Predicting student status using machine learning by analyzing classroom behaviors with X-API data

Abdelamine Elouafi[1], Ilyas Tammouch[1], Souad Eddarouich[2], Raja Touahni[1]
[1]Faculty of Science/Telecommunications Systems, and Decision Engineering Laboratory, Ibn Tofail University, Kenitra, Morocco
[2]Regional Educational Center, Rabat, Morocco

Article Info

Article history:

Received Jul 11, 2024
Revised Sep 30, 2024
Accepted Oct 7, 2024

Keywords:

Decision tree
K-nearest neighbors
Logistics regression
Random forest
SVM
The performance of students
XGBoost

ABSTRACT

We explore the emergence and growing significance of educational data mining, a field dedicated to extracting valuable insights from vast datasets gathered from diverse educational environments. Utilizing the experience API (XAPI) and the Kalboard 360 online learning platform, our research presents a novel behaviorally based student performance model that evaluates the influence of student interactions on academic results. We create reliable models for precisely projecting academic success by utilizing machine learning techniques including logistic regression, k-nearest neighbors (KNNs), support vector machines (SVM), decision trees, random forests (RF), and XGBoost. The outcomes show a notable increase in categorization accuracy. Through the personalization of instruction, formative assessment support, and proactive identification of each student's unique needs to maximize their learning experience, this approach holds the potential to improve educational processes.

Corresponding Author:

Abdelamine Elouafi
Faculty of Science/Telecommunications Systems, and Decision Engineering Laboratory
Ibn Tofail University
Kenita, Morocco
Email: abdelamine.elouafi@uit.ac.ma

1. INTRODUCTION

Today, the volume of educational data is rapidly expanding, marking the emergence of a new field: educational data mining. This field focuses on developing methods to address educational challenges by uncovering hidden insights from data collected across various educational environments [1]. L'existing literature has mainly aimed at predicting academic performance by exploring the impact of students ‘external environment on their academic achievement studies have notably used institutional bases and international assessments, such as TIMSS, PISA and PIRLS, to identify the key factors influencing this performance [2], [3]. However, our contribution stands out by focusing on the impact of behavioral traits on student performance [4], by integrating data collected through the Kalboard 360 platform and applying advanced data mining methods, our study aims to comprehensively how these specific behaviors directly influence academic success.

The following sections of this article will demonstrate how our innovative methodological approach fills these gaps by providing valuable insights to improve educational strategies and guide future research in this crucial area [5], we apply six machine learning algorithms: decision tree, random forests (RF), k-nearest neighbors (KNNs) and support vector machines (SVM) to build a robust academic performance model [6], [7]. The goal of this study is to promote the continuous improvement of teaching methods, particularly by helping teachers in the diagnostic and summative assessment phases to analyze student behavior in the classroom.

Much study has been conducted on forecasting student performance and behavior in the classroom, includes a variety of innovative ways and tools for achieving goals, gathering information, making decisions, and making recommendations. Some of the information used as a source for this article is included below. The authors concluded that the school administration and atmosphere have an impact on students' academic performance [8], [9] On the other hand, The authors found that the teacher is primarily responsible for students' performance [10].

The authors [1], [11] presented a case study that examined students' learning preferences through educational information mining. This research aimed to demonstrate how data mining methods might enhance students' academic performance in postsecondary education. The course database, which contains students' academic and personal records, is where the data set was gathered. In reference [12], [13] the authors use the expectation maximization algorithm (EM-clustering) to assign pupils to five groups according to how well they performed. The algorithm is based on maximum likelihood parameter estimates in probabilistic models.

In other work Shannaq *et al*. [14], [15] employed a categorization method to determine the total number of enrolled students by examining the essential traits that can affect the students' loyalty. Following their extraction of 2069 sample records from the student database, the authors applied a decision tree technique to build a classification model and pinpoint the key attributes that could affect students. This research enables the university administration to provide the necessary materials for newly registered students in higher education institutions.

In Yaacob *et al*. [16], [17] employed the k-means clustering algorithm to use a database to forecast students' learning activities, such as tests and quizzes in class. The instructor of the class will receive the gathered data prior to the last exam. With timely interventions, this study helps teachers lower failing rates and raise student accomplishment. In conclusion, numerous studies have investigated employing data mining approaches to solve educational challenges. nonetheless, a dearth of studies has illuminated how students behave during the learning process and how this affects their academic achievement. The impact of students' interactions with the e-learning system will be the main topic of this study. Additionally, schools will benefit from the extracted knowledge by improving students' academic performance. additionally, to support administrators in enhancing learning systems.

The paper is organized as follows: the section's initial introduction: An outline of the research issue, including the study's history and purpose, is provided in this part. It also includes an overview of the paper's primary research topics and goals. The research strategy and methodology of the study are covered in section 2: method, along with details on the data gathering and analysis processes. Section 3: results and discussion: this section includes important facts and statistics as well as a discussion of the study's conclusions. Along with highlighting any patterns or trends found, it also discusses the significance of the findings for the field and suggests topics for further research.

2. METHOD

2.1. Proposed model

The suggested approach to assess student performance is collected via the experience api (XAPI) from the Kalboard 360 e-learning platform. The raw data is transformed into a CSV file to make additional analysis easier. The data are then preprocessed to guarantee their dependability and integrity. Data cleaning, duplicate entry removal, missing value management, variable standardization, handling outliers, normalization, and data transformation are some of the methods used to get the data ready for further analysis. Feature selection finds a subset of relevant features in the preprocessed data after it has been preprocessed. Multiple classifier techniques, such as logistic regression, KNN, SVM, decision trees, RF, and XGBoost, then use this subset as input.

These classifiers produce predictions and categorize new occurrences by using the selected characteristics to find patterns and relationships in the data. The preprocessed data is split into two subsets: the training set and the testing set to assess the effectiveness and capacity for generalization of the trained models. While the test set is intended to evaluate data that has yet to be seen, the training set is used to train algorithms for classification. After evaluating the results, we succeeded in predicting the students' status, which helps teachers make decisions. We adhered to the procedures depicted in Figure 1.

2.2. Data collected

This article employs data obtained from the Kalboard 360 E-Learning system using experience API (XAPI) [18], [19]. The training and learning architecture (TLA)'s XAPI component keeps account of learner

experiences and actions, such reading articles or watching training videos. With the use of the experience API, learning activity providers can specify the student, the activity, and the items that comprise a learning experience [20], [21]. In this study, variables that may have an impact on academic success are evaluated and student behavior is tracked throughout the learning process using X-API. There are 480 student records with 17 attributes in the data collected set. Three main categories comprise the features presented in the Table 1:
– Demographic characteristics like gender, place of birth, and nationality.
– Academic background includes stage, grade, semester, and section.
– Behavioral aspects include raising hands in class, accessing resources, participating in conversations, and reviewing messages and announcements.

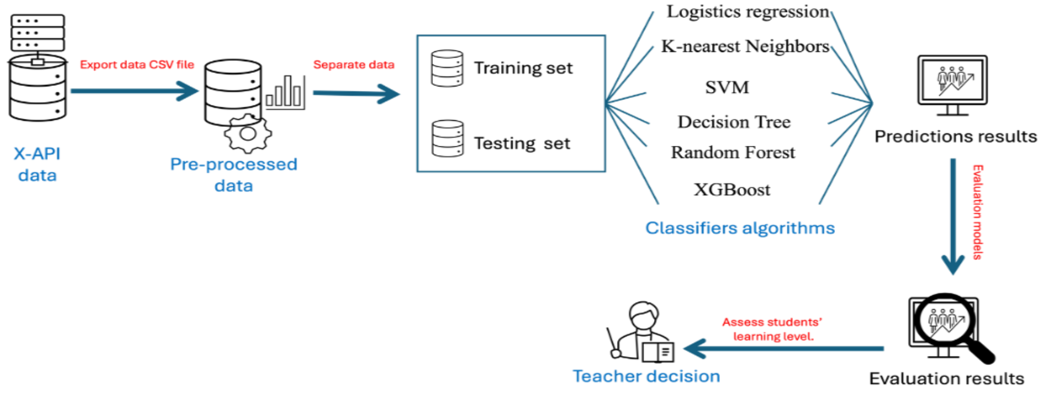

Figure 1. The workflow processes

Table 1. Description of variables

Attribute category	Attribute	Description
Demographical attribute	Gender Relation Nationality Place of birth	The student's gender (male or female). Contact parent of the student, such as father or mother. Student nationality. Place of birth for the student.
Academic background attribute	Stage ID	Students are classified into stages, such as low level, middle level, and high level.
	Section ID Semester Grade ID Topic Parents answering	Sections to which students belong include (A, B, and C). Semesters of the school year such as (First or Second). The student's grade corresponds to (G-01, G-02). Topics covered in class include science, math, English, and Arabic… Parents Answering Survey (Yes or No)
	Parent school gratification	Parent satisfaction at school (Good or Bad)
	Student absence days	The number of days absent (Under 7 or Above 7)
Behavioral attribute	Raised hand Visited course Discussion group Announcements view	How students react when using the Kalboard 360 e-learning system.
Features	Class performance	Three categories are used to classify students: lower level, middle level, and high level.

2.3. Preparation data

We employ specific preprocessing techniques to enhance the data quality after completing the data collection task [7], [22]. Data preparation, which includes data transformation, data reduction, data purification, and feature selection, is thought to be an important phase in the knowledge discovery process [23]. Data cleaning is performed on this data set to eliminate noise and missing values. Following cleaning, the data set now contains 480 records. The dataset contains 305 males and 175 females. Stage ID has 199 lower levels, 248 middle levels, and 99 high levels. Furthermore, the students are divided into three sections: 283 students from section A, 167 students from section B, and 30 students from section C. Topic attributes include: 95 students are associated to the IT topic, 21 to Math, 45 to English, 30 to Biology, 24 to Chemistry, 24 to Geology, 25 to Spanish, 22 to Quran, 51 to science, 19 to History, and 59 to Arabic topic. Relation attribute includes 283 students, their contact person is the father and 197 students, the contact person is their mother.

2.4. The objective of paper

The primary goal of this study is to analyze students' classroom behavior to identify the characteristics that have an important impact on their academic achievement. Next, we aim to identify the model that most accurately represents student performance using powerful machine learning algorithms. Our goal is to boost prediction accuracy. and this model will optimize intervention mechanisms to provide students with individualized help adapted to their individual needs. This helps professors in the evaluation of students in the classroom.

3. RESULTS AND DISCUSSION

This section gives an extensive overview of our experimental design, including classroom behavior analysis and variable identification that affects student status. We describe in detail the precise settings and instruments employed to carry out an exhaustive assessment of our suggested model. We then go over the model's performance evaluation's findings. In conclusion, we go into detail about these findings and emphasize how well the model works to forecast student achievement based on behavior in the classroom.

3.1. Correlation

An important stage in our research was to investigate the correlations [24] between various features using heat map methodologies. The Figure 2 visually depicts these correlations, focusing on the linkages between various traits and student status. A heat map highlights potential patterns or linkages that could have an impact on student outcomes while offering a thorough and understandable picture of how various elements interact. With these characteristics, it is possible to obtain information that has a direct impact on student performance, which has been the objective of several previous studies. However, these studies have mainly focused on characteristics influencing student performance in terms of knowledge. In contrast, our research emphasizes the impact of classroom behavior on student performance.

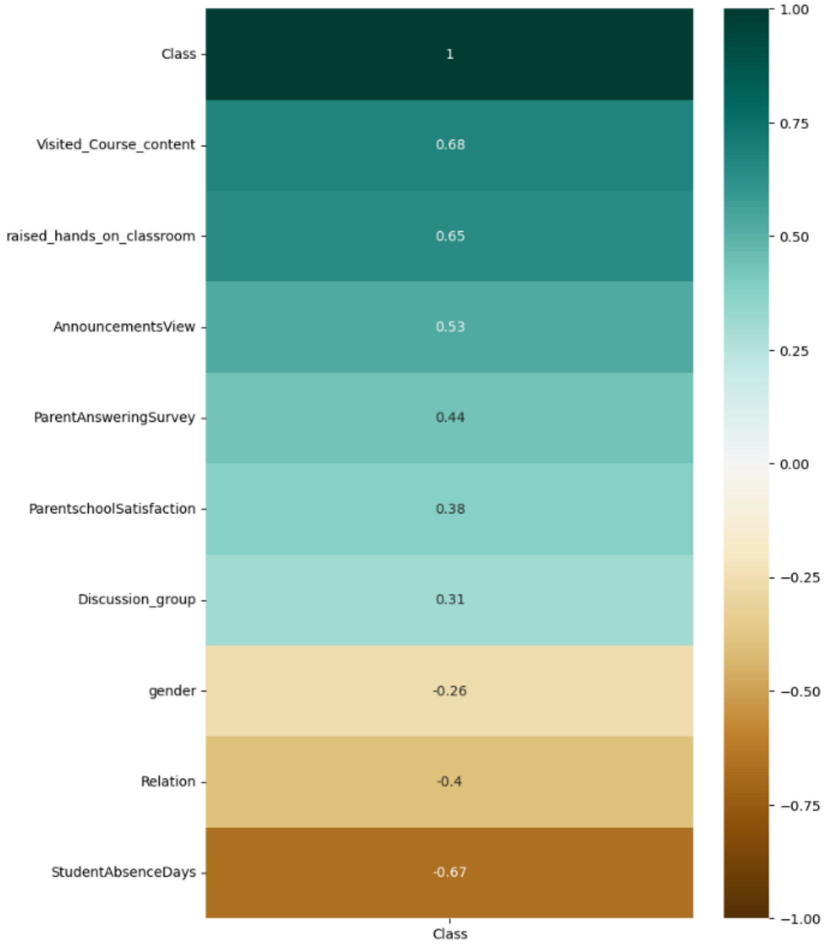

Figure 2. Qualities associated with the student's status

With this heatmap, we can quickly determine which features have the most significant influence on student status:

- Four most impactful features (positively):
  a. Students' permanent visit to course content has a positive impact on their status.
  b. Raising your hand in class to participate has a positive impact on students' status.
  c. The number of interactions students have with posted announcements has a favorable impact on their status.
  d. Group discussion appears to influence student learning.
- Two most impactful features (negatively):
  a. Student absences have an inverse impact on their learning.
  b. The relationship between parents and students also influences the status of students.

In this work, we will focus on classroom behavior, considering the following characteristics: raised hand, visited course, discussion groups, and announcement view. These associations are represented visually in Figure 3, which also shows the distribution of student status and the relationships between each pair of attributes.

To analyze the diagram representing the density of student status according to the variables raised hand, visited course, announcement view, and discussion group, several key observations emerge. Figure 4 graphically represents these distributions. First, it is notable that students with a low level of participation, especially in the "low level" group, are strongly concentrated in the lower ranges of the variables. For example, many observations show that for "raised hand" and "visited course," these students are mainly located between 0 and 20 on a scale of 0 to 100. On the other hand, students with a high level of participation ("high level") cluster mainly towards the upper ends of the variables, such as 80 to 100 for "raised hand" and 75 to 94 for "visited course". for "announcement view", although most students show a low level of viewing announcements (0–20), those with intermediate and high levels more evenly distribute their observations over a wider range of 20 to 90. Finally, for the "discussion group," a significant part of the observations is concentrated around 0 to 20, with a more spread distribution for intermediate and high levels between 20 and 90. These observations suggest a potential correlation between the level of participation of students and their statutes, pointing toward further exploration of the impacts of active engagement on academic outcomes and student well-being.

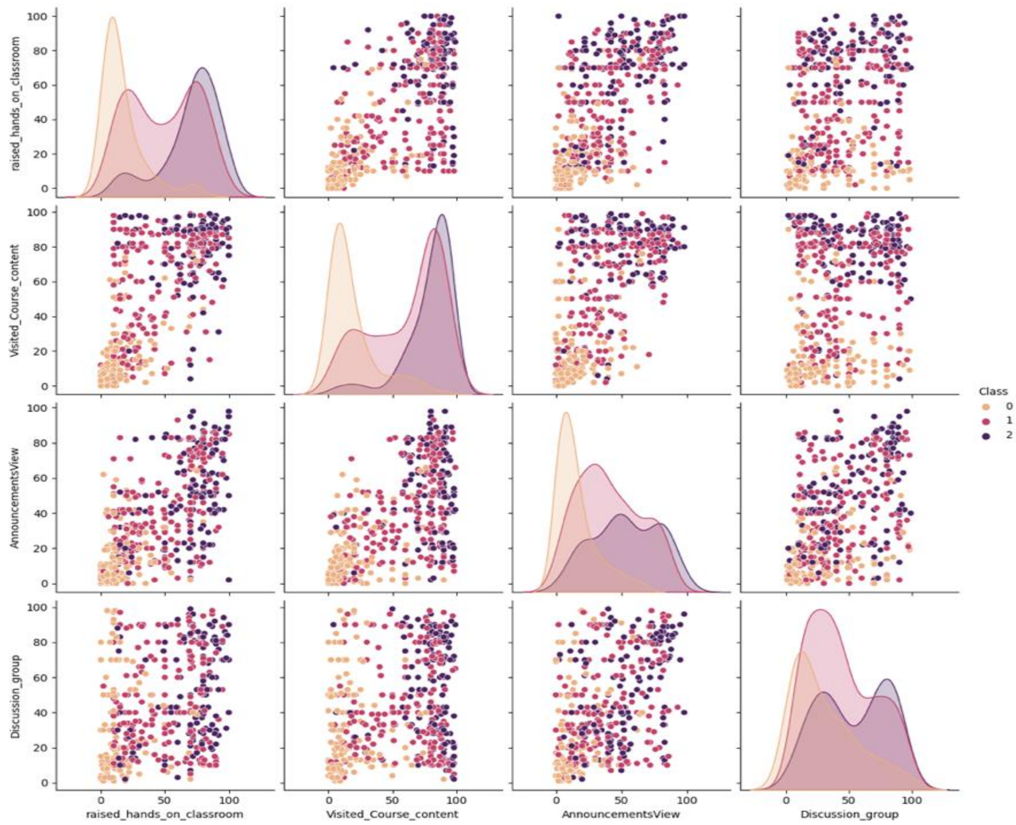

Figure 3. The relationship between feature pairs and the distribution of student status

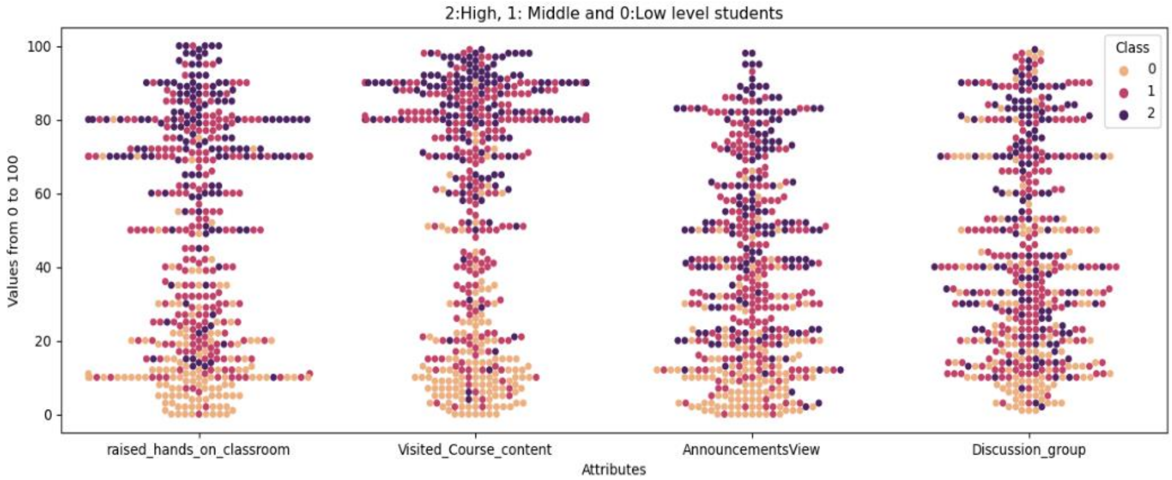

Figure 4. Distribution of students according to class behavior

3.2. Model construction

During the model building stage, we created predicted models of academic achievement using six different machine learning techniques. Among these techniques are SVM, KNNs, RF, and decision trees. While decision trees are useful for capturing nonlinear relationships in data, RF combine several decision trees to manage overfitting and increase prediction accuracy. KNNs, appropriate for localized patterns in the data, uses similarity metrics to predict outcomes based on the closest training samples. To maximize the margin between classes for reliable predictions, SVM categorize data points into distinct classes using kernel approaches. To maximize model parameters and guarantee generalizability across a variety of datasets, cross-validation techniques were employed throughout the training and fine-tuning of each algorithm. This comprehensive approach allowed us to explore various facets of student performance prediction, leveraging the strengths of each algorithm to uncover meaningful insights from educational data [25], [26].

3.3. Comparison of the results of classification algorithms

After applying classification techniques to the dataset, our study demonstrates that the results are distinct depending on the classification algorithms used. Table 2 presents the results using different classification algorithms (Logistic regression, KNNs, Support vector machine, decision tree, RF, and XGBoost). There are notable differences in the machine learning models' capacity to classify at three different levels of categorization, as shown by the comparative table of performance measures. With an overall accuracy of 84% and high F1 scores, notably 0.95 for the high-level category, logistic regression stands out. KNNs, on the other hand, produce fewer good findings, with an accuracy of 69% and generally lower F1 scores. With balanced recall and F1 scores, the support vector machine achieves 72% accuracy, which is a little less than the logistics regression. Although Decision-Tree's overall accuracy is just 60%, it performs exceptionally well at a high level, with a high F1-score of 0.84. Both RF and XGBoost demonstrate strong results; RF attains 83% accuracy, while XGBoost stands out for having a flawless F1-score of 1.00 at the intermediate level. In conclusion, the choice of model depends on specific priorities regarding accuracy and the ability to handle different levels of classification, with logistic regression and XGBoost recommended for their overall strong performance.

Table 2. best hyper-parameters after a comprehensive grid search.

Method	Accuracy	Recall			F1-score		
		Lower level	Middle level	High level	Lower level	Middle level	High level
Logistic regression	0.84	0.89	0.95	0.74	0.86	0.89	0.79
KNNs	0.69	0.60	0.77	0.72	0.69	0.74	0.66
Support vector machine	0.72	0.74	0.86	0.62	0.76	0.76	0.65
Decision-Tree	0.60	0.49	0.86	0.56	0.52	0.84	0.54
RF	0.83	0.93	0.80	0.78	0.86	0.85	0.80
XGBoost	0.84	0.71	1.00	0.85	0.81	0.92	0.80

4. CONCLUSION AND PERSPECTIVES

Our findings provide conclusive evidence that classroom behavior has a powerful impact on students' performance, comparable to other socio-economic factors. These results suggest that teachers can leverage predictive models such as XGBoost and logistic regression to evaluate and monitor student behavior. Key behavioral indicators include active participation (raised hands), course visits, involvement in discussion groups, and engagement with announcements. Raising hands reflects a student's willingness to participate or ask questions, while course visits indicate their use of online learning resources. Engagement with announcements and participation in discussion groups show active involvement in the course content. These four traits help teachers better understand student behavior in the classroom and enable them to adapt and deal with challenges that students may encounter. Teachers can use data on these four features, gathered from each class, to predict student achievement and develop remediation or reinforcement plans. Integrating these findings with other research on improving learner performance will enable us not only to address teaching-related challenges but also to evaluate and enhance educational systems.

REFERENCES

[1] A. El-Halees, "Mining students data to analyze learning behavior: a case study." 2009. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8cc1bbbafc6cf1fe90b3b5dfb65d4dbd0d3e11da

[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[3] M. M. Abuteir, A. M. El-Halees, M. M. A. Tair, and A. M. El-Halees, "Mining educational data to improve students' performance: a case study educational data mining view project opinion mining view project mining educational data to improve students' performance: a case study," vol. 2, no. 2, 2012, [Online]. Available: http://www.esjournals.org

[4] A. Ahmadi, A. Ziapour, J. Y. lebni, and N. Mehedi, "Prediction of academic motivation based on variables of personality traits, academic self-efficacy, academic alienation and social support in paramedical students," *Community Health Equity Research and Policy*, vol. 43, no. 2, pp. 195–201, Jan. 2023, doi: 10.1177/0272684X211004948.

[5] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, Sep. 2021, doi: 10.3390/educsci11090552.

[6] S. P. Das and S. Padhy, "A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 1, pp. 97–111, Jan. 2018, doi: 10.1007/S13042-015-0359-0.

[7] A. Elouafi, I. Tammouch, S. Eddarouich, and R. Touahni, "Evaluating various machine learning methods for predicting students' math performance in the 2019 TIMSS," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 565–574, Apr. 2024, doi: 10.11591/ijeecs.v34.i1.pp565-574.

[8] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[9] A. Harris, *Teaching and learning in the effective school*. Routledge, 2019. doi: 10.4324/9780429398117.

[10] M. C. F. Raguro, A. C. Lagman, L. P. Abad, and P. L. S. Ong, "Extraction of LMS student engagement and behavioral patterns in online education using decision tree and k-means algorithm," in *ACM International Conference Proceeding Series*, New York, NY, USA: ACM, Jan. 2022, pp. 138–143. doi: 10.1145/3512353.3512373.

[11] B. Kumar and S. Pal, "Mining educational data to analyze students performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, 2011, doi: 10.14569/ijacsa.2011.020609.

[12] M. F. Mustapha, A. N. I. Zulkifli, O. Kairan, N. N. S. M. Zizi, N. N. Yahya, and N. M. Mohamad, "The prediction of student's academic performance using RapidMiner," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, pp. 363–371, Oct. 2023, doi: 10.11591/ijeecs.v32.i1.pp363-371.

[13] A. F. N. Alrammahi and K. B. S. Aljanabi, "A new approach for improving clustering algorithms performance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 3, pp. 1569–1575, Dec. 2020, doi: 10.11591/ijeecs.v20.i3.pp1569-1575.

[14] B. Shannaq, Y. Rafael, and V. Alexandro, "Student relationship in higher education using data mining techniques," *Global Journal of Computer Science and Technology*, vol. 10, no. 11, pp. 54–59, 2010.

[15] A. Polyzou and G. Karypis, "Grade prediction with models specific to students and courses," *International Journal of Data Science and Analytics*, vol. 2, no. 3–4, pp. 159–171, Dec. 2016, doi: 10.1007/s41060-016-0024-z.

[16] T. Ilyas, A. Elouafi, and E. Souad, "Centroid competitive learning approach for clustering and mapping the social vulnerability in Morocco," *International Journal of Advanced And Applied Sciences*, vol. 9, no. 9, pp. 70–77, Sep. 2022, doi: 10.21833/ijaas.2022.09.009.

[17] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1584–1592, Dec. 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

[18] A. K. Hamoud *et al.*, "A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 2, pp. 1105–1116, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1105-1116.

[19] A. E. W. Widarta, A. Luthfi, and C. Kusuma Dewa, "Prediction of student performance based on behavior using e-learning during the Covid-19 pandemic using support vector machine," *Sinkron*, vol. 9, no. 1, pp. 332–345, Jan. 2024, doi: 10.33395/sinkron.v9i1.12857.

[20] J. Armani, "Shaping learning adaptive technologies for teachers: a proposal for an adaptive learning management system," in *Proceedings - IEEE International Conference on Advanced Learning Technologies, ICALT 2004*, IEEE, 2004, pp. 783–785. doi: 10.1109/ICALT.2004.1357656.

[21] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Education and Information Technologies*, vol. 26, no. 1, pp. 205–240, Jan. 2021, doi: 10.1007/s10639-020-10230-3.

[22] N. Agnihotri and S. K. Prasad, “Hybrid logistic regression support vector model to enhance prediction of bipolar disorder,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, pp. 1294–1300, Nov. 2024, doi: 10.11591/ijeecs.v36.i2.pp1294-1300.

[23] Y. D. Lan, “A hybrid feature selection based on mutual information and genetic algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 214–225, Jul. 2017, doi: 10.11591/ijeecs.v7.i1.pp214-225.

[24] C. Reimann, P. Filzmoser, K. Hron, P. Kynčlová, and R. G. Garrett, “A new method for correlation analysis of compositional (environmental) data – a worked example,” *Science of the Total Environment*, vol. 607–608, pp. 965–971, Dec. 2017, doi: 10.1016/j.scitotenv.2017.06.063.

[25] M. Jebbari, B. Cherradi, S. Hamida, M. A. Ouassil, T. El Harrouti, and A. Raihani, “Enhancing learner performance prediction on online platforms using machine learning algorithms,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 1, pp. 343–353, Jul. 2024, doi: 10.11591/ijeecs.v35.i1.pp343-353.

[26] B. Mahesh, “Machine learning algorithms - a review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/art20203995.

BIOGRAPHIES OF AUTHORS

Abdelamine Elouafi is currently a doctoral candidate at Ibn Tofail University, located in Kenitra, Morocco. Concurrently, he serves as a secondary school teacher at the Ministry of Education in Morocco. He earned his master's degree in decision-Making Informatics in 2019 from Sultan Moulay Slimane University in Beni Mellal, Morocco. Additionally, in 2016, he obtained a professional teaching license from the Higher Normal School of Fes, showcasing his commitment to education. In 2014, he achieved a license degree in Computer Engineering, Electronics, and Automation from Sultan Moulay Slimane University. His scientific interests are diverse, spanning areas such as machine learning, deep learning, data analysis, and system evaluation. For any inquiries, he can be contacted at email: abdelamine.elouafi@uit.ac.ma.

Ilyas Tammouch is currently a doctoral student at Ibn Tofail University, located in Kenitra, Morocco. At the same time, his scientific interests are varied and cover areas such as machine learning, deep learning, data analysis, and systems evaluation. For any inquiries, he can be contacted at email: ilyas.tammouch@uit.ac.ma.

Souad Eddarouich is currently an Associate Professor of Computer Science at the Regional Center for Education and Training Professions in Rabat (CRMEF), he also serves as a researcher in artificial intelligence within the ERMID research team. His expertise spans various areas, including machine learning, deep learning, data analysis, and system evaluation. For any inquiries, she can be contacted at email: eddarouichsouad@gmail.com.

Raja Touahni is currently a Professor in Higher Education at the Faculty of Science, Telecommunications Systems, and Decision Engineering Laboratory, Ibn Tofail University, Kenitra, Morocco. His areas of expertise encompass various subjects such as machine learning, deep learning, data analysis, and system evaluation. For any inquiries, she can be contacted at email: touahni.raja@uit.ac.ma.