MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 4, pp. 1135–1144 (2025)

Predicting Student Performance in Moroccan Secondary Education: A Machine Learning Framework for Academic Pathway Guidance

Sammah S.[1], Ait Daoud M.[1,2], Achtaich K.[1], Tragha A.[1]

[1] *LTIM, Department of Computer Science, Faculty of Sciences Ben M'sick, Hassan II University of Casablanca, Morocco*
[2] *ORDIPU, Faculty of Sciences Ben M'sick, Hassan II University of Casablanca, Morocco*

(Received 28 March 2025; Revised 19 October 2025; Accepted 25 October 2025)

This study addresses the lack of region-specific tools for academic counseling in Morocco by proposing a machine learning framework to predict student performance across secondary education pathways. Using academic records of students from the Greater Casablanca region, we evaluate four models – Random Forest, Support Vector Machine (SVM), Decision Tree, and Linear Regression – following a methodology that integrates data preprocessing, feature selection, and synthetic data enrichment to address class imbalance. The Random Forest algorithm achieved an accuracy rate of 75.20%, significantly outperforming the other models. By linking predictive outcomes to actionable academic guidance, the proposed framework enables educators to recommend pathways tailored to individual student strengths, thus addressing a critical gap in Morocco's education system.

Keywords: *data mining; student performance; predictive modeling.*
2010 MSC: 68T05 **DOI:** 10.23939/mmc2025.04.1135

1. Introduction

Education plays a fundamental role in the development of individuals as well as society as a whole [1]. In today's context, data analysis is increasingly essential in various sectors, including education. This study, positioned at the intersection of data science and big data analytics, explores how technological advancements can be leveraged to forecast and enhance student performance. By analyzing academic results over multiple sessions, the primary objective is to develop a predictive model capable of projecting students' future performance based on their past results. More specifically, this study focuses on predicting success in various fields using results from two consecutive academic sessions.

Advancements in data science have profoundly transformed our approach to information, enabling the extraction of meaningful insights from large datasets [2]. This study aims to apply these techniques in the educational domain to assist teachers, institutions, and students in making more informed decisions. By employing predictive models in this context, it becomes possible to identify key factors influencing academic performance and implement proactive measures to improve student outcomes.

The approach adopted in this study is based on the analysis of a database containing students' academic performance over three consecutive sessions. Each record includes the grades obtained in various subjects during the corresponding session. Using this data, the predictive model is designed to estimate students' future performance in specific subjects based on their past results. This tool may serve as a valuable asset for students, teachers, and academic advisors, supporting data-driven decisions that optimize academic success and guide career pathways more effectively.

2. Literature review

The field of education has been significantly influenced by the advent of big data, data analytics, and machine learning. These technological advancements are revolutionizing how educators and researchers approach the enhancement of educational processes and academic performance prediction. By leveraging data-driven methodologies, researchers are expanding traditional boundaries allowing for more

refined analyses of student data and forecasting of academic outcomes. Just as predictive analytics are crucial in business decision-making, these tools enable more precise and proactive assessments of educational trajectories.

The significance of this technological shift is particularly evident in academic performance prediction. For instance, [3] analyzed the efficacy of algorithms such as Naïve Bayes, Decision Trees, and Multilayer Perceptron in forecasting student success. Their study utilized data from a University of Tuzla survey and an enrollment database, employing Weka software to evaluate classifier performance. Their findings showed that the Naïve Bayes algorithm achieved 76.65% accuracy with a short training time but relatively high error rates.

In a related study, [4] applied knowledge discovery methods to develop an educational intervention plan aimed at reducing dropout rates by 14%. By leveraging logistic regression models with activity grades as predictive parameters, their approach utilized an iterative procedure evaluating student performance weekly. Their proposed model, LOGIT-Act, demonstrated superior performance compared to SVM, FFNN, PESFAM, and SEDM, achieving accuracy, precision, recall, and specificity of 97.13%, 98.95%, 96.73%, and 97.14%, respectively.

Similarly, [5] introduced a machine learning framework to identify students at risk of not graduating. Utilizing data from two schools across two districts, they experimented with five machine learning models: SVM, Random Forest, Logistic Regression, AdaBoost, and Decision Tree. Their results indicated that Random Forest yielded the highest predictive accuracy, ranking students based on estimated risk scores.

At UniSZA University in Malaysia, [6] adopted machine learning techniques to predict first-year student performance. Their dataset included nine attributes such as gender, race, GPA, and family income, covering 399 students between 2006 and 2014. By evaluating Decision Tree, rule-based classifiers, and Naïve Bayes, they found that rule-based classifiers yielded the highest accuracy of 71.3%.

Feature selection methods were central to the study by [7], which aimed to identify students at risk in a norm-based grading system. Their analysis utilized six machine learning classifiers, achieving 88% accuracy with a Naïve Bayes classifier when using 16 selected features. Likewise, [8] compared SVM and KNN classifiers on data from the University of Minho, which contained 33 attributes, demonstrating that SVM achieved high accuracy via cross-validation experiments.

Further research by [9] introduced an RTV-SVM classifier to predict at-risk students based on academic performance. This algorithm achieved high accuracy rates of 93.8% and 93.5% for predicting at-risk and marginal students, respectively, while reducing training time by 59%.

Another comparative study [10] evaluated Support Vector Machines (SVM) and Artificial Neural Networks (ANN) on a dataset of 6130 students. Their results showed that SVM achieved an accuracy of 84.54% in predicting academic performance. Additionally, [11] compared eight machine learning models to predict student performance in an Indian technical college, with Random Forest demonstrating the highest accuracy at 93.8%. Similarly, [12] investigated resampling techniques for predicting student dropout, concluding that Random Forest combined with SVM-SMOTE balancing achieved the best accuracy of 77.97%.

Fuzzy neural networks were explored by [13], who integrated metaheuristic optimization techniques based on gas solubility to enhance early performance prediction. This innovative approach yielded an accuracy of 96.04%, demonstrating its potential in academic forecasting.

Furthermore, [14] analyzed the effectiveness of Decision Trees (DT), K-Nearest Neighbors (KNN), and Random Forest (RF) in predicting student performance using self-generated data. Their approach resulted in an average accuracy of 75%, reinforcing the effectiveness of combining multiple algorithms for enhanced predictive accuracy.

Artificial Neural Networks (ANN) were the focus of [15], which examined the retention rates of first-year students at Columbus State University between 2005 and 2010. Their findings showed that a two-layer ANN achieved an accuracy of 89% in predicting second-year retention.

Moreover, [16] developed a predictive model to identify student dropouts using decision tree variants, where ID3 achieved the highest accuracy of 90.9%, outperforming C4.5, CART, and ADT.

Another study by [17] emphasized time-dependent variables in predicting student performance in online learning environments. Using data from 330 students from a Learning Management System (LMS), they compared CART, Logistic Regression, and AdaBoost, with CART achieving over 95% accuracy.

Additionally, [18] reviewed data mining techniques for dropout prediction, using data from 189 students and feature selection via genetic algorithms. They found that the 3NN classifier was most effective, achieving an accuracy of 87%.

In an online learning context, [19] examined ANN, Decision Trees, and Bayesian Networks across 62 375 students, concluding that Decision Trees provided superior predictive efficiency.

Further research by [20] focused on human-interpretable features for predicting low academic performance. Using data from the University of Minnesota, they applied SVM, Random Forest, Gradient Boosting, and Decision Trees, achieving an accuracy exceeding 75% in detecting at-risk students.

In Morocco, [21] developed an acceptance model for the e-orientation platform "orientation-chabab.com," testing four algorithms (Naïve Bayes, J48, NLMT, SimpleLogistic) with WEKA software, where J48 yielded the highest classification accuracy. Another Moroccan study by [22] explored learning differences in blended learning environments using data from the FOAD_FSBM e-learning platform, emphasizing statistical analyses of student performance.

Additionally, a study by [23] proposed developing an automatic machine learning tool to analyze student performance and guide improvement suggestions, addressing the challenge of diverse performance levels among students.

Research conducted by [24] utilized a dataset from higher education to evaluate various machine learning algorithms, including DT, SVM, and boosting algorithms. The study found that boosting algorithms, particularly LightGBM and CatBoost, outperformed traditional classifiers. Furthermore, [25] examined the potential of data from LMS platforms, specifically Moodle, to predict student performance by analyzing behavioral data. Several machine learning techniques were applied, with multilayer perceptron neural networks (MLPNNs) achieving the highest accuracy of 93%. In another research effort, [26] aimed to find a decision tree alternative to Random Forest, achieving 97% precision using Decision Tree evaluations, compared to Random Forest's 93%. Lastly, [27] investigated the relationship between midterm and final exam grades using machine learning models, achieving classification accuracies between 70%–75%.

In summary, this research emphasizes the need for optimal algorithm selection to accurately predict student performance and success. Notably, models based on DT, SVM, and Random Forest demonstrate high accuracy, leveraging the complexity of the data. Given these findings, exploring hybrid approaches that combine these strengths could lead to robust predictions of student performance across various fields.

3. Research methodology

To facilitate the understanding of our system, we have developed a visual representation of its architecture. The image below (see Figure 1) highlights the organization of the various modules and key components that interact synchronously to achieve our goals. It provides a valuable overview of the system's underlying structure, enabling a better understanding of its complexity and internal functioning.

3.1. Data collection

As part of this study, we combined five distinct datasets, covering a period of five years, from 2012 to 2016. The first step in our approach was to collect data from the ERP (Enterprise Resource Planning) database of the HSI, specifically for the Grand Casablanca region. Each dataset, corresponding to one year, contains 41 tables with essential information about students, their grades, their courses, and

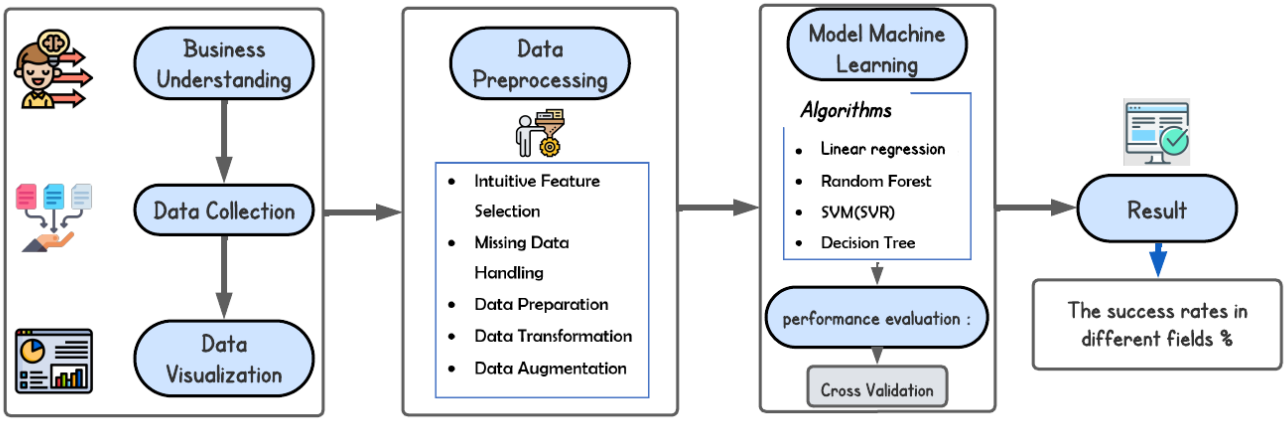

Fig. 1. System architecture.

other relevant data. This information was carefully analyzed to identify the most relevant data for our study. Subsequently, we merged these five datasets in order to systematically group and organize the information. This allowed us to obtain a comprehensive and precise view of the evolution of educational data during the study period.

3.2. Data preprocessing

Data preprocessing [28] is a critical step in exploring and preparing raw data to ensure its usability for analysis. In real-world scenarios, datasets often come with gaps, inconsistencies, and missing essential behaviors or trends. Furthermore, they are prone to various errors, such as outliers or noise, which can significantly affect the accuracy of the analysis. To address these challenges, we applied a comprehensive preprocessing pipeline, as illustrated in Figure 2, to clean and transform our raw dataset into a structured and reliable resource for further analysis.

3.2.1. Intuitive feature selection

Fig. 2. Data preprocessing pipeline.

As part of the data preprocessing stage, we adopt an approach based on intuitive feature selection. This crucial step relies on a deep understanding of business logic and aims to identify the most relevant variables while eliminating those that do not add value to the predictive model.

Rather than relying solely on automated methods, we leverage our domain expertise to analyze complex relationships between different variables. This approach allows us to prioritize features with a direct impact on prediction and avoid including non-significant data.

For example, certain features such as students' first and last names were excluded, as they represent personal information with no influence on academic performance. Similarly, the student's gender was disregarded, as it plays no relevant role in our study context.

Conversely, we retained several essential variables for predictive analysis:

— **Student ID** to ensure individualized performance tracking.
— **Subject name** to study the impact of the learning domain on results.
— **Student grades**, a central element of our analysis.
— **Academic level**, allowing us to examine performance variations across different classes.
— **Academic session**, useful for observing performance trends over multiple years.

By refining the dataset at this initial stage, we enhance the model's efficiency by reducing noise and focusing on variables with a real impact on expected outcomes. This methodology also promotes greater model interpretability, facilitating its adoption in a decision-making context.

3.2.2. Missing data handling

In our data preprocessing process, we have adopted a rigorous approach to handling missing values. Rather than removing them, which could lead to significant information loss, we have chosen to replace them with a constant value, set at 10, across the entire dataset.

This choice ensures data consistency and integrity while minimizing the impact of missing values on model performance. Using a fixed value helps prevent biases introduced by other imputation methods, such as the mean or median, which can distort distribution balance, especially in heterogeneous datasets.

Moreover, this strategy simplifies data analysis and result interpretation by ensuring uniform handling of missing values. It also enhances model stability during training, preventing unpredictable variations that could arise from imputations based on fluctuating statistics.

By integrating this method, we ensure that the dataset remains usable without compromising its analytical quality. Thus, our approach helps optimize the model's performance and robustness while minimizing potential distortions.

3.2.3. Data preparation

Data preparation is an essential step in ensuring the reliability and relevance of subsequent analyses. We implemented a series of operations aimed at structuring and optimizing our dataset, including harmonizing labels, standardizing formats, and removing irrelevant elements.

— **Label Harmonization and Grouping**: One of the major challenges in preprocessing lies in the heterogeneity of designations and categories within the dataset. To ensure better semantic consistency, we harmonized the labels by grouping similar values under a unified nomenclature. For example, different names referring to the same subject were merged: "ARABE", "Langue Arabe", and "LANGUE ARABE" were standardized under a single label, "LANGUE ARABE". Similarly, the designations "HIST-GEO" and "HISTOIRE-GEOGRAPHIE" were standardized as "HISTOIRE GEOGRAPHIE". This homogenization facilitates data interpretation and reduces inconsistencies that could distort statistical analyses or affect the performance of predictive models.

— **Removal of Irrelevant Data**: We also performed targeted filtering of academic levels in the NOMNIVEAU column, retaining only those relevant to the scope of our study. This rigorous selection process eliminates out-of-context observations, ensuring better data homogeneity and reducing the risk of biased analyses.

3.2.4. Data transformation

As part of our study, we performed a series of data transformations to prepare them for predictive analysis. This transformation was carried out in multiple steps to ensure that the data aligned with our analytical objectives.

a) **Selection of Students and Sessions.** We selected a subset of students for whom we have data over three consecutive academic sessions, starting with their first session in TCS (Common Scientific Track). This selection allows us to focus the analysis on students with complete tracking, thereby enhancing the reliability of our predictive model. However, the initial data structure, where each grade was recorded individually by subject and session, made analysis and result interpretation more complex. To ensure better coherence and usability, we transformed the data into an aggregated format.

b) **Data Aggregation and Pivoting.** Initially, each row in the dataset represented an individual grade assigned to a student for a specific subject and session. A single student could therefore appear multiple times with different grades for the same subject, corresponding to distinct assessments (e.g., continuous assessment, final exam) (see Figure 3). This structure made analysis more complex, as it did not provide a consolidated view of students' performance by subject and session. To address this, we applied data pivoting, grouping all grades for the same subject and session into a single column.

```
dat.head(50)
```

	ELEVE	ELEVE_NOM	ELEVE_PRENOM	ELEVE_SEXE	Nom_matiere	NOTES	NOMNIVEAU	SESSION
0	1			MASCULIN	LANGUE ARABE	13.0	1e A.C.	2011-2012
1	1			MASCULIN	PHYSIQUE-CHIMIE	15.0	1e A.C.	2011-2012
2	1			MASCULIN	1ERE LANGUE ETRANGERE	17.0	1e A.C.	2011-2012
3	1			MASCULIN	LANGUE ARABE	12.0	1e A.C.	2011-2012
4	1			MASCULIN	MATHEMATIQUES	12.0	1e A.C.	2011-2012
5	1			MASCULIN	S.V.T	11.0	1e A.C.	2011-2012
6	1			MASCULIN	2EME LANGUE ETRANGERE	10.0	1e A.C.	2011-2012
7	1			MASCULIN	HISTOIRE GEOGRAPHIE	11.0	1e A.C.	2011-2012
8	1			MASCULIN	EDUCATION ISLAMIQUE	16.0	1e A.C.	2011-2012
9	1			MASCULIN	S.V.T	15.0	1e A.C.	2011-2012
10	1			MASCULIN	PHYSIQUE-CHIMIE	15.0	1e A.C.	2011-2012
11	1			MASCULIN	MATHEMATIQUES	16.0	1e A.C.	2011-2012
12	1			MASCULIN	HISTOIRE GEOGRAPHIE	14.0	1e A.C.	2011-2012
13	1			MASCULIN	1ERE LANGUE ETRANGERE	16.0	1e A.C.	2011-2012
14	1			MASCULIN	EDUCATION ISLAMIQUE	15.0	1e A.C.	2011-2012
15	1			MASCULIN	LANGUE ARABE	15.0	1e A.C.	2011-2012
16	1			MASCULIN	LANGUE ARABE	14.0	1e A.C.	2011-2012
17	1			MASCULIN	S.V.T	12.0	1e A.C.	2011-2012
18	1			MASCULIN	MATHEMATIQUES	11.0	1e A.C.	2011-2012
19	1			MASCULIN	EDUCATION ISLAMIQUE	9.0	1e A.C.	2011-2012
20	1			MASCULIN	2EME LANGUE ETRANGERE	16.0	1e A.C.	2011-2012
21	1			MASCULIN	PHYSIQUE-CHIMIE	16.0	1e A.C.	2011-2012
22	1			MASCULIN	S.V.T	13.0	1e A.C.	2011-2012

Fig. 3. Dataset before.

To eliminate redundant information, we calculated the average grade for each student, subject, and session. Instead of multiple rows representing different assessments, each student is now represented by a single row per session, with columns corresponding to subjects and containing the computed average grades.

After the transformation, the table became as follows (Figure 4). This restructuring enhances data

```
dta.head(50)
```

	ELEVE	SESSION	NOMNIVEAU	ANGLAIS	FRANÇAIS	HISTOIRE GEOGRAPHIE	MATHEMATIQUES	PHYSIQUE-CHIMIE	SVT
0	17	2013-2014	TCS	5.50	12.50	18.50	16.75	15.00	14.75
1	17	2014-2015	1e Sc. Exp.	15.50	12.00	8.62	17.50	11.00	12.50
2	17	2015-2016	2e Sc. Exp.	14.00	15.00	10.00	10.25	18.00	13.25
3	110	2011-2012	TCS	12.50	7.00	15.83	8.20	12.00	8.00
4	110	2012-2013	1e Sc. Exp.	14.90	13.17	5.40	11.12	9.05	14.10
5	110	2013-2014	2e Sc. Exp.	13.25	16.00	10.00	15.00	8.12	9.62
6	160	2011-2012	1e Sc. Exp.	14.75	4.60	7.14	7.00	7.50	5.20
7	160	2012-2013	1e Sc. Exp.	16.50	13.33	6.05	10.25	7.85	12.70
8	160	2013-2014	2e Sc. Exp.	15.25	17.00	10.00	7.25	14.00	9.19
9	192	2011-2012	TCS	10.00	8.40	11.67	2.17	9.33	6.67
10	192	2012-2013	1e Sc. Exp.	13.80	13.50	4.60	8.12	7.45	13.00
11	192	2013-2014	2e Sc. Exp.	10.00	17.00	10.00	9.75	6.50	4.38
12	205	2011-2012	TCS	13.25	7.80	16.00	9.00	12.00	9.00
13	205	2012-2013	1e Sc. Exp.	15.30	12.00	9.65	10.38	7.70	12.70
14	205	2013-2014	2e Sc. Exp.	15.25	14.00	10.00	6.50	12.00	14.12
15	206	2011-2012	TCS	13.50	10.80	14.50	9.33	12.33	9.00
16	206	2012-2013	1e Sc. Exp.	16.10	13.08	11.60	9.94	8.25	13.40

Fig. 4. Dataset after.

readability and usability, enabling more coherent analyses and facilitating the training of predictive models.

3.2.5. Data augmentation

When working with real-world data, its availability and quality often pose a major challenge in machine learning. In our case, the limited amount of usable data, due to incomplete or missing information, risked affecting the reliability of our predictive model. To address this constraint, we implemented a data augmentation strategy, which involves generating synthetic data to enrich our initial dataset while preserving its structure and characteristics.

The main objective was to create artificial students with an academic trajectory similar to that of real students, including three consecutive sessions. This approach enhances the model's generalization ability by diversifying the student profiles and covering a broader range of academic performances.

Our data augmentation process is based on the following steps:

— **Identification of the session structure:** We analyzed the sequence of existing academic sessions to ensure consistency in the integration of new students.
— **Generation of synthetic students:** We created new student IDs and incorporated them into the database.
— **Assignment of sessions:** Each synthetic student was assigned three consecutive sessions, following the model of real students, ensuring a coherent academic progression.
— **Definition of academic levels:** To maximize profile diversity, we assigned synthetic students to different academic levels, including TCS, 1e Eco. Gest, 1e Sc. Math., 1e Sc. Exp., 2e Sc. Exp., 2e Sc. Eco, 2e SGC, and 2e Sc. Math.

Thanks to this approach, we not only compensated for the lack of real data but also improved the representativeness of student profiles, allowing our model to better capture educational dynamics.

4. Results

Model construction is a fundamental step in the data mining process. While data preprocessing and understanding play a key role in the final model's performance, selecting the appropriate model is a decisive step. It involves experimenting with multiple models and choosing the one that provides the highest accuracy for the specific machine learning task. This rigorous selection ensures that the model can generalize effectively to new data and produce reliable predictions.

In our study, we selected four algorithms based on a literature review: Linear Regression, Random Forest, SVM (SVR), and Decision Tree. These models were chosen for their effectiveness in predicting academic performance and their adaptability to our dataset.

One of the major challenges of this study is accurately predicting students' success rates based solely on their grades. This task is particularly complex because academic success is influenced by numerous factors, such as teaching methods, student engagement, and individual differences in learning. Nevertheless, by leveraging the available data, we aim to extract meaningful patterns that can help anticipate future performance.

To ensure the reliability of predictions and prevent overfitting, we adopted a rigorous evaluation approach based on 5-fold cross-validation. This technique provides more robust estimates by testing the model on different data partitions, ensuring better generalization capability.

Finally, we compared the performance of the models using cross-validation to identify the one that offers the best predictions for our problem. The Table 1 presents the results obtained.

Table 1. Model performance comparison.

Algorithm	Performance (%)
Linear Regression	72.18
Random Forest	75.20
Support Vector Machine (SVM) / Support Vector Regression (SVR)	60.46
Decision Tree	73.93

5. Conclusion

The primary objective of this study was to design a predictive model capable of anticipating student success rates in different academic tracks, relying exclusively on students' academic performance from the previous two years. Following an in-depth literature review, we selected and tested several machine learning models to identify the one offering the best performance.

A rigorous evaluation, based on cross-validation, allowed us to compare these models and obtain significant results. The Random Forest algorithm proved to be the most effective, achieving an accuracy rate of approximately 75.20%. However, this performance remained limited due to data constraints. Although we were able to utilize information from students who had completed a full academic path – from the common core to the second year of high school, including all three academic sessions – the overall dataset was still insufficient to ensure fully reliable modeling.

To address this constraint, we applied a data augmentation technique, which enhanced the model's robustness and improved prediction accuracy. Nevertheless, while this approach partially mitigated the impact of the limited size of the dataset, it did not entirely overcome the data shortcomings.

To improve this study and mitigate these limitations, several research avenues can be considered. Expanding the student sample and ensuring a more comprehensive coverage of their academic trajectories would be necessary. Additionally, accounting for educational and socio-economic disparities across different regions could help assess the model's generalizability. Incorporating other explanatory variables, such as student motivation or socio-economic background, could also enhance prediction accuracy. Finally, exploring ensemble or hybrid models that help reduce variance and improve generalization, would be a promising direction to strengthen the robustness of the results.

Ultimately, this study highlights the importance of having a sufficient and high-quality dataset for developing reliable predictive models. While the results obtained are promising, they underscore the challenges associated with data limitations and open the door to future research aimed at refining predictions and optimizing student guidance.

[1] Apple M. W. Can education change society? Routledge, New York (2012).

[2] Van Der Aalst W. Data Science in Action. Process Mining. Springer, Berlin, Heidelberg (2016).

[3] Osmanbegovic E., Suljic M. Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business. **10** (1), 3–12 (2012).

[4] Burgos C., Campanario M. L., de la Peña D., Lara J. A., Lizcano D., Martínez M. A. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering. **66**, 541–556 (2018).

[5] Lakkaraju H., Aguiar E., Shan C., Miller D., Bhanpuri N., Ghani R., Addison K. L. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1909–1918 (2015).

[6] Ahmed A. B. E. D., Elaraby I. S. Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology. **2** (2), 43–47 (2014).

[7] Marbouti F., Diefes-Dux H. A., Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education. **103**, 1–15 (2016).

[8] Al-Shehri H., Al-Qarni A.,Al-Saati L., Batoaq A., Badukhen H., Alrashed S. Student performance prediction using Support Vector Machine and K-Nearest Neighbor. 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE). 1–4 (2017).

[9] Chui K. T., Fung D. C. L., Lytras M. D., Lam T. M. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior. **107**, 105584 (2020).

[10] Nieto Y., García-Díaz V., Montenegro C., Crespo R. G. Supporting academic decision making at higher educational institutions using machine learning-based algorithms. Soft Computing. **23**, 4145–4153 (2019).

[11] Aggarwal D., Mittal S., Bali V. Significance of non-academic parameters for predicting student performance using en-semble learning techniques. International Journal of System Dynamics Applications (IJSDA). **10** (3), 38–49 (2021).

[12] Ghorbani R., Ghousi R. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access. **8**, 67899–67911 (2020).

[13] Hussain K., Talpur N., Aftab M. U., Zakria. A Novel Metaheuristic Approach to Optimization of Neuro-Fuzzy System for Students' Performance Prediction. Journal of Soft Computing and Data Mining. **1** (1), 1–9 (2020).

[14] Wakelam E., Jefferies A., Davey N., Sun Y. The potential for student performance prediction in small cohorts with minimal availa-ble attributes. British Journal of Educational Technology. **51** (2), 347–370 (2020).

[15] Plagge M. Using artificial neural networks to predict first-year traditional students second year re-tention rates. ACMSE '13: Proceedings of the 51st annual ACM Southeast Conference. 17, 1–5 (2013).

[16] Pal S. Mining educational data to reduce dropout rates of engineering students. International Journal of Information Engineering and Electronic Business. **4** (2), 1–7 (2012).

[17] Hu Y.-H., Lo C.-L., Shih S.-P. Developing early warning systems to predict students' online learning performance. Computers in Human Behavior. **36**, 469–478 (2014).

[18] Yukselturk E., Ozekes S., Türel Y. K. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. European Journal of Open, Distance and e-Learning. **17** (1), 118–133 (2014).

[19] Tan M., Shao P. Prediction of student dropout in e-Learning program through the use of machine learn-ing method. International Journal of Emerging Technologies in Learning. **10** (1), 11–17 (2015).

[20] Bolisetti S., Sankepally P. R., Kunta S. R. R., Lakkireddy R. R., Vemula M. K. Student Performance Analyser Using Supervised Learning Algorithms. EasyChair Preprint №5747 (2021).

[21] Ihya R., Aitdaoud M., Namir A., Guerss F. Z., Haddani H. Using Machine Learning Algorithms to Predict the E-orientation Systems Acceptancy. Innovations in Smart Cities Applications Edition 3. 117–130 (2020).

[22] Ait Daoud M., Namir A., Talbi M. FSLSM-Based Analysis of Student Performance Information in a Blended Learning Course Using Moodle LMS. Open Information Science. **8** (1), 20220163 (2024).

[23] Ramya P., Balakrishnan S., Kannan M. Recommendation system to improve students performance using machine learning. IOP Conference Series: Materials Science and Engineering. **872**, 012038 (2020).

[24] Villar A., de Andrade C. R. V. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence. **4** (1), 2 (2024).

[25] Shawareb N., Ewais A., Dalipi F. Utilizing Data Mining Techniques to Predict Students Performance using Data Log from MOODLE. 2564–2588 (2024).

[26] Krishna C. V., Ashokkumar S. Prediction of academic performance of students using novel classification technique of decision tree comparing with random forest. AIP Conference Proceedings. **2871** (1), 020008 (2024).

[27] Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments. **9** (1), 11 (2022).

[28] García S., Ramírez-Gallego S., Luengo J., Benítez J. M., Herrera F. Big data preprocessing: methods and prospects. Big Data Analytics. **1** (1), 1–22 (2016).

Прогнозування успішності учнів у середній освіті в Марокко: структура машинного навчання для керівництва академічним шляхом

Самма С.[1], Айт Дауд М.[1,2], Ахтаіч К.[1], Трага А.[1]

[1]*Лабораторія LTIM, кафедра інформатики, факультет наук Бен М'сіка, Університет Хассана II Касабланки, Марокко*
[2]*Лабораторія ORDIPU, Факультет наук Бен М'сіка, Університет Хассана II Касабланки, Марокко*

Це дослідження спрямоване на вирішення проблеми відсутності регіональних інструментів для академічного консультування в Марокко шляхом пропозиції фреймворку машинного навчання для прогнозування успішності учнів у різних напрямках середньої освіти. Використовуючи академічні дані учнів із регіону Великої Касабланки, ми оцінюємо чотири моделі (Random Forest, SVM, Decision Trees, Linear Regression), дотримуючись методології, яка включає попередню обробку даних, вибір ознак і збагачення синтетичними даними для вирішення проблеми дисбалансу класів. Алгоритм Random Forest показав результативність на рівні 75.20%, значно перевершивши інші моделі. Пов'язуючи прогнози з практичними рекомендаціями щодо навчання цей фреймворк надає педагогам можливість рекомендувати індивідуальні навчальні траєкторії, враховуючи сильні сторони кожного учня, що усуває критичний пробіл в системі освіти Марокко.

Ключові слова: *видобуток даних; успішність студентів; прогностичне моделювання.*