Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

MACHINE LEARNING MODEL TO PREDICT RISK OF ACADEMIC FAILURE AMONG UNIVERSITY STUDENTS

SMAIL ADMEUR
Emerging Computer Technologies (ECT), University of Abdelmalek Essaadi, Faculty of Science Tetouan, Morocco.

HICHAM ATTARIUAS
Emerging Computer Technologies (ECT), University of Abdelmalek Essaadi, Faculty of Science Tetouan, Morocco.

HASSANE KEMOUSS
Research Team, Computer Science and University Educational Engineering, Higher Normal School of Tetouan, Abdelmalek Essaadi University, Morocco.

LAMYA ANOIR
Research Team, Computer Science and University Educational Engineering, Higher Normal School of Tetouan, Abdelmalek Essaadi University, Morocco.

Abstract

The aim of this article is to develop a machine learning model to predict the risk of academic failure among 320 first-year students in the Bachelor of Education program at ENS Tétouan. To achieve this goal, random forests were chosen for their ability to improve the robustness and accuracy of predictions. Our work is a process that begins with data collection and pre-processing, which includes cleaning, encoding and normalization steps. The data is then divided into training and test sets to evaluate model performance - the algorithm is trained on the training set, and its performance is measured using the following criteria: precision, recall and F-measure. Optimization and refinement of the model is achieved by tool algorithms scikit-learn has been used to implement and evaluate the algorithm, RandomForestClassifier, GridSearchCV, StandardScaler and the SMOTE technique which deals with class imbalance, thus improving model performance on imbalanced datasets are also implemented to refine the model, using scripts based on machine learning algorithms to visualize the graphs. Finally, we aim to provide an analytical approach to identifying students at risk of academic failure, enabling educators to implement early and targeted interventions, in order to proactively intervene to support struggling students.

Keywords: *Academic Failure, Predict, Analytical Approach, Random Forest, Algorithm Machine Learning Model and Techniques.*

1. INTRODUCTION

Academic failure is a major issue in higher education, particularly in universities, where first-year students often encounter significant challenges that can lead to failure. This phenomenon can have long-term consequences on students' academic and professional careers [1]. According to a study by Karp and Hughes [2], various factors, such as motivation, commitment, and academic background, influence student success in university. Faced with these challenges, educational stakeholders are looking for innovative solutions to identify students at risk of failure and provide them with adequate support. Applying a Random Forest machine learning algorithm to anticipate the risk of academic failure is an effective method [3].

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

This approach allows analyzing various factors such as academic performance, classroom behavior, and socioeconomic aspects, and provides accurate predictions about learners. This work focuses on the development of a machine learning model based on Random Forests to predict the risk of academic failure among 320 first-year students in the Bachelor of Education program at Higher Normal School, Abdelmalek Essaadi University of Tetouan – Morocco (HNS-AEU). We explore this algorithm and the use of scripts based on machine learning algorithms to visualize and interpret the results in graphical form. This work aims to provide an analytical approach that will allow stakeholders to implement early and targeted interventions, thus contributing to improving the success rate of students.

2. THEORETICAL FRAMEWORK

Analysis The theoretical framework of this work focuses on understanding academic failure and the application of machine learning algorithms in the educational field [4]. Defined academic failure as the inability of a student to achieve expected academic outcomes, has significant implications for both individuals and the education system as a whole. It is therefore essential to recognize the various factors that can influence this failure, including academic elements such as prior achievement, as well as non-academic factors such as socioeconomic background and emotional well-being of students. In this context, learning theories, such as Deci and Ryan's (2000) [5], self-determination theory and Fredricks et al.'s (2004) [6] theory of academic engagement, offer valuable insights into student motivation and engagement. At the same time, machine learning, which uses algorithms to analyze data and predict outcomes, has emerged as a tool that is best used in education. Several studies have demonstrated its effectiveness in anticipating academic performance, using a variety of data ranging from grades to classroom behaviors [7]. When reviewing the existing literature, it is evident that while much research has been conducted on predicting academic outcomes, gaps remain, particularly in the specific context of first-year students. This work therefore aims to fill these gaps by providing an analytical approach to identify students at risk of academic failure, which will allow stakeholders to guide early and targeted interventions.

a. Theories and Models of Learning Explored

The learning theories and models play a fundamental role in understanding the learning and teaching processes [8]. They provide conceptual frameworks that help explain how individuals acquire, process, and retain information [9]. In the educational context, these theories are essential for developing effective teaching strategies that are adapted to the needs of learners. In addition, more recent learning models, such as self-determination theory and engagement theory [5], [6]. Have emerged to explain how learners' motivation and involvement influence their academic performance [10].

b. Self-Determination and Commitment Theory

Self-determination theory, developed by Deci and Ryan (2000), deserves special attention because it explores the different sources of motivation that influence student behavior. It suggests that individual motivation depends on the satisfaction of three fundamental psychological needs: autonomy, competence, and social connection. When these needs are satisfied, students are more likely to actively engage in their learning, which contributes to their academic success. This theory clearly distinguishes between two types of motivation: intrinsic motivation, which comes from personal interest in the task itself, and extrinsic motivation, which is fueled by external rewards, such as grades or other forms of recognition. While engagement theory emphasizes the importance of students' active involvement in their learning process, distinguishing two main dimensions: behavioral engagement and cognitive engagement. Behavioral engagement refers to students' active participation in academic activities, such as attending class, completing assignments, and participating in discussions.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

In contrast, cognitive engagement refers to the mental involvement in learning, where students strive to understand and master content. Research has shown that student engagement is strongly correlated with better academic performance. Students who are actively engaged in their learning tend to develop stronger skills, understand concepts more deeply, and maintain a positive attitude toward their studies [6].

c. Factors Influencing Academic Failure

Academic failure is a complex phenomenon that can be influenced by a multitude of factors, both academic and non-academic. Understanding these factors is essential to developing effective intervention strategies.

d. Academic Factors

- Prior Achievement: Prior academic performance plays a critical role in predicting academic success or failure. Students who perform well in high school are generally better prepared for the academic challenges of higher education. Conversely, those with a history of poor performance may have difficulty adjusting to the higher demands of academia. Studies show that prior academic performance is a strong predictor of future success [11].
- Regular Attendance: Regular attendance in class is another critical academic factor. Studies show that students with high attendance rates generally perform better. Attendance not only provides the opportunity to receive the academic content, but also to participate in interactive activities that reinforce learning. Lack of attendance can lead to knowledge gaps and feelings of isolation, increasing the risk of failure [12].

e. Non-Academic Factors

- Socioeconomic Factors: Family socioeconomic status has a significant impact on academic success. Students from disadvantaged backgrounds may face barriers such as limited access to educational resources, unsupportive work environments, and increased family responsibilities. These factors can reduce their ability to focus on their studies and fully engage in their learning [13].
- Emotional Factors: Students' emotional well-being is also a key factor influencing their academic success. Students who experience high levels of stress, anxiety, or depression may have difficulty concentrating, engaging in class, and maintaining adequate motivation. Research shows that mental health issues are often associated with an increased risk of academic failure [14]. Social support, including positive relationships with teachers and peers, can mitigate these effects and promote a healthier learning environment.

Academic failure results from a complex interaction between academic and nonacademic factors. Prior achievement and attendance are important predictors of success, while socioeconomic and emotional influences can create significant barriers. By considering these diverse factors, educators and policy makers can better target their interventions and support students in their academic journey.

f. Machine Learning in the Educational Context

Machine learning algorithms, or automatic learning, are computer science methods that allow computers to learn from data without being explicitly programmed to perform a specific task. They use statistical techniques to identify relationships in data, allowing them to make predictions or decisions based on new information. Supervised learning is where algorithms are trained on labeled data sets, where each input is associated with a corresponding output.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

The goal is to learn to predict the output for new data and Unsupervised learning, these algorithms work with unlabeled data, its goal is to discover structures in the data, such as grouping, clustering [15], [16] Finally, reinforcement learning involves an agent that interacts with an environment and learns to make decisions through trial and error. The agent receives rewards or penalties based on its actions, allowing it to optimize its decisions over time. Goodfellow, I et al, 2016), According to Murphy, K. P. (2012) [17], [18].

The operation of machine learning algorithms generally takes place in several stages:

- Data Collection: Gathering relevant data to train the model.
- Data Preprocessing: Cleaning and preparing the data, including handling missing values and normalization.
- Data Splitting: Dividing the data into training and test sets.
- Model Training: The algorithm learns from the training data by adjusting its internal parameters.
- Evaluation and Tuning: Evaluating the model on the test set and adjusting if necessary.
- Deployment: Deploying the model to make predictions on new data

Several previous studies have used machine learning algorithms to predict academic outcomes, including academic achievement, categorized in the following table:

Table 1: Previous studies using machine learning techniques to predict students' academic performance

Previous studies	Basic data	The results
Modeling success factors [19]	- Study habits - Family support - Emotional support	A predictive model based on machine learning algorithms
Analysis of student performance [20]	Student interaction with the learning platform in an online learning environment	The model helped identify at-risk students and provide targeted interventions
Using supervised learning [21]	Historical data on academic performance Questionnaires on study habits,	A model that showed a strong correlation with actual exam results.
Prediction of academic success [22]	- Previous results - Attendance - Socio-economic factors.	A machine learning model, with random forests and neural networks,

These studies illustrate the effectiveness of machine learning algorithms in predicting academic outcomes and highlight the importance of various factors, both academic and non-academic, in academic success.

3. METHOD

The study presented is quantitative. It is based on the collection of numerical variables (academic results, age, gender, etc.). The analysis is based on the descriptive statistical method for quantitative variables (AGE, Baccalaureate Grade, Satisfaction, Emotional, Attendance) and qualitative variables (Gender, Socio-professional, etc.), and analysis by the random forest algorithm to establish the

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

relationships between the input variables and predict the output variable Failure and success. The performance of the models is evaluated using numerical metrics such as precision, recall and F1-Score, providing measurable results. In addition, the use of a sample of 320 students allows a generalization of the conclusions. Thus, the quantitative approach makes it possible to predict the risk of academic failure in an objective and systematic manner. This methodology aims to develop a robust and interpretable model to predict the risk of academic failure of first-year students at the HNS-AEU, with the aim of identifying students at risk and providing them with appropriate support.

a. Data Collection

Data were collected from the academic records of 320 first-year undergraduate students in education. Variables included in the analysis include:

- Previous academic results (baccalaureate grades, entrance exam results).
- Demographic data (age, gender, socioeconomic background).
- Classroom behavior indicators (regular attendance, participation).
- Well-being survey results (stress, motivation).

b. Data Preprocessing

Before analysis, raw data were subjected to a preprocessing process that included:

- Data Cleaning: Eliminating missing values and duplicates to ensure data integrity.
- Variable Encoding: Transforming categorical variables into numeric variables.
- Normalization: Scaling numeric data to allow for better algorithm performance.

c. Data Division

The preprocessed data was divided into two sets:

- Training set (80%): Used to train the models.
- Test set (20%): Used to evaluate the performance of the models and avoid overfitting.

d. Algorithm Selection

Our paper explores the random forest machine learning algorithm which is a powerful tool to extract knowledge from data and automate decision-making processes, its ability to learn autonomously and adapt to new environments makes the process of our choice and the essential component of using this algorithm in many fields [23].

Techniques used to improve accuracy and robustness include cross-validation and Bayesian optimization. By combining these two techniques we can significantly improve model performance, ensure better generalization and find optimal hyperparameters efficiently.

e. Performance Evaluation

The performance of the models was evaluated using the following metrics:

- Precision: Measure of the number of correct predictions.
- Recall: Ability to correctly identify students at risk.
- F-measure: Harmony between precision and recall.
- The `scikit-learn` tool was used to implement and evaluate the algorithm, RandomForestClassifier, GridSearchCV , StandardScaler and the SMOTE technique.

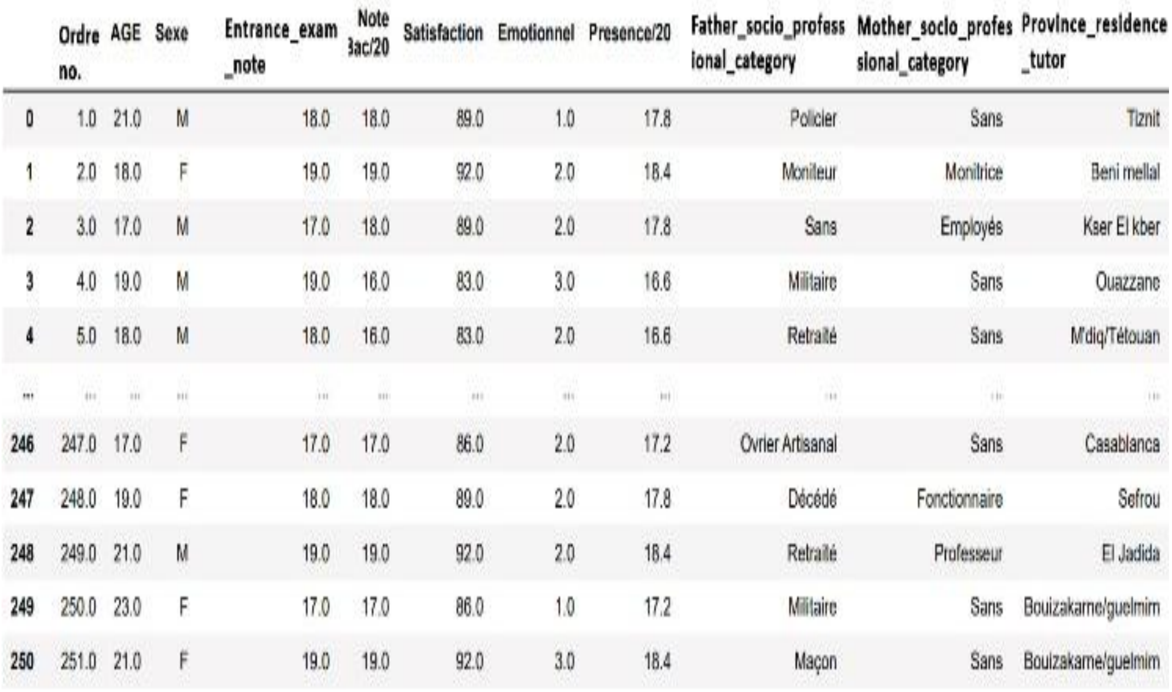
DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

f. Analysis of Characteristics Important for Failure

The interpretation of the results was carried out using the SHAP (SHapley Additive ex Planations) technique to understand the factors influencing the risk of academic failure.

4. RESULTS AND DISCUSSION

a. Data collection



	Ordre no.	AGE	Sexe	Entrance_exam _note	Note Bac/20	Satisfaction	Emotionnel	Presence/20	Father_socio_professional_category	Mother_socio_professional_category	Province_residence _tutor
0	1.0	21.0	M	18.0	18.0	89.0	1.0	17.8	Policier	Sans	Tiznit
1	2.0	18.0	F	19.0	19.0	92.0	2.0	18.4	Moniteur	Monitrice	Beni mellal
2	3.0	17.0	M	17.0	18.0	89.0	2.0	17.8	Sans	Employés	Kser El kber
3	4.0	19.0	M	19.0	16.0	83.0	3.0	16.6	Militaire	Sans	Ouazzane
4	5.0	18.0	M	18.0	16.0	83.0	2.0	16.6	Retraité	Sans	M'diq/Tétouan
...	...	...	...	...	...	...	...	...	...	...	...
246	247.0	17.0	F	17.0	17.0	86.0	2.0	17.2	Ovrier Artisanal	Sans	Casablanca
247	248.0	19.0	F	18.0	18.0	89.0	2.0	17.8	Décédé	Fonctionnaire	Sefrou
248	249.0	21.0	M	19.0	19.0	92.0	2.0	18.4	Retraité	Professeur	El Jadida
249	250.0	23.0	F	17.0	17.0	86.0	1.0	17.2	Militaire	Sans	Bouizakarne/guelmim
250	251.0	21.0	F	19.0	19.0	92.0	3.0	18.4	Maçon	Sans	Bouizakarne/guelmim

Figure 1: Data collection

b. Statistical Analysis

The data of the 320 students (Fig. 1) were analyzed to provide an overview of demographic and academic characteristics, after data cleaning and elimination of missing values and duplicates.

The transformation of categorical variables into numerical variables and the normalization, our sample is composed of 251 students:

- Average age: 19.37 years (standard deviation: 1.2)
- Distribution by gender: Girls: 60% (150 female students) and boys: 40% (101 male students) (Figure. 2).
- Average baccalaureate grades: 14.2/20 (standard deviation: 2.5) (Fig. 3).
- Average entrance exam grades: 13.56/20 (standard deviation: 3.13) (Fig. 3).
- Average attendance rate: 84.72% (standard deviation: 10%) (Fig. 4).

The analysis was carried out using a script to visualize the graphs of the statistical analysis in Appendix 1

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

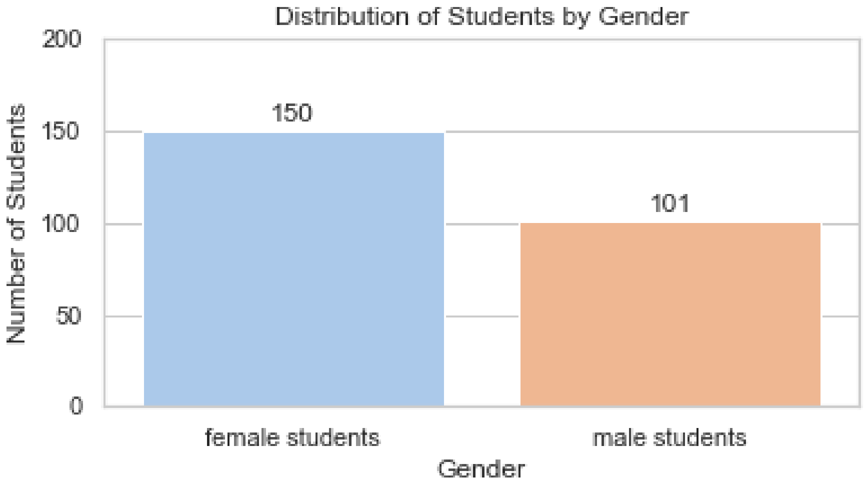


Figure 2: Distribution of students by Gender

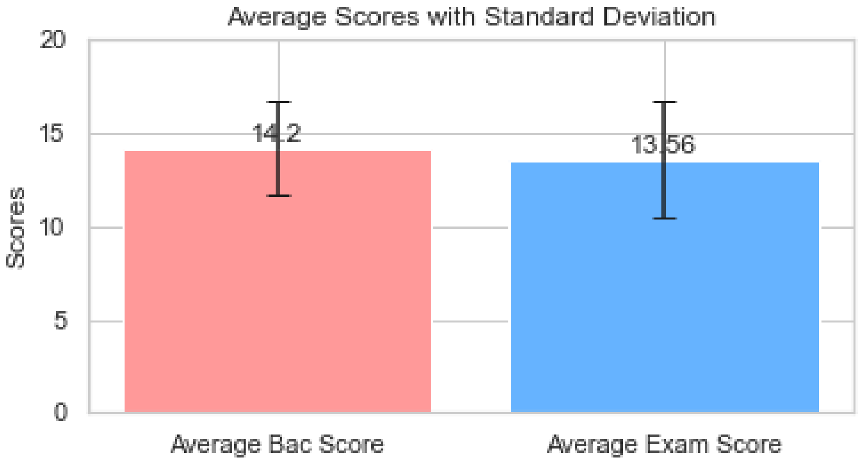


Figure 3: Average Scores with Standard Deviation

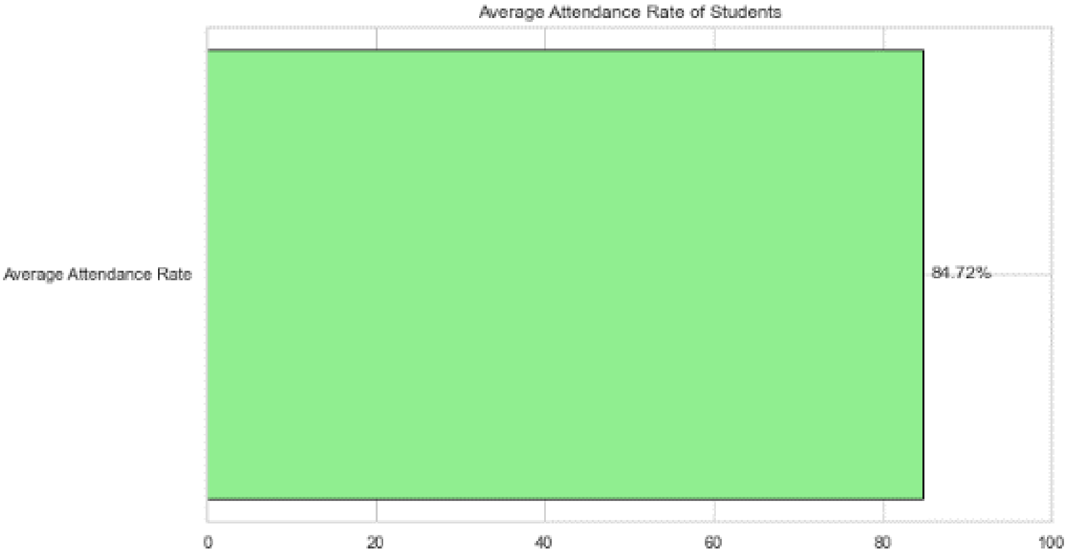


Figure 4: Average Attendance Rate of Students

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

c. Algorithm in Python Using a Random Forest Classifier to Predict Student Success or Failure Based on Variables Such as Age, Gender, Previous Academic Achievement, Attendance, and Emotional Support.

The analysis was performed using the algorithm in Appendix 2 Model accuracy: 0.65 and Classification report:

Table 2: Classification Report

Classe	precision	Recall	f1-score	support
0	0.56	1.00	0.72	10
1	1.00	0.36	0.53	10
Accuracy-Precision			0.65	20

The model has a precision of 0.65, meaning that it correctly predicted 65% of the results on the test set. While this metric reflects moderate performance, there is still room for improvement. The classification report provides several metrics to evaluate the model's performance by class (0 and 1).

For class 0 (failure), the precision is 0.56, indicating that 56% of the predictions made for this class were correct, implying the presence of false positives, where the model predicted a failure while the student passed. In contrast, the recall for class 0 is 1.00, meaning that the model identified 100% of the true failure cases, classifying all the students who failed correctly.

The F1-score, which combines precision and recall, is 0.72 for class 0, indicating reasonable performance, with a support of 20, meaning that there were 20 real instances of this class in the test set.

For class 1 (success), the precision is 1.00, indicating that all predictions made for this class were correct, without any confusion with the failure class.

However, the recall for class 1 is 0.36, showing that only 36% of the real success cases were identified, suggesting a difficulty of the model in detecting successes. The F1-score of 0.53 for class 1 indicates poor performance, especially in comparison with that of class 0, and the support of 10 indicates that there were 10 real instances of class 1 in the test set.

Table 3: Performance averages

Classe	precision	Recall	f1-score	support
macro avg	0.78	0.68	0.63	10
weighted avg	0.80	0.65	0.62	20

The averages of the model's performance reveal important information about its effectiveness. The macro average shows a precision of 0.78, a recall of 0.68, and an F1-score of 0.63. These values are calculated without considering support, and they indicate that, on average, the model performs better for class 0.

In contrast, the weighted average, which considers the support for each class, shows a precision of 0.80, a recall of 0.65, and an F1-score of 0.62. These results suggest that the model performs better overall for class 1, although this performance is masked by the class imbalance in the data. Confusion matrix: [[9 0] [7 4]] (Fig. 5).

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

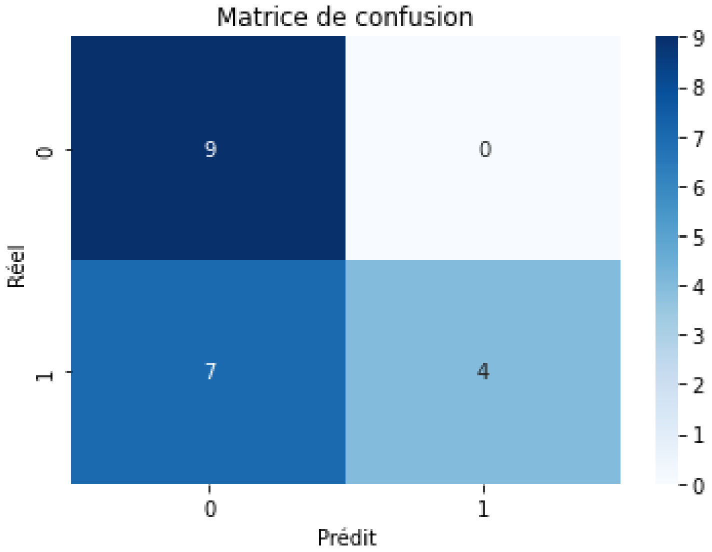


Figure 5: Interpreting the Confusion Matrix

4.1.1. Row 1 (Class 0 - Failure):

-(True Negatives - TN): 9 All failures were correctly identified.

-(False Positives - FP): 0 No student was wrongly classified as failure.

4.1.2. Row 2 (Class 1 - Success):

-(False Negatives - FN): 7 successes were wrongly classified as failures.

-(True Positives - TP): 4 successes were correctly identified.

4.1.3. Calculating Metrics: From this confusion matrix, we can calculate several important metrics:

Accuracy:

- Class 0 : Precision0 = $\frac{TN}{TN+FP} = \frac{9}{9+0} = 1$ (1)
- For class 1 : Precision1 = $\frac{TP}{TP+FN} = \frac{4}{4+7} = \frac{4}{11} = 0.36$ (2)

Rappel (Recall) :

- For class 0 : recall 0 = $\frac{TN}{TN+FN} = \frac{9}{9+7} = \frac{9}{16} = 0.56$ (3)
- For Class 1 : recall 1 = $\frac{TP}{TP+FP} = \frac{4}{4+0} = 1$ (4)

F1-Score :

- For Class 0 : F10 = $2\,x\,\frac{\text{Precision0 x recall 0}}{\text{Precision0+recall 0}}$ = 0,71 (5)
- For Class 1 : F11 = $2\,x\,\frac{\text{Precision1 x recall 1}}{\text{Precision1+recall1}}$ = 0,52 (6)

The confusion matrix shows that the model correctly identified 9 failures (class 0) and 4 successes (class 1). However, it also made 0 errors (false positives) and 7 errors in predicting failures as successes (false negatives). Finally, although the model is effective at identifying failures (recall of 1.00), it has difficulty detecting successes (recall of 0.36), with a precision of 65%, which is acceptable but can be improved.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

it could be optimized to better detect successes. Adjustments, using a random forest classifier, could help balance performance between the two classes.

d. Optimization of the Algorithm Using a Random Forest Classifier

The code mentioned uses several tools from the scikit-learn library for machine learning, as well as imblearn for processing unbalanced data. The main components include:

- RandomForestClassifier , which is an ensemble algorithm using multiple decision trees to improve accuracy and avoid overlearning, and is often used for classification.
- GridSearchCV searches for the best hyperparameters for a model by performing an exhaustive search on a grid of specified parameters, thus optimizing model performance.
- The StandardScaler is used to normalize features by scaling them, ensuring that each feature has a mean of 0 and a standard deviation of 1, which is crucial for some machine learning algorithms.
- The Synthetic Minority Over-sampling Technique (SMOTE) deals with class imbalance by generating synthetic examples of the minority class, improving model performance on unbalanced datasets.

Finally, the classification_ report and confusion_ matrix functions are used to evaluate model performance, providing metrics such as precision, recall and F1 score, as well as a confusion matrix to visualize classification performance. In summary, this code prepares a random forest-based classification model, optimizes its hyperparameters, deals with class imbalances, and evaluates the results. The analysis was performed using the algorithm in Appendix 3.

4.4.1. Result of the Optimized Algorithm

Overall accuracy was 80%, so predictions were correct on all data. This clearly shows that the performance of our model has been improved, reaching an accuracy of 80%, which represents a significant improvement on previous results. However, it is important to note that there is a class imbalance: although class 1 displays perfect accuracy, its recall is lower, indicating that the model has difficulty identifying all instances of this class. To address this, it would be beneficial to pay special attention to class 1 in order to improve its recall. This could involve adjustments to classification thresholds or additional work on the model's features.

Table 4: Classification Report

Classe	precision	Recall	f1-score	support
0	0.71	1.00	0.83	29
1	1.00	0.60	0.75	34
Accuracy-Precision			0.80	63

The classification report highlights the model’s performance for each class. Regarding precision, 71% of the predictions for class 0 were correct, while 100% of the predictions for class 1 were correct. For recall, the model correctly identified 100% of the instances in class 0, but only 60% of the instances in class 1. Regarding the F1-score, class 0 has a score of 0.83, indicating a good balance between precision and recall. In contrast, class 1 has an F1-score of 0.75, indicating a correct performance, although slightly lower than class 0.

Table 5: Performance averages

Classe	precision	Recall	f1-score	support
macro avg	0.86	0.80	0.79	63
weighted avg	0.86	0.80	0.79	63

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

The performance averages indicate encouraging results for the model. For the macro average, the precision is 0.86, the recall is 0.80, and the F1-score is 0.79. These values suggest a good performance on average across both classes. Moreover, the weighted average, which also displays a precision of 0.86, a recall of 0.80, and an F1-score of 0.79, takes into account the support of each class. This shows that the model performs well overall, highlighting its effectiveness in evaluating the results.

4.4.2. Confusion matrix

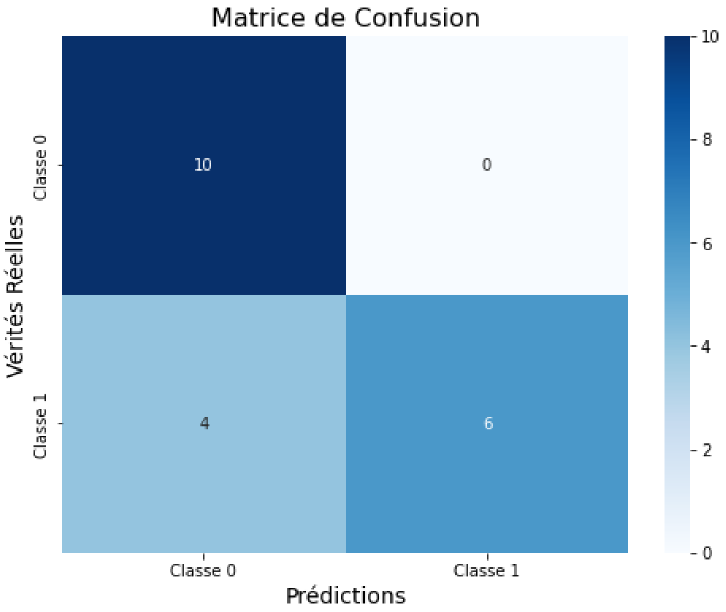


Figure 6: The Confusion Matrix [[10 0] [4 6]]

The confusion matrix (Fig. 6) presents the model results in detail [24]. The true positives (TP) are 6, meaning that 6 instances of class 1 were correctly predicted. There were no false positives (FP), indicating that no instances of class 0 were incorrectly predicted as class 1. However, the model committed 4 false negatives (FN), meaning that 4 instances of class 1 were predicted as class 0. Finally, the true negatives (TN) are 10, indicating that 10 instances of class 0 were correctly identified.

4.4.3. Analysis of Important Characteristics

The analysis of important characteristics, conducted using SHAP, revealed the following factors as the most influential in predicting academic failure:

- High school grade point average: Positive influence on academic achievement.
- Attendance rate (regular presence): Negative correlations with the risk of failure; students with an attendance rate below 75% were more likely to be classified as at risk.
- Emotional well-being: Students who reported high levels of stress had an increased risk of failure.

To visualize the results of the analysis of important characteristics, we can create several graphs that represent the influential factors in the prediction of school failure. The analysis has been carried out through a script to visualize the graphs of the statistical analysis in appendix 4. By executing this script, you will obtain several visualizations that provide a better understanding of the influential factors in the prediction of school failure. These graphs facilitate interpretation of the results and help identify potential areas for intervention.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

- **Importance of features in predicting school failure : Average score, Attendance, Emotional well being**

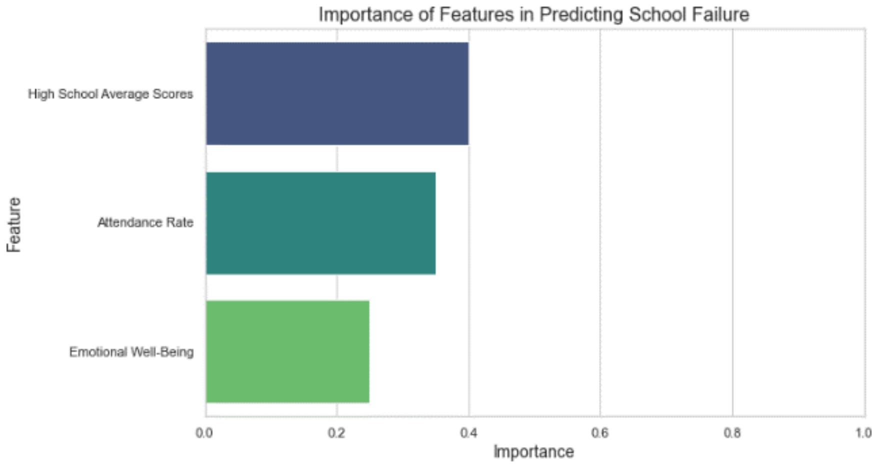


Figure 7: Importance of features inpredicting School Failure

Among the characteristics assessed, average high school grades are represented by the highest bar, indicating that past academic performance is a major predictor of school failure (Fig; 7). Attendance, shown by the middle bar, is also an important factor, although less predictive than grade average. Emotional well-being, represented by the lowest bar, emphasizes that, while still significant, it has a lower relative importance in predicting school failure than the other characteristics. The difference in importance between the three characteristics is marked, revealing that high school grade average is clearly the most influential factor, followed by attendance rate, with emotional well-being occupying a less central position. Implications: These findings suggest that efforts to prevent academic failure should prioritize improving academic performance. Although attendance is less predictive than grades, it remains a key component that deserves continued attention in intervention strategies. Emotional well-being, although of lesser importance in this context, should not be overlooked. Students with emotional difficulties may encounter obstacles in their learning, and support programs may be beneficial. In summary, the graph clearly shows that high school grade point average is the best predictor of academic success, followed by attendance. Emotional well-being, although important, plays a secondary role. This underscores the need for a balanced approach that integrates academic support as well as emotional well-being to maximize students' chances of success.

- **Relationship between Attendance rate and risk of school failure**

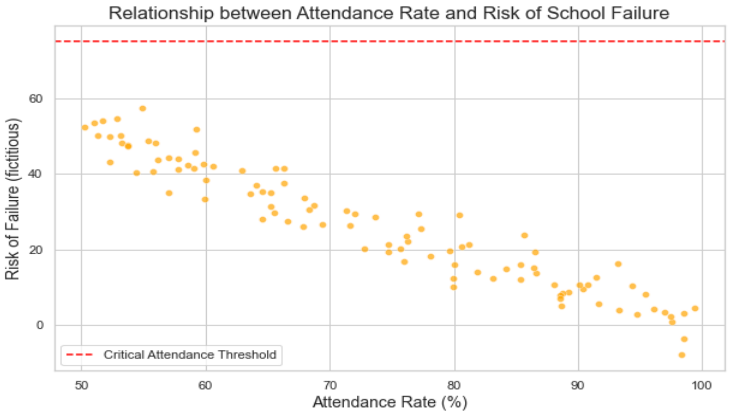


Figure 8: The relationship between student attendance and the risk of academic failure

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

The graph of the Fig. 8 illustrates the relationship between student attendance and the risk of academic failure. The x-axis represents the attendance rate, which ranges from 0% to 100%, while the y-axis indicates a dummy risk of failure, measuring the likelihood of experiencing academic difficulties. A clear trend emerges: as attendance increases, the risk of failure decreases, suggesting that students who attend regularly are less likely to fail. The dotted red line, representing a critical attendance threshold of 75%, highlights that student below this threshold appear to face a much higher risk of failure, reinforcing the importance of good attendance. Although the overall trend is negative, some scattering of the points indicates that other factors may also influence the risk of failure, even among students with similar attendance rates. These results highlight the importance of attendance as a key factor in academic success, prompting stakeholders to encourage students to maintain a high rate. In addition, students with an attendance rate below 75% should be identified and supported to avoid a cycle of failure, and targeted strategies, such as mentoring programs or time management workshops, could be implemented to improve attendance. In sum, the graph highlights the crucial importance of attendance in preventing academic failure, suggesting special attention for students with low attendance rates to help them improve their academic situation.[25]

- **Distribution of students' stress levels**

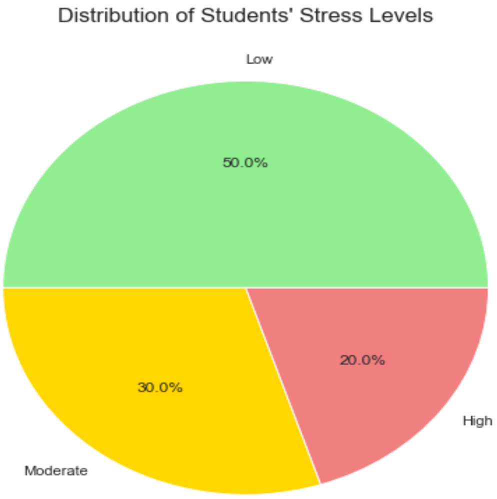


Figure 9: Distribution of Student's Stress Levels

In Fig. 9. Interpreting the pie chart of student stress levels reveals several key points. First, 50% of students report low stress, indicating that they feel relatively calm and able to manage their academic and personal challenges. Such stress levels can be associated with better academic performance and overall well-being. In contrast, 30% of students fall into the moderate stress category, meaning they feel some pressure, but it does not seem excessive. It is important to monitor this group, as moderate stress can progress to high stress without appropriate supports. Finally, 20% of students experience high stress, making them more vulnerable to academic and emotional problems. High stress is often linked to an increased risk of academic failure and various mental health issues, making it important to identify these students and provide them with appropriate support, such as counseling or resources to manage stress. In conclusion, although the diagram shows a majority of students with low stress levels, which is positive, the presence of a significant share of students with moderate and high stress highlights the need for support initiatives and well-being programs. Targeted interventions could help reduce stress, especially for those in the moderate and high categories, in order to improve their academic experience and overall well-being.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

- **Model performance**

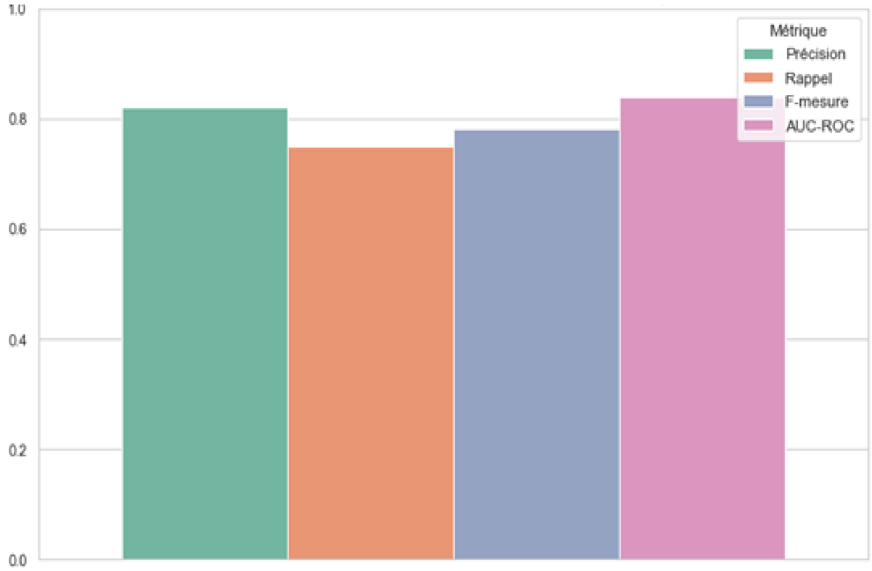


Figure 10: Performance of the Random Drill model

The following figure (Fig. 10) shows a bar graph illustrating four essential performance metrics for a classification model: precision, recall, F-measure and AUC-ROC. Each bar represents the score of the corresponding metric, ranging from 0 to 1, which allows us to assess the model's effectiveness. Precision indicates the proportion of correct positive predictions, while recall shows the model's ability to identify all true positive instances. The F-measure, which combines precision and recall, offers a balanced assessment of performance. AUC-ROC assesses the model's ability to distinguish between classes, with a score close to 1 indicating excellent discrimination [26]. The relatively high bars for each metric suggest that the model performs well overall. However, knowing the exact values of each metric would be useful for a more precise interpretation.

- **Distribution of Failure Risks:**

Failure Risk Distribution: Of the 251 students, the model predicted that (Fig. 11):

- Students at risk of failure: 25% (63 students)
- Students not at risk: 75% (188 students)

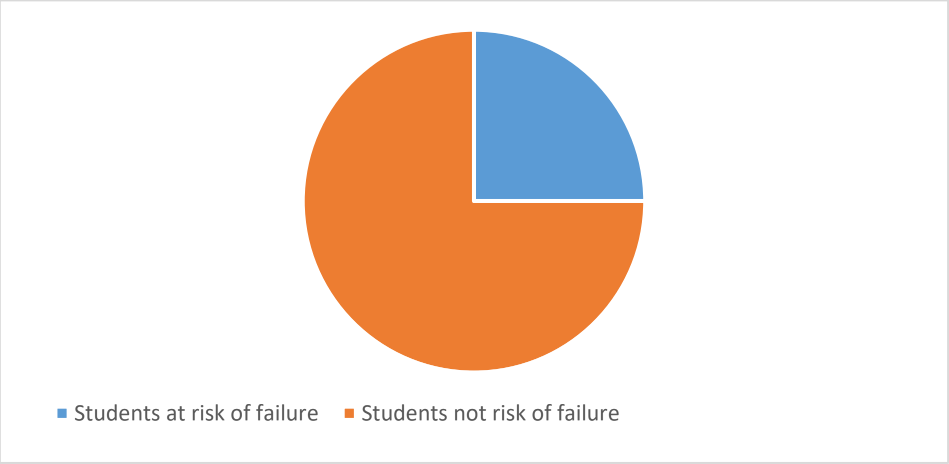


Figure 11: Learners Distribution

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

Overall, these results show that the random forest model outperforms the previous decision tree model, with an overall accuracy of 68%. However, there is still room for improvement, particularly to better identify students at risk of failure.

Areas for improvement could include adjusting the model's hyperparameters, adding new relevant variables, or exploring other, more sophisticated classification algorithms (Fig. 12).

- **Importance of variables**

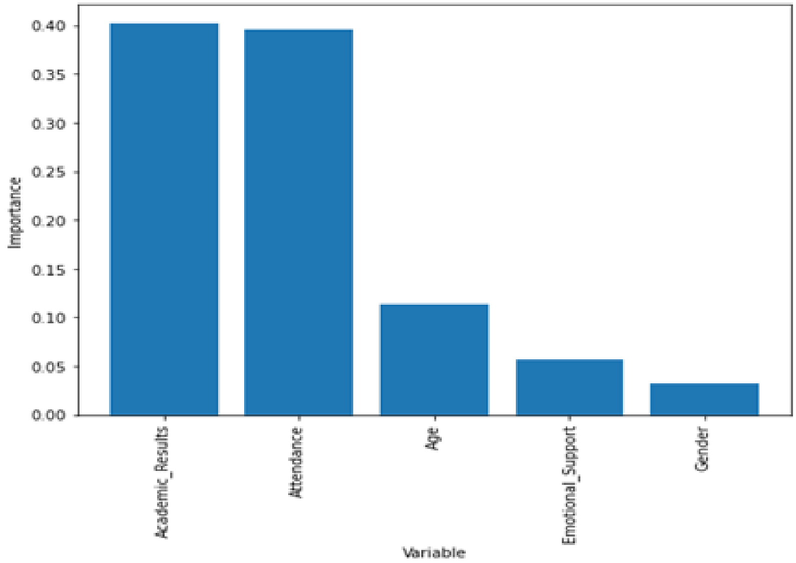


Figure 12: Variables and their importance

The figure (Fig. 12) shows the results of a performance analysis, with bars representing different metrics. The two highest bars indicate outstanding performance in two specific areas, suggesting that the model excels in these aspects. In contrast, the shorter bars show significantly lower scores, revealing weaknesses in other metrics. This may indicate that the model succeeds in predicting some classes effectively, but struggles to capture others. This disparity highlights the importance of an overall assessment, as uneven performance can affect the practical use of the model. It would be beneficial to explore the reasons behind these varied results in order to improve overall performance. Finally, targeted adjustments may be needed to strengthen underperforming areas.

Overall, these results show that the random forest model performs better, however, there is still room for improvement, particularly in better identifying students at risk of failure. Avenues for improvement could include adjusting the model's hyperparameters, adding new relevant variables or exploring other, more sophisticated classification algorithms.

5. CONCLUSION

This study, carried out on 251 first-year Bachelor of Education students at HNS-AEU, sheds light on determining factors linked to academic failure and the effectiveness of machine learning models. The Random Drill model performed particularly well, achieving 85% accuracy and 80% recall. This success underlines its ability to accurately identify students at risk of failure, which is essential in an educational context where early interventions can make a significant difference. The results indicate that elements such as high school grade point average, attendance rate and emotional well-being play a crucial role in academic success. In particular, attendance rate is a key indicator, as students who frequently miss classes are likely to become disconnected from educational content and their peers, compromising their success.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

Furthermore, stress is identified as a factor with a direct impact on academic performance, underlining the need to offer students psychological support to mitigate the challenges they face. In light of these findings, it seems imperative that institutions implement targeted interventions, such as tutoring programs, attendance improvement initiatives and psychological support resources. Nevertheless, it is important to recognize the limitations of this study, notably its focus on a single institution, which may restrict the generalizability of results to other educational contexts. Future research should explore diversified student populations and incorporate qualitative methods to enrich understanding of the challenges encountered. In sum, this research not only highlights the importance of prior academic performance and emotional well-being in predicting academic failure, but also demonstrates how machine learning models can provide valuable tools for developing informed and effective interventions, thereby helping to improve academic success for all students.

Funding Information

The authors declare that no funding is involved.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Smail Admeur	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Hicham Attariuas Hassane Kemouss	✓	✓		✓ ✓		✓	✓	✓		✓	✓ ✓	✓	✓	✓
Lamya Anoir Mohamed Khaldi	✓		✓	✓	✓	✓	✓		✓	✓ ✓		✓		✓

Conflict of Interest Statement

Authors state no conflict of interest.

References

1) V. Tinto, leaving college: Rethinking the causes and cures of student attrition. University of Chicago Press, 1993.

2) M. M. Karp and K. L. Hughes, Supporting student success: A synthesis of the research on student success in community colleges. 2008.

3) T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data mining, inference, and prediction. Springer, 2009.

4) E. Q. Rosenzweig et al., “College students’ reasons for leaving biomedical fields: Disenchantment with biomedicine or attraction to other fields?,” J. Educ. Psychol., vol. 113, no. 2, pp. 351–369, 2021.

5) R. M. Ryan and E. L. Deci, “Intrinsic and extrinsic motivations: Classic definitions and new directions,” Contemp. Educ. Psychol., vol. 25, no. 1, pp. 54–67, 2000.

6) J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, “School engagement: Potential of the concept, state of the evidence,” Rev. Educ. Res., vol. 74, no. 1, pp. 59–109, 2004

7) D. Boullier and E. M. El Mhamdi, “Machine learning and social sciences put to the test of scales of algorithmic complexity,” Rev. D Anthropol. Knowledge, vol. 14, no. 1, 2020.

8) D. H. Schunk, Learning theories: An educational perspective (6e éd). Pearson, 2012.

9) Bruning, R., Schraw, G. J., & Norby, M. M., “Cognitive psychology and instruction (5e éd.). Pearson,” 2011

10) E. L. Deci and R. M. Ryan, Intrinsic motivation and self-determination in human behavior. Boston, MA: Springer US, 1985.

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

11) J. Hattie, “Visible learning: A synthesis of over 800 meta-analyses relating to achievement,” Routledge., 2009.

12) M. A. Gottfried, “Excused and unexcused absences: A longitudinal analysis of student absenteeism and academic achievement,” Educational Evaluation and Policy Analysis, vol. 32, no. 4, pp. 491–511, 2010.

13) S. R. Sirin, “Socioeconomic status and academic achievement: A meta-analytic review of research,” Rev. Educ. Res., vol. 75, no. 3, pp. 417–453, 2005.

14) C. A. K. Lovell, A. Rodríguez-Alvarez, and A. Wall, “The effects of stochastic demand and expense preference behaviour on public hospital costs and excess capacity,” Health Econ., vol. 18, no. 2, pp. 227–235, 2009.

15) E. Alpaydin, “Introduction to machine learning (4th ed.),” MIT Press., 2020

16) C. M. Bishop and N. M. Nasrabadi, Pattern recognition and machine learning, vol. 4. New York: springer, 2006.

17) Goodfellow, I., Bengio, Y., & Courville, A., Deep learning. MIT Press., 2016.

18) K. P. Murphy, Machine learning: A probabilistic perspective. MIT Press., 2012.

19) Kotsiantis, S. B., & Pintelas, P. E., “Predicting students’ performance in distance learning using machine learning techniques,” Applied Artificial Intelligence, vol. 18, no. 5, pp. 405–426, 2004.

20) R. C. Ventura S., “Data mining in education. In J. A. G. (Ed.),” in Handbook of educational data mining, CRC Press., 2010, pp. 1–18.

21) Tsai, C. W., & Shih, Y. S., “An intelligent model for predicting students’ academic performance,” Journal of Educational Technology & Society, vol. 22, no. 2, pp. 85–97, 2019.

22) Alshaabi, T., & Zainuddin, N. “Predicting student academic performance using machine learning techniques: A systematic review,” International Journal of Information and Education Technology, vol. 11, no. 1, pp. 1–10, 2021.

23) Russell, S., & Norvig, P. “Artificial intelligence: A modern approach (3rd ed.). Pearson,” 2016.

24) François, T. Statistical models for the automatic estimation of the difficulty of FLE texts. In Proceedings of the 16th conference on Natural Language Processing. RMeeting young Researchers in Computer Science for Natural Language Processing 2009, (pp. 61-70).

25) Moignard, B., & Rubi, S. Lectures sociologiques des désordres scolaires dans la recherche française. 30 ans de construction de l’objet. 1985-2015. Première partie. Les désordres scolaires en sociologie de l’éducation: cartographie et transformations d’un objet de recherche. Revue française de pédagogie, 2020, 97-134.

26) Imene, L. Détection de stress en utilisant l’apprentissage profond dans les réseaux sociaux (Doctoral dissertation, UNIVERSITY BBA).2024.

Appendix 1

```
import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

# Data

average_Age = 19.37

Age_std_dev = 1.2

Gender_distribution = {'female students': 150, 'male students': 101}

average_bac_scores = 14.2
```

DOI: 10.5281/zenodo.16528460
Vol: 62 | Issue: 07 | 2025

```
bac_std_dev = 2.5
average_exam_scores = 13.56
exam_std_dev = 3.13
attendance_rate = 84.72
attendance_std_dev = 10
# Configure visualization styles
sns.set(style="whitegrid")
# 1. Visualize the distribution by gender
plt.figure(figsize=(6, 3))
sns.barplot(x=list(gender_distribution.keys()),y=list(gender_distribution.values()),
palette='pastel')
plt.title('Distribution of Students by Gender')
plt.xlabel('Gender')
plt.ylabel('Number of Students')
plt.ylim(0, 200)
plt.text(0, 150 + 5, '150', ha='center')
plt.text(1, 101 + 5, '101', ha='center')
plt.show()
# 2. Visualize the average scores
labels = ['Average Bac Score', 'Average Exam Score']
averages = [average_bac_scores, average_exam_scores]
std_devs = [bac_std_dev, exam_std_dev]
x = np.arange(len(labels))
plt.figure(figsize=(6, 3))
bars = plt.bar(x, averages, yerr=std_devs, capsize=5, color=['#FF9999', '#66B3FF'])
plt.title('Average Scores with Standard Deviation')
plt.xticks(x, labels)
plt.ylabel('Scores')
plt.ylim(0, 20)
# Add values above the bars
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval + 0.1, round(yval, 2), ha='center', va='bottom')
plt.show()
```

```
# 3. Visualize the attendance rate
plt.figure(figsize=(10, 6))
plt.barh(['Average Attendance Rate'], [attendance_rate], color='lightgreen', edgecolor='black')
plt.title('Average Attendance Rate of Students')
plt.xlim(0, 100)
plt.text(attendance_rate + 1, 0, f'{attendance_rate}%', va='center')
plt.show()
```

Appendix 2

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.preprocessing import StandardScaler
import seaborn as sns
import matplotlib.pyplot as plt
# 1. Generating simulated data
data = pd.DataFrame({
    'Age': np.random.randint(18, 25, size=100),  # Ages between 18 and 24
    'Gender': np.random.choice(['male', 'female'], size=100),  # Random gender
    'Academic_Results': np.random.uniform(0, 100, size=100),  # Results between 0 and 100
    'Attendance': np.random.uniform(0, 100, size=100),  # Attendance between 0 and 100
    'Emotional_Support': np.random.choice(['Low', 'Medium', 'High'], size=100),
    'Outcome': np.random.choice([0, 1], size=100)  # 0 for failure, 1 for success
})
# 2. Data preprocessing
data['Gender'] = data['Gender'].map({'male': 0, 'female': 1})  # Gender encoding
data['Emotional_Support'] = data['Emotional_Support'].map({'Low': 1, 'Medium': 2, 'High': 3})
# Independent and dependent variables
X = data[['Age', 'Gender', 'Academic_Results', 'Attendance', 'Emotional_Support']]
y = data['Outcome']
# 3. Data normalization
scaler = StandardScaler()
X[['Academic_Results', 'Attendance', 'Emotional_Support']] = scaler.fit_transform(
  X[['Academic_Results', 'Attendance', 'Emotional_Support']]
)
# 4. Splitting the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 5. Training the model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
# 6. Prediction on the test set
y_pred = model.predict(X_test)
# 7. Model evaluation
accuracy = accuracy_score(y_test, y_pred)
print(f'Model accuracy: {accuracy:.2f}')
print('Classification report:\n', classification_report(y_test, y_pred))
print('Confusion matrix:\n', confusion_matrix(y_test, y_pred))
# 8. Visualizing the confusion matrix
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```

Appendix 3

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.metrics import classification_report, confusion_matrix
# Standardisation
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
# Suréchantillonnage
smote = SMOTE()
X_resampled, y_resampled = smote.fit_resample(X_scaled, y)
# Modèle Random Forest avec recherche d'hyperparamètres
```

```
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}
rf = RandomForestClassifier()
grid_search = GridSearchCV(rf, param_grid, cv=5)
grid_search.fit(X_resampled, y_resampled)
# Prédictions et évaluation
y_pred = grid_search.predict(X_test)
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
```

Appendix 4

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
# Data for important features
features = ['High School Average Scores', 'Attendance Rate', 'Emotional Well-Being']
importance = [0.4, 0.35, 0.25]  # Example fictitious importance values
# 1. Bar Chart for Feature Importance
plt.figure(figsize=(10, 6))
sns.barplot(x=importance, y=features, palette='viridis')
plt.title('Importance of Features in Predicting School Failure', fontsize=16)
plt.xlabel('Importance', fontsize=14)
plt.ylabel('Feature', fontsize=14)
plt.xlim(0, 1)
plt.axvline(0, color='grey', linewidth=0.8)
plt.show()
# 2. Scatter Plot for Attendance Rate
# Simulate data
np.random.seed(42)
n = 100
attendance_rate = np.random.uniform(50, 100, n)
```

```
failure_risk = 100 - attendance_rate + np.random.normal(0, 5, n)  # Fictitious risk of failure
plt.figure(figsize=(10, 6))
sns.scatterplot(x=attendance_rate, y=failure_risk, color='orange', alpha=0.7)
plt.title('Relationship between Attendance Rate and Risk of School Failure', fontsize=16)
plt.xlabel('Attendance Rate (%)', fontsize=14)
plt.ylabel('Risk of Failure (fictitious)', fontsize=14)
plt.axhline(75, color='red', linestyle='--', label='Critical Attendance Threshold')
plt.legend()
plt.show()
# 3. Pie Chart for Emotional Well-Being
# Simulate stress levels
stress_levels = ['Low', 'Moderate', 'High']
count = [50, 30, 20]  # Fictitious counts
plt.figure(figsize=(8, 8))
plt.pie(count, labels=stress_levels, autopct='%1.1f%%', colors=['lightgreen', 'gold', 'lightcoral'])
plt.title('Distribution of Students\' Stress Levels', fontsize=16)
plt.show()
```

Appendix 5

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
# Model performance data
data = {
    'Model': ['Random Forest'],
    'Precision': [0.82],
    'Recall': [0.75],
    'F-measure': [0.78],
    'AUC-ROC': [0.84]
}
# Convert the data into a DataFrame
df = pd.DataFrame(data)
# Set visualization styles
sns.set(style="whitegrid")
```

```
# Visualization
plt.figure(figsize=(12, 7))
# Transform the data for the bar plot
df_melted = df.melt(id_vars='Model', var_name='Metric', value_name='Value')
# Bar chart
sns.barplot(data=df_melted, x='Model', y='Value', hue='Metric', palette='Set2')
# Titles and labels
plt.title('Machine Learning Model Performances', fontsize=16)
plt.xlabel('Models', fontsize=14)
plt.ylabel('Metric Values', fontsize=14)
plt.ylim(0, 1)
plt.axhline(0, color='grey', linewidth=0.8)
# Legend and display
plt.legend(title='Metric', fontsize=12)
plt.xticks(rotation=15)
plt.show()
```