

An approach-based ensemble methods to predict school performance for Moroccan students

Abdallah Maiti¹, Abdallah Abarda², Mohamed Hanini¹

¹Laboratory of Computing, Networks, Mobility and Modelling (IR2M) FST, Hassan First University of Settat, Settat, Morocco

²Laboratory LM2CE, Faculty of Economic Sciences and Management, Hassan First University of Settat, Settat, Morocco

Article Info

Article history:

Received Jan 18, 2025

Revised Apr 9, 2025

Accepted Jul 2, 2025

Keywords:

Deep learning

Ensemble methods

Machine learning algorithms

School performance

Stacking

ABSTRACT

Education is a key factor in Morocco's development, with school performance serving as a critical measure of the education system's quality. However, disparities in student outcomes remain, influenced by socio-economic, demographic, and infrastructural factors. Our study aims to develop a predictive model to assess and improve school performance in Morocco using ensemble machine learning techniques, focusing on the stacking approach. Data from the Massar platform includes variables such as gender, age, type of school, parental occupation, academic results, and residential area. After rigorous data cleaning and preprocessing, a stacking model was created by combining predictions from five base models: random forest, gradient boosting, k-nearest neighbors (KNN), support vector machine (SVM), and multi-layer perceptron (MLP). A random forest meta-model was used to integrate these results. The experimental results of the paper demonstrate the effectiveness of our approach. The stacking model achieved an accuracy of 78.70%, surpassing the individual base models. The meta-model demonstrated strong reliability, achieving an F1 score of 78.62% while reducing false negatives and ensuring balanced predictions. Among the base models, neural networks showed the best performance, achieving the highest predictive accuracy. This research highlights the potential of stacking methods for predicting school performance. Incorporating additional variables, such as parental education and teacher attributes, could further refine the model and enhance Morocco's educational outcomes.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Abdallah Maiti

Laboratory of Computing, Networks, Mobility and Modelling (IR2M) FST

Hassan First University of Settat

Settat, Morocco

Email: Abdallah.maiti@uhp.ac.ma

1. INTRODUCTION

Morocco is vigorously pursuing sustainable development, wherein education occupies a pivotal position among the nation's strategic priorities. Education is acknowledged as a fundamental catalyst for economic and social advancement, contributing substantially to the mitigation of inequalities, the enhancement of life satisfaction, and the fostering of a fair and thriving society. A principal metric employed to evaluate the efficacy of the Moroccan education system is school performance, which serves as a reflection of the success of public policies and their consequent effects on students' academic outcomes and trajectories [1].

This research investigates the complex phenomenon of school performance in Morocco, concentrating on the primary factors that affect student outcomes at three critical educational junctures: the 6th grade of primary education, the 3rd grade of middle education, and the baccalaureate. These educational stages represent crucial milestones in a student's academic progression and yield significant insights into the broader dynamics of the Moroccan education system. Over time, disparities in educational performance have been underscored through national and international evaluations, such as the National Assessment of Student Achievement (PNEA), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PEARL). These assessments provide valuable data regarding the strengths and challenges of the education system, facilitating the monitoring of student outcomes across various phases of their academic experiences. By juxtaposing results on both national and international scales, these assessments assist in pinpointing areas necessitating enhancement and establish a basis for informing educational reforms and policies. A noteworthy trend observed in recent years pertains to the variability in baccalaureate pass rates (which correspond to a “high school diploma” level in the United States or “A-Level” in the United Kingdom). In 2021, the pass rate attained 81.83%, signifying a phase of improvement. Conversely, in 2023, this rate declined to 73.99%, indicative of a marked reduction in performance [2]. This variability implies that advancements may not be consistently maintained and could be inequitably distributed, thereby underscoring deeper systemic issues within the education framework that warrant attention.

Educational inequalities in Morocco are indicative of both systemic and contextual obstacles, encompassing insufficient resources, overcrowded educational settings, and disparate access to high-quality educational opportunities across various regions [3], [4]. In particular, rural locales experience significant deficiencies in infrastructural development, while persistent gender disparities continue to impede the academic advancement of numerous students. In response to these challenges, this study seeks to clarify the synergistic effects of socio-economic, regional, and educational determinants on student performance. The primary aim is to evaluate the manner in which these variables influence academic outcomes by constructing a predictive model through the application of machine learning methodologies. This model is intended to facilitate the early identification of students who are at risk of either failure or success, thereby enabling educational authorities to implement targeted, data-informed interventions.

Recent scholarly articles have endeavored to scrutinize and forecast academic achievement utilizing machine learning methodologies. Zhang *et al.* [5] examined the efficacy of deep learning frameworks in predicting educational results. Although their models demonstrated commendable accuracy, they were marred by limitations related to interpretability and substantial computational demands, rendering them challenging to implement within educational settings characterized by resource constraints. Nafea *et al.* [6] employed ensemble learning techniques, such as random forests and gradient boosting, to enhance predictive robustness. Nevertheless, their methodology did not adequately encompass the socio-economic and contextual variables affecting academic success, thereby constraining the broader applicability of their findings. Concurrently, Moroccan scholars have investigated the influence of local determinants on academic performance. Benjelloun and Bouargane [7] underscored the significance of socio-economic conditions and educational infrastructure in relation to school dropout rates within Morocco. However, their investigation was confined to descriptive statistical assessments, lacking the employment of sophisticated predictive frameworks. Likewise, Elbouknify *et al.* [8] endeavored to integrate supervised classification techniques, including support vector machines (SVMs) and random forests, to predict students at risk, yet their model was hampered by considerable class imbalance, diminishing its efficacy in identifying students facing academic challenges. Additional research, such as that conducted by Cho *et al.* [9], sought to tackle the problem of imbalanced datasets in forecasting academic performance via the SMOTE technique. However, their study fell short of providing a comprehensive comparative analysis of various resampling strategies. In contrast, Deb *et al.* [10] concentrated on predictive models for university-level dropout, thereby leaving a notable gap concerning earlier educational stages.

To address the shortcomings of previous research, our study introduces a more holistic and context-aware methodology adapted to the Moroccan educational environment. We conduct an in-depth analysis of key factors affecting student performance, such as socio-economic status, availability of educational resources, geographic disparities, and individual student characteristics. In contrast to earlier studies that often relied on single algorithms or limited variables, we propose a hybrid machine learning approach. Our methodology integrates several models through ensemble learning techniques, with a particular focus on Stacking. This strategy combines the strengths of various algorithms - such as decision trees, k-nearest neighbors (KNN), SVMs, and neural networks - to enhance both accuracy and robustness. To tackle the issue of class imbalance, which often limits the ability to detect students at risk, we employ the SMOTE technique, ensuring more balanced and reliable predictions. The present investigation is predicated upon data derived from the Moroccan Massar platform, which constitutes a comprehensive educational information system

encompassing detailed academic, demographic, and socio-economic variables. This extensive dataset facilitates an in-depth examination of the complex interrelationships that affect academic outcomes. Our aims are twofold: Initially, we endeavor to construct a representative dataset that integrates both the structural and contextual dimensions inherent in the Moroccan education system. Subsequently, we aim to develop an intelligent predictive model capable of identifying students at increased risk of academic failure. The ultimate objective is to provide actionable insights that can assist policymakers in formulating equitable and efficacious educational interventions.

The subsequent sections of the article are organized as follows: “method” delineates the research methodology, encompassing data collection, preprocessing, and the formulation of the predictive model utilizing ensemble methods; “results and discussion” elucidates the findings of the study, scrutinizing the influence of various determinants on academic performance and examining their ramifications within the framework of Moroccan educational policies; and “conclusion” encapsulates the principal contributions of the study, provides practical recommendations and suggests avenues for future research aimed at improving educational outcomes in Morocco.

2. METHOD

This study follows a systematic and well-organized methodology, starting with data collection and preprocessing, then identifying key features through detailed analysis. The predictive modeling leverages advanced ensemble techniques, particularly stacking, which integrates various basic models. A meta-model combines their outputs to boost accuracy by capitalizing on their complementary strengths.

2.1. Data preprocessing and cleaning

This section details the data preprocessing and cleaning steps applied to student records from the Massar platform, with a focus on preparing the dataset for analyzing the impact of student characteristics on school performance.

2.1.1. Used dataset

The data used of this study is derived from the Massar platform, an integrated information system that modernizes school management across primary, middle, and secondary education institutions in Morocco. This dataset includes a total of 19,115 students. Among these, 9,140 students were successful, while 9,975 failed. The collected data encompasses key parameters essential for analyzing the factors influencing academic success:

- Gender: student's gender.
- Age: age of the student.
- Educational level: students' stage of education (e.g., primary, middle school, or high school).
- Educational cycle: classification of the institution based on the level of education it provides (primary school, middle school, or high school).
- Type of institution: legal status of the school (public or private).
- Parents' occupations: father's and mother's professions, serving as indirect indicators of the family's socioeconomic status.
- Residential area: the student's living environment (rural or urban).
- School performance: the final result of the student, used as the target variable (successful or failed).

It is important to note that the educational levels of the parents are not directly available on the Massar platform. However, parental occupations are used as indirect indicators for this variable.

The primary aim of this study is to evaluate the impact of personal, educational, and social characteristics on the school performance of Moroccan students. Simultaneously, a predictive model is developed based on these data to identify the key determinants of success. These features, identified in the literature as significantly influential [11]-[13] provide insights into ways to optimize student performance and enhance the effectiveness of the national education system.

2.1.2. Data cleaning and preprocessing

Data cleaning and preprocessing constitute fundamental components in the realms of data analysis and machine learning, serving to ensure the quality, consistency, and reliability of the dataset, all of which are imperative for efficacious modeling [14], [15]. These methodologies contribute to the mitigation of biases and errors, thereby enhancing the discernment of significant patterns and the development of robust predictive models. The preprocessing phase generally encompasses the following procedures:

- Handling missing values: the issue of absent data is rectified through imputation methodologies to attenuate the impact of incomplete information. The SimpleImputer class from the scikit-learn library is

employed utilizing the most_frequent strategy, substituting missing entries with the mode of the corresponding feature, thus minimizing the distortion of the dataset's distribution [16].

- Converting categorical variables to indicator variables: categorical variables undergo transformation into numerical representations via the generation of dummy variables. This process is facilitated by the `pd.get_dummies()` function, which produces distinct columns for each category, rendering the data amenable to machine learning models [17].
- Label encoding: label encoding is applied to ordinal or target variables by converting categorical labels into numerical values. This transformation is performed using the `LabelEncoder` class from `scikit-learn`, enabling machine learning algorithms to effectively process categorical target variables.
- Data normalization: data normalization is executed through the utilization of the `StandardScaler` class from `scikit-learn` to standardize the scales of the features. This process guarantees that each feature attains a mean of 0 and a standard deviation of 1, which is essential for algorithms that exhibit sensitivity to feature magnitudes, thereby averting biases arising from disparate scales [18].

By systematically addressing these dimensions of preprocessing, the data is rendered adequately prepared for analytical purposes. This not only bolsters the reliability and interpretability of the results but also lays a robust groundwork for the advancement of accurate and generalizable predictive models.

2.2. Operating principle

Stacking, also referred to as “stacked generalization,” constitutes an ensemble methodology that amalgamates predictions derived from multiple foundational models by employing a meta-model, thereby augmenting both precision and resilience. This approach harnesses the complementary strengths of the foundational models through a two-tiered hierarchical framework: the lower and higher levels.

- Lower level: a diverse array of foundational models, known as base learners, are trained in isolation on the identical training dataset. These models encompass a broad spectrum of algorithms, including decision trees, neural networks, SVMs, KNNs, random forest, gradient boosting, and multi-layer perceptron (MLP) algorithms. Each individual model produces predictions for the test set, which subsequently serve as novel features for the subsequent level.
- Higher level: the predictions generated by the base models are combined to form a new dataset. A meta-model, or meta-learner, is then trained on this dataset to learn how to optimally combine the results of the base models. A key advantage of stacking is its ability to capture diverse and complementary aspects of the data, thereby reducing bias and variance in individual models, while minimizing the risk of overfitting.

2.3. Proposed models

This section presents a structured and methodical approach for implementing machine learning techniques with the aim of improving predictive performance. The methodology ensures that each phase of the project is carried out in a coherent and efficient manner, enabling both the rigorous evaluation of individual models and their integration into a unified ensemble through stacking. The main stages of the proposed pipeline are outlined below:

- i) Data division: the dataset was meticulously partitioned into two subsets to facilitate effective model training and accurate performance assessment. Specifically, 80% of the data was designated for training the machine learning models, thereby providing ample information for the learning process. The remaining 20% was allocated for testing, thereby ensuring an impartial evaluation of the models' performance on previously unseen data. This division reconciles the necessity for robust training with precise performance evaluation.
- ii) Base models: a total of five machine learning algorithms were identified as foundational models, establishing the groundwork for the stacking methodology. Each model was trained independently on the training dataset, yielding a diverse spectrum of predictive capabilities. The selected models are as follows:
 - Random forest classifier (RFC): this technique amalgamates the predictions of multiple decision trees to enhance overall accuracy and diminish variance. By generating random subsets of the data for each tree, RFC guarantees diverse decision-making, which aids in averting overfitting and improves performance on complex datasets [19].
 - Gradient boosting classifier (GBC): this classifier enhances predictions by iteratively rectifying the errors of preceding models. Each subsequent tree is trained to minimize the residuals from earlier models, enabling the classifier to elucidate complex relationships within the data. It is particularly effective in applications requiring high levels of accuracy and precision in classification [20].

- KNN: this non-parametric approach classifies a data point based on the predominant class among its nearest neighbors. It is particularly adept at modeling nonlinear relationships within low-dimensional data spaces. Its simplicity and efficacy render it a favored option for localized classification endeavors [21].
- SVMs: these algorithms are constructed to identify the optimal hyperplane that demarcates classes with the maximum margin, ensuring a distinct separation within the feature space. This attribute is especially advantageous in classification tasks where the classes are well-defined yet may exhibit overlap [22].
- MLPs: these artificial neural networks (ANNs) consist of multiple hidden layers situated between the input and output layers. They possess the capacity to learn complex representations of data through these hidden layers, enabling them to model intricate nonlinear relationships. This versatility renders MLPs highly effective for managing diverse and heterogeneous datasets [23].

The predictions from the five base models were combined into a new dataset to leverage the strengths of each model. This ensemble was then used as input for the meta-model, which refines the final decision, improving accuracy and robustness. The approach enhances performance by reducing individual model errors and increasing the generalization ability of the final model, as shown in Figure 1. The RFC was selected as the meta-model due to its efficiency in handling complex data. By aggregating multiple decision trees, it reduces biases and errors from base models, boosting prediction accuracy. Although logistic regression and extreme gradient boosting (XGBoost) were tested, random forest outperformed them in cross-validation, making it the best choice. To evaluate generalization, the meta-model was applied to the test set for prediction generation. A 10-fold cross-validation ensured a more reliable evaluation and minimized overfitting. Performance was assessed using a confusion matrix, calculating precision, recall, F1 score, and accuracy.

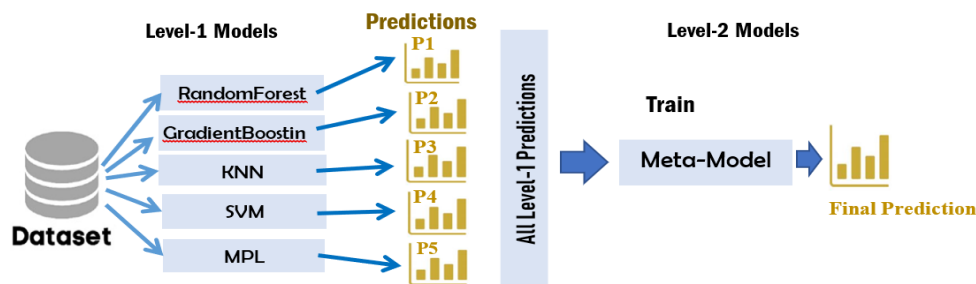


Figure 1. Architecture of the proposed model

3. RESULTS AND DISCUSSION

3.1. Presentation of the modeling results

This study used a stacking ensemble approach to improve academic performance prediction by combining multiple machine learning models. Specifically, the predictions of five baseline models (random forest, gradient boosting, KNN, SVM, and MLP) were aggregated by a random forest-based meta-model. The performances of these models was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score, as shown in Figure 2, which compares these metrics across different algorithms. These metrics, calculated from confusion matrices as shown in Figure 3, are summarized in Tables 1 and 2. Among the baseline models, MLP achieved the highest accuracy at 78.60%, followed by random forest (77.50%). Gradient boosting, KNN, and SVM achieved slightly lower accuracies of 76.20%, 76.00%, and 77.00%, respectively. All models demonstrated relatively balanced classification of students between success (Class 1) and failure (Class 0), although variations in recall highlighted some uncertainty in detecting at-risk students.

Training the meta-model (random forest) on the baseline model predictions led to a modest but notable improvement in performance, with the final accuracy reaching 78.70%, slightly surpassing that of the best-performing baseline model (MLP at 78.60%). For Class 0 (failure), the meta-model achieved a precision of 80.54% and a recall of 77.66%. For Class 1 (pass), the precision and recall were 76.71% and 79.67%, respectively. This improvement highlights the effectiveness of stacking in reducing classification errors and compensating for the weaknesses of individual models. In particular, the meta-model optimized recall, improving the identification of students at risk of failure and reducing false negatives, an important factor in school settings.

However, the success of the stacking approach depends on the quality of the base models. If the base models produce highly correlated errors or exhibit poor individual performance, the meta-model may not completely overcome these limitations. Analysis of our model performance highlights the relevance of ensemble learning for predicting academic achievement in the Moroccan educational context. By combining several base models (random forest, gradient boosting, KNN, SVM, and MLP) in a random forest-based meta-model, we achieved an overall accuracy of 78.70% and a recall of 79.89% for class 1 (high-performing students). These results suggest that model aggregation improves predictive robustness and mitigates errors inherent in individual algorithms. However, the absence of key explanatory variables, such as parental education or extracurricular activities, may explain why our model's performance remained slightly below 80%.

Our research indicates that academic achievement is influenced by a confluence of individual, institutional, and familial factors. An examination of misclassification patterns indicates that students exhibiting intermediate profiles - those whose attributes reside within the spectrum of success and failure - present the most significant challenges in terms of predictive accuracy. This observation underscores the intricate nature of educational outcomes and emphasizes the necessity for further investigation into additional variables and methodologies in subsequent research endeavors. Although the model demonstrates a commendable performance, with an accuracy rate falling short of 80%, such outcomes can be ascribed to a multitude of factors. Firstly, the complexity inherent in forecasting academic success is compounded by the myriad influences that remain unaccounted for within our dataset. Psychological dimensions, personal motivation, familial support, and the educational environment are all pivotal elements that were excluded from the analytical framework. Moreover, classification challenges stem from the convergence in feature distributions between cohorts of successful and failing students. Unlike tasks characterized by distinctly separated categories, the attributes of students in both classifications frequently overlap, engendering considerable uncertainty in predictive outcomes.

Furthermore, the omission of critical variables such as parental educational attainment, sibling count, or educator qualifications constrains the model's capacity to thoroughly encapsulate the complexities associated with academic success. This deficiency in data may elucidate the reasons behind the accuracy remaining below the 80% threshold, despite the utilization of sophisticated machine learning methodologies. Nonetheless, the model fulfills its principal objective: to develop a dependable instrument for identifying students at risk of academic underachievement. With an achieved accuracy of 78.70%, the model provides significant insights, albeit without guaranteeing flawless classification. In addition to predictive capabilities, the study elucidates essential trends and determinants affecting academic success. The performance analysis accentuates the constraints imposed by the available data and the critical nature of the absent variables. These findings indicate potential avenues for future inquiry, particularly with respect to more exhaustive data collection efforts. While the model is not without its shortcomings, it represents a valuable asset for educators and policymakers, enabling the early identification of students requiring targeted interventions.

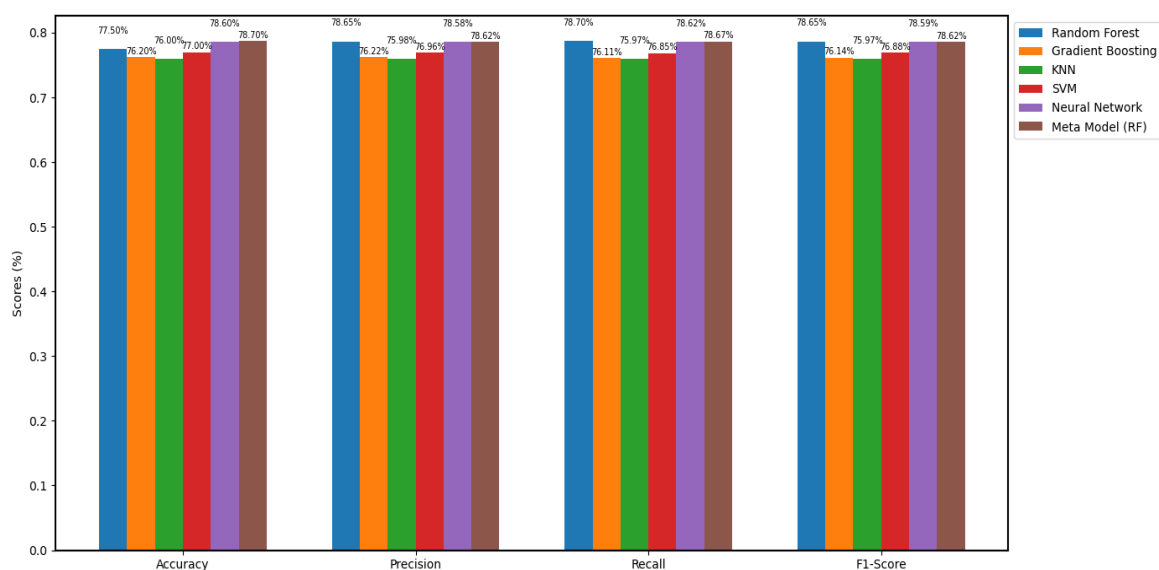


Figure 2. Comparative evaluation of classification algorithms across different metrics

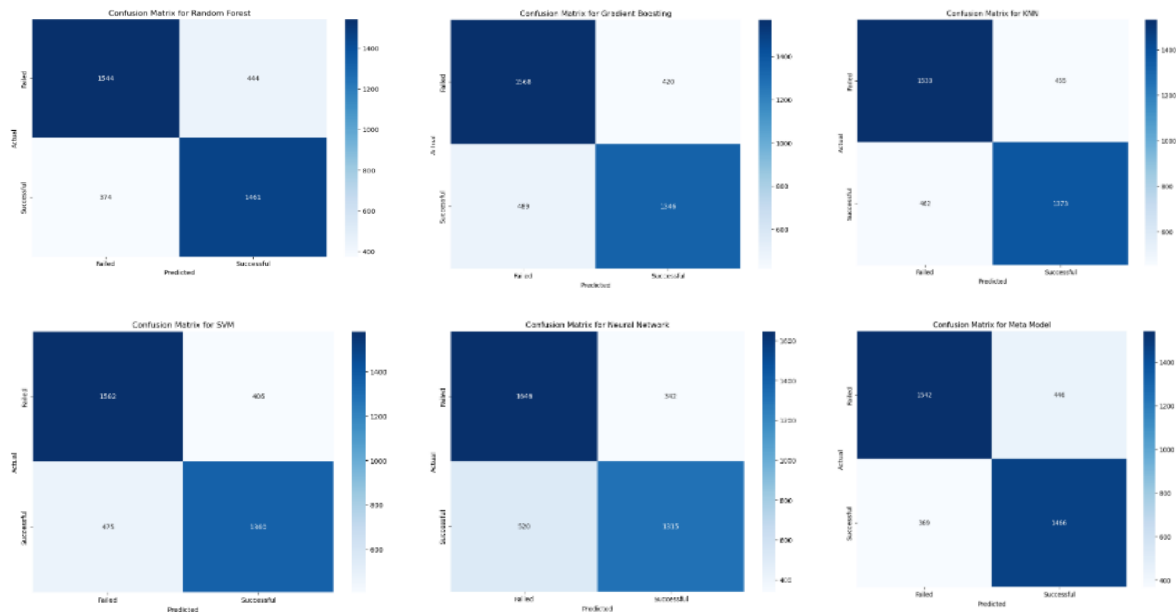


Figure 3. Confusion matrix obtained by the proposed approach

Table 1. Results obtained by lower level classifiers in terms of accuracy, precision, recall, and F1 score

Classifiers	Classes	Accuracy	Precision	Recall	F1 score
Random forest	0	0.775	0.805846	0.77666	0.790984
	1		0.767174	0.797275	0.781935
Gradient boosting	0	0.762	0.762275	0.788732	0.775278
	1		0.762174	0.733515	0.74757
KNN	0	0.760	0.768421	0.771127	0.769772
	1		0.751094	0.748229	0.749659
SVM	0	0.770	0.769081	0.795775	0.7822
	1		0.770102	0.741144	0.755346
Neural network (MLP)	0	0.786	0.801236	0.782696	0.79185
	1		0.770335	0.789646	0.77987

Table 2. Classification report for the meta model

Model	Classes	Accuracy	Precision	Recall	F1 score
Meta model	0	0.787	0.805425	0.77666	0.790781
(Random forest)	1		0.767051	0.79673	0.781609

3.2. Comparison of our approach with the literature

Research on predicting academic performance in Morocco remains relatively scarce, making our study valuable due to its adaptation to the local context and its innovative methodological approach. Our research is distinguished by a context-aware and novel framework that incorporates a diverse set of variables while leveraging a two-level ensemble learning strategy to predict student success. Compared to previous studies, our approach reveals key differences in terms of feature selection, algorithm choices, and performance metrics, specifically regarding accuracy, precision, and recall [24], [25].

Abubakaria *et al.* [26] focused on behavioral and academic factors such as parental involvement, school satisfaction, and absenteeism. Their ANN-based model achieved a stable accuracy of 76.8%, optimized using Adam and stochastic gradient descent (SGD). However, their study relied primarily on conventional contextual indicators, limiting its scope. In contrast, Ouatik *et al.* [27] incorporated a broader feature set, including personal attributes (gender, age), academic indicators (grade point average (GPA), grades), psychological aspects, and interactions within virtual learning environments (VLEs). Their application of SVM, KNN, and C4.5 resulted in a maximum accuracy of 87% using the sequential minimal optimization (SMO)-SVM algorithm. Hajar *et al.* [28] primarily analyzed academic and behavioral variables, such as classroom participation and access to educational resources. Their XGBoost model outperformed logistic regression, achieving an accuracy of 84.38% compared to 82.29% for logistic regression. However, they did not provide a detailed breakdown of model performance for different student groups, which limits

insights into the model's strengths and weaknesses in predicting success and failure. Similarly, Yagci [29] and Asad *et al.* [30] reported relatively lower performance, with decision tree and neural network models reaching a maximum accuracy of 70.8%. These results suggest that their feature selection and model choices may not have been optimal for capturing the complexity of student success predictors.

Rogers *et al.* [31] utilized Moodle log data to identify key predictive indicators, including submission activity and quiz participation. Their random forest model achieved an area under the curve-receiver operating characteristic (AUC-ROC) score of 0.77 in training and 0.73 in testing, demonstrating the potential of tree-based models for educational data analysis. However, their reliance on Moodle-specific variables limits the applicability of their findings to broader academic settings. Noor *et al.* [32] adopted a qualitative approach, examining students' perceptions of mobile learning. They identified factors such as ease of use (69%) and interactivity (53%) as influential, but their methodology lacks predictive power, as it does not employ machine learning models to forecast academic success. Waheed *et al.* [33] explored deep learning techniques, particularly convolutional neural networks (CNNs) and long short-term memory (LSTM) models, to predict academic success based on student engagement in online learning platforms. Their study reported an accuracy of 81.2%, but their precision and recall scores varied significantly across different student groups, indicating potential class imbalance issues.

Our investigation differentiates itself by incorporating a distinctive amalgamation of individual characteristics (such as gender and age), contextual determinants (including school type and rural versus urban environments), and familial variables (pertaining to parental occupations), all specifically adapted to the Moroccan educational framework. From a methodological perspective, we adopted a two-tiered ensemble learning architecture, wherein foundational models (encompassing random forest, gradient boosting, KNN, SVM, and MLP) were synthesized into a meta-model predicated on random forest. This methodology leverages the synergistic advantages of these models, culminating in an aggregate accuracy of 78.70%, a precision of 80.54%, and a recall of 77.67% for the first class. While certain studies have reported elevated accuracy metrics (84.38% and 87%, respectively), their selection of features and datasets lacks diversity, potentially constraining their applicability to broader contexts. In contrast, our research demonstrates competitive efficacy while ensuring enhanced contextual significance and wider applicability through its extensive feature array and sophisticated ensemble methodologies. In comparison to deep learning methodologies [32], [33], our approach provides enhanced interpretability and achieves an equitable balance between precision and recall, thereby facilitating more dependable classifications of both high-achieving and at-risk students. These findings highlight the significance of our contribution and provide a foundation for future research to expand datasets, incorporate additional variables, and explore even more advanced methodological frameworks.

4. CONCLUSION, LIMITATIONS, AND PERSPECTIVES

This research emphasizes the efficacy of stacking methodologies in forecasting academic achievement in Morocco. By combining multiple classification models into a random forest-based meta-model, our approach achieved strong predictive accuracy, successfully identifying students at risk while reducing false negatives. These results demonstrate the benefits of ensemble learning for educational data analysis, especially in environments where early intervention can play a critical role in improving student outcomes. Nonetheless, certain limitations must be acknowledged. The lack of important variables, such as parental education levels, number of siblings, and teacher-related factors like experience and training, may have constrained the model's capacity to fully capture the range of influences on student performance. This limitation reflects broader challenges tied to data availability and quality, which continue to hinder research in the field of education.

Beyond predictive accuracy, the study opens new paths for exploration and enhancement. Expanding the dataset to include richer socioeconomic and pedagogical indicators could lead to more refined predictions and a deeper understanding of the drivers of academic success. Moreover, exploring alternative modeling strategies, such as deep learning or hybrid boosting techniques, may further improve performance and adaptability across diverse student populations. Ensuring fairness in predictive models also emerges as a key priority. Tackling potential biases in both data and algorithms is vital to ensure that all students receive equitable consideration, regardless of their background or location. Ultimately, this research lays the groundwork for broader discussions about the role of artificial intelligence in education, with the aim of informing more inclusive, context-sensitive policy decisions that support student achievement.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abdallah Maiti	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Abdallah Abarda	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Mohamed Hanini	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [A. M.]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.




REFERENCES

- [1] J. Amaghouss and M. Zouine, "A critical analysis of the governance of the Moroccan education system in the era of online education," in *Socioeconomic Inclusion During an Era of Online Education*, IGI Global, 2022, pp. 156–176.
- [2] M. Bettah and A. Abbaia, "The determinants of school performance in selected MENA countries: what role of the education system's characteristics?," *revistamultidisciplinar.com*, vol. 6, no. 1, pp. 17–38, Jan. 2024, doi: 10.23882/rmd.24178.
- [3] P. Barrett, A. Treves, T. Shmis, D. Ambasz, and M. Ustinova, *The impact of school infrastructure on learning: a synthesis of the evidence*. Washington, DC: World Bank, 2019.
- [4] L. Darling-Hammond, "Inequality in teaching and schooling: how opportunity is rationed to students of color in America," *The Right Thing to Do, The Smart Thing to Do*, pp. 208–233, 2001, [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK223640/#ddd00122>.
- [5] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, "Educational data mining techniques for student performance prediction: method review and comparison analysis," *Frontiers in Psychology*, vol. 12, Dec. 2021, doi: 10.3389/fpsyg.2021.698490.
- [6] A. A. Nafea, M. Mishlish, A. M. haban Shaban, M. M. AL-Ani, K. M. A. Alheeti, and H. J. Mohammed, "Enhancing student's performance classification using ensemble modeling," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 4, pp. 204–214, 2023.
- [7] R. Bentaibi and N. Bouargane, "Contribution to the understanding of the phenomenon of school dropout in the Moroccan rural context (in French: Contribution à la compréhension du phénomène du décrochage scolaire dans le contexte rural Marocain)," *Revue Marocaine de l'Évaluation et de la Recherche Educative*, vol. 8, no. 8, pp. 301–317, 2022.
- [8] I. Elbounkify *et al.*, "AI-based identification and support of at-risk students: a case study of the Moroccan education system," *arxiv preprint: 2504.07160*, Apr. 2025, [Online]. Available: <http://arxiv.org/abs/2504.07160>.
- [9] C. H. Cho, Y. W. Yu, and H. G. Kim, "A study on dropout prediction for university students using machine learning," *Applied Sciences*, vol. 13, no. 21, p. 12004, Nov. 2023, doi: 10.3390/app132112004.
- [10] S. Deb, M. S. R. Sammy, A. N. Tusher, M. R. S. Sakib, M. F. Hasan, and A. I. Aunik, "Predicting student dropout: a machine learning approach," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jun. 2024, pp. 1–7, doi: 10.1109/ICCCNT61001.2024.10726161.
- [11] J. Liu, P. Peng, and L. Luo, "The relation between family socioeconomic status and academic achievement in China: A Meta-analysis," *Educational Psychology Review*, vol. 32, no. 1, pp. 49–76, 2020, doi: 10.1007/s10648-019-09494-0.
- [12] Koukouch, "The Moroccan school system: In the absence of family commitment, public school, remains 'selective' and 'meritocratic' if not 'exclusive' (in French: Le système scolaire marocain: En l'absence de l'engagement familial, l'école publique, demeure «sélective» et «méritocratique» si ce n'est pas «exclusive»), *Revue Marocaine de l'Évaluation et de la Recherche en Éducation*, 2021.
- [13] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: a statistical and data mining approach," *International Journal of Computer Applications*, vol. 63, no. 8, pp. 35–39, 2013, doi: 10.5120/10489-5242.
- [14] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, Apr. 2019, vol. 2019-April, pp. 2086–2090, doi: 10.1109/ICDE.2019.00245.
- [15] A. Rahman, "Statistics-based data preprocessing methods and machine learning algorithms for big data analysis," *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 44–65, 2019.
- [16] J. W. Grzymala-Busse and W. J. Grzymala-Busse, "Handling missing attribute values," in *Data Mining and Knowledge Discovery Handbook*, 2006, pp. 37–57.
- [17] P. Kedar, P. Taher S, and P. Chinmay D, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, 2017.
- [18] D. Borkin, A. Némethová, G. Michalčonok, and K. Maiorov, "Impact of data normalization on classification model accuracy," *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, vol. 27, no. 45, pp. 79–84, 2019, doi: 10.2478/rput-2019-0029.
- [19] L. Andy and W. Matthew, "Classification and regression by randomforest," *R News*, vol. 2, pp. 18–22, 2002.




- [20] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Aug, pp. 785–794, doi: 10.1145/2939672.2939785.
- [21] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for KNN classification," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 3, pp. 1–19, May 2017, doi: 10.1145/2990508.
- [22] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare applications: a review," *Information*, vol. 15, no. 4, p. 235, Apr. 2024, doi: 10.3390/info15040235.
- [23] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons," in *Computational intelligence: a methodological introduction*, Springer, 2022, pp. 53–124.
- [24] H. Mohamed and S. Md Nasir, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, 2015.
- [25] A. Maiti, A. Abarda, M. Hanini, and A. Oussous, "An optimal model combining SqueezeNet and machine learning methods for lung disease diagnosis," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 20, no. 1, p. e15734056258742, Jan. 2024, doi: 10.2174/0115734056258742230920062315.
- [26] M. S. Abubakari, F. Arifin, and G. G. Hungilo, "Predicting students' academic performance in educational data mining based on deep learning using TensorFlow," *International Journal of Education and Management Engineering*, vol. 10, no. 6, pp. 27–33, 2020, doi: 10.5815/ijeme.2020.06.04.
- [27] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Students' orientation using machine learning and big data," *International journal of online and biomedical engineering*, vol. 17, no. 1, pp. 111–119, 2021, doi: 10.3991/ijoe.v17i01.18037.
- [28] A. Hajar, J. Adil, and Y. Ali, "Predicting factors affecting student's performance in a learning management system," *Indian Journal of Computer Science and Engineering*, vol. 12, no. 6, pp. 1771–1779, 2021, doi: 10.21817/indjcs/2021/v12i6/211206015.
- [29] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
- [30] R. Asad, S. Arooj, and S. U. Rehman, "Study of educational data mining approaches for student performance analysis," *Technical Journal*, vol. 27, no. 1, pp. 68–81, 2022.
- [31] J. K. Rogers, T. C. Mercado, and R. Cheng, "Predicting student performance using Moodle data and machine learning with feature importance," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 37, no. 1, pp. 223–231, Jan. 2025, doi: 10.11591/ijeecs.v37.i1.pp223-231.
- [32] A. S. M. Noor, M. N. Y. Atoom, and M. A. Jalil, "Toward mobile learning at Jordanian higher education institutions," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 3, pp. 1538–1545, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1538-1545.
- [33] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, p. 106189, Mar. 2020, doi: 10.1016/j.chb.2019.106189.

BIOGRAPHIES OF AUTHORS






Abdallah Maiti    holds a Ph.D. in artificial intelligence and statistics from the Faculty of Science and Technology of Settat, Hassan First University, Morocco. He is a Moroccan statistical engineer and graduate of the Institut National de la Statistique et de l'Économie Appliquée de Rabat (INSEA). Throughout his academic and research career, he has contributed to the field by publishing several articles in indexed journals, notably in machine learning, deep learning and computer vision. He has participated in national and international conferences. He can be contacted at email: maiabdel@gmail.com.



Abdallah Abarda    serves as a professor of statistics and data analysis within the Faculty of Economics and Management at Settat. Since the year 2024, he has assumed the role of Director of the Laboratory of Mathematical Modeling and Economic Computing (LM2CE). He possesses a degree in Statistical Engineering from INSEA, located in Rabat, as well as a PhD in Statistics conferred by Ibn Tofail University. His scholarly research is concentrated on statistical analysis and artificial intelligence, and he has authored in excess of 40 scientific articles published in indexed journals. He presides over the International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI) and has participated as a member of the Technical Program Committee (TPC) and organizing committee for a variety of international conferences and workshops. He can be contacted at email: abardabdallah@gmail.com.



Mohamed Hanini    Professor in the Department of Mathematics and Computer Science at the Faculty of Science and Technology. He obtained his Ph.D. in the fields of mathematics and computer science in 2013. He has made substantial contributions, as author and co-author, to a series of scientific publications in the fields of modeling and performance evaluation of communication networks, cloud computing, network security and artificial intelligence. His academic involvement includes participation in a multitude of international conferences, where he has assumed the responsibilities of Technical Program Committee (TPC) member and Organizing Committee member for various international conferences and workshops, in addition to fulfilling the role of reviewer for several international journals. He can be contacted at email: mohamed.hanini@uhp.ac.ma.