamen, China.

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. *IOP Conference Series: Materials Science and Engineering, 928*(3), Article 032019. https://doi.org/10.1088/1757-899X/928/3/032019

Hawk, F., & Shah, A. (2007). Decision sciences journal of innovative education. *Using Learning Style Instruments to Enhance Student Learning*, (5), 1–19. https://doi.org/10.1111/j.1540-4609.2007.00125.x

Henrie, C. R., Bodily, R., Larsen, R., & Graham, C. R. (2018). Exploring the potential of LMS log data as a proxy measure of student engagement. *Journal of Computing in Higher Education, 30*, 344–362. https://doi.org/10.1007/s12528-017-9161-1

Hermanto, Y. B., & Srimulyani, V. A. (2021). The challenges of online learning during the covid-19 pandemic. *Jurnal Pendidikan Dan Pengajaran, 54*(1), 46–57. https://ejournal.undiksha.ac.id/index.php/JPP/article/view/29703.

Hoa, T. T. V., Huyền, P. T., & Hoa, N. Q. (2020). Trần, Thị Vân Hoa. "Đại dịch COVID-19: Cơ hội và thách thức cho Giáo dục Đại học Việt Nam.". *TẠP CHÍ KINH TẾ VÀ PHÁT TRIỂN, 274*(4), 64–74.

Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems, 101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Hovlid, E., Husabø, G., Valestrand, E. A., & Hartveit, M. (2022). Learning team-based quality improvement in a virtual setting: A qualitative study. *BMJ Open, 12*(6), Article e061390. https://doi.org/10.1136/bmjopen-2022-061390

Jiang, S., Williams, A., Schenke, K., Warschauer, M., & O'dowd, D. (2014). Predicting MOOC performance with week 1 behavior. *Educational data mining*. https://api.semanticscholar.org/CorpusID:21612658.

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching, 8*(3), 198. https://api.semanticscholar.org/CorpusID:61790532.

Kebritchi, M., Lipschuetz, A., & Santiague, L. (2017). Issues and challenges for teaching successful online courses in higher education: A literature review. *Journal of Educational Technology Systems, 46*(1), 4–29. https://doi.org/10.1177/0047239516661713

Latif, E., & Miles, S. (2013). Class attendance and academic performance: A panel data analysis. *Economic Papers: A journal of applied economics and policy, 32*(4), 470–476. https://doi.org/10.1111/1759-3441.12054

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 3093. https://doi.org/10.3390/app9153093

Liu, D., Zhang, Y., Zhang, J., Li, Q., Zhang, C., & Yin, Y. (2020). Multiple features fusion attention mechanism enhanced deep knowledge tracing for student performance prediction. *IEEE Access, 8*, 194894–194903. https://doi.org/10.1109/ACCESS.2020.3033200

Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. *Journal of Educational Technology & Society, 21*(2), 220–232. http://www.jstor.org/stable/26388400.

Marburger, D. R. (2001). Absenteeism and undergraduate exam performance. *The Journal of Economic Education, 32*(2), 99–109. https://doi.org/10.1080/00220480109595176

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. http://repository.alt.ac.uk/id/eprint/629.

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

Mensah, F. K., & Kiernan, K. E. (2010). Gender differences in educational attainment: Influences of the family environment. *British Educational Research Journal, 36*(2), 239–260. https://doi.org/10.1080/01411920902802198

Picciano, A. G. (2021). Theories and frameworks for online education: Seeking an integrated model. In *A guide to administering distance learning* (pp. 79–103). Brill. https://doi.org/10.1163/9789004471382_005.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences, 10*(3), 1042. https://doi.org/10.3390/app10031042

Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education, 163*, Article 104108. https://doi.org/10.1016/j.compedu.2020.104108

Rodgers, J. R. (2002). Encouraging tutorial attendance at university did not improve performance. *Australian Economic Papers, 41*(3), 255–266. https://doi.org/10.1111/1467-8454.00163

Romer, D. (1993). Do students go to class? Should they? *The Journal of Economic Perspectives, 7*(3), 167–174. https://doi.org/10.1257/jep.7.3.167

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Sánchez, E. M. T., Miguélañez, S. O., & Abad, F. M. (2019). Explanatory factors as predictors of academic achievement in PISA tests. An analysis of the moderating effect of gender. *International Journal of Educational Research, 96*, 111–119. https://doi.org/10.1016/j.ijer.2019.06.002

Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., Barroso, J., & de Jesus Filipe, V. M. (2022). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *International conference on technology and innovation in learning, teaching and education* (pp. 52–68). https://doi.org/10.1007/978-3-031-22918-3_5

Shi, J., Li, C., & Yan, X. (2023). Artificial intelligence for load forecasting: A stacking learning approach based on ensemble diversity regularization. *Energy, 262*, Article 125295. https://doi.org/10.1016/j.energy.2022.125295

Stanca, L. (2006). The effects of attendance on academic performance: Panel data evidence for introductory microeconomics. *The Journal of Economic Education, 37*(3), 251–266. https://doi.org/10.3200/JECE.37.3.251-266

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405. https://doi.org/10.1080/09645290701409939

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing & Communications, 19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, Article 106189. https://doi.org/10.1016/j.chb.2019.106189

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009). Predicting NDUM student's academic performance using data mining techniques. *2009 Second International Conference on Computer and Electrical Engineering, 2*, 357–361. https://doi.org/10.1109/ICCEE.2009.168

Wu, T., Zhang, W., Jiao, X., Guo, W., & Hamoud, Y. A. (2021). Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. *Computers and Electronics in Agriculture, 184*, Article 106039. https://doi.org/10.1016/j.compag.2021.106039

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 742–753. https://doi.org/10.1109/JSTSP.2017.2692560

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yuniastari, R., & Silva, A. M.d. (2022). The advantages and disadvantages of offline and emergency remote online general English classes. *Language Circle: Journal of Language and Literature, 16*(2), 394–412. https://doi.org/10.15294/lc.v16i2.31861

Zajac, M. (2009). Using learning styles to personalize online learning. *Campus-Wide Information Systems, 26*(3), 256–265. https://doi.org/10.1108/10650740910967410