IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 October 2023, accepted 25 December 2023, date of publication 28 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348091

RESEARCH ARTICLE

Development of an Early Warning System to Support Educational Planning Process by Identifying At-Risk Students

MUSTAPHA SKITTOU, MOHAMED MERROUCHI, AND TAOUFIQ GADI
Laboratory Mathematics, Computer and Engineering Sciences (MISI), Faculty of Sciences and Technics, Hassan First University of Settat, Settat 26000, Morocco
Corresponding author: Mustapha Skittou (m.skittou@uhp.ac.ma)

ABSTRACT The development of data analysis techniques and intelligent systems has had a considerable impact on education, and has seen the emergence of the field of educational data mining (EDM). The Early Warning System (EWS) has been of great use in predicting at-risk students or analyzing learners' performance. Our project concerns the development of an early warning system that takes into account a number of socio-cultural, structural and educational factors that have a direct impact on a student's decision to drop out of school. We have worked on an original database dedicated to this issue, which reflects our approach of seeking exhaustiveness and precision in the choice of dropout indicators. The model we built performed very well, particularly with the K-Nearest Neighbor (KNN) algorithm, with an accuracy rate of over 99.5% for the training set and over 99.3% for the test set. The results are visualized using a Django application we developed for this purpose, and we show how this can be useful for educational planning.

INDEX TERMS Early warning system, machine learning, KNN, educational planning, dropout.

I. INTRODUCTION

The evolutionary path of IT practice has taken a new form with the advent of intelligent systems, especially predictive and recommendation systems. And with the explosion of data and the entry into the era of Big Data [1], these systems have found more opportunities to flourish and achieve the most remarkable results.

Early warning systems (EWS) are one of the most famous types of intelligent systems, and have benefited from the considerable leap forward in computing methods and technologies used, as well as the development of hardware infrastructures.

The EWS is a predictive system that aims to support decision-making by giving a proactive view of the future situation by analyzed data. EWS are used in almost all fields, and their role lies in detecting anomalies in real systems and warning decision-makers of the seriousness of situations, so that they can anticipate their intervention to remedy the problems posed, or at least limit the negative effects and consequences.

An EWS can be defined by a number of active keywords [2], [3]: Collect, Analyze, Detect, Prevent, Alert, Notify. Each word indicates one of the key stages of an EWS, hence its action model, which consists of a set of layers or steps. The first step involves the continuous monitoring of relevant indicators and the collection of data in real or near-real time. The next step involves the analysis and processing of the data collected. This process involves examining the data to identify early indicators, patterns or deviations from the norm. Various techniques such as advanced algorithms, statistical models or artificial intelligence can be used to identify potential anomalies during this analysis. The third step is 'Alert and Notify', once the system detects an irregularity or potential hazard, it quickly triggers an alert to inform the parties concerned. Then, in the fourth step 'Risk assessment', professionals and supervisors evaluate the reliability and severity of the warning. They examine existing data and information to understand the characteristics of the risk, its potential consequences and possible actions to minimize its

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

impact. The communication and dissemination process plays an essential role in risk management and response. Once a risk has been assessed and verified, it is crucial to share the relevant information with all parties concerned, including stakeholders, decision-makers and the general public. The final layer of the process involves response and action. And finally, once the early warning system provides the necessary information, appropriate measures are taken to reduce risks, prevent crises or minimize adverse effects.

All these steps form an iterative process, as we are always aiming for perfection of the system, given that EWSs are used in highly critical areas and that the effect of their Outputs can avert disasters in some cases, whether in the near future or in the long term. That's why we're constantly striving for perfection.

A. EMIS AND EDM

The Education Management Information System (EMIS) can be defined as a specific information system whose main aim is to support decision-making in the educational field, by first enabling the collection and storage of all types of educational data: Fundamental data, Business data and Statistical data. Then, to analyze these data, each with the appropriate technology. Finally, to produce visual reports or results to provide decision-makers with ideas on all dimensions of the education system.

Educational Data Mining (EDM) is the whole range of technologies applied to education data in order to extract value from them [4]. This includes techniques relating to statistics, data mining and machine learning. EDM [5], [6] is considered the heart of the EMIS, since it is the machine that enables data processing, builds data models and profiles of education stakeholders, and knows all the interconnections and relationships between these data.

The education management information system is undoubtedly a data-driven system, as everything is concentered on data. The analysis of this data through EDM techniques and its outputs enables us to map out future strategies and support decision-making in the field of education.

The EWS is part of the overall EMIS framework, or considered as an integral part of the Learning Management System (LMS), since it serves as a tool for transforming and processing educational data for specific purposes. And by incorporating EDM, these purposes happen to be varied, so EWS changes facet and output. Among the main topics addressed by EDM is the analysis of Student's performance [6] through the analysis of students' progression and grades in the subjects or courses they have taken. Another theme is the study of learner behavior [7], which will be observed either in terms of the learner's experience in his or her school or academic environment, with all its variables acting without context, or in relation to his or her interaction with school content (courses and modules) and the program curriculum. The EDM is also concerned with identifying students at risk of dropping out [4], a phenomenon much studied in education systems because of its negative impact on the future of students and the quality of their learning. Naturally, the study of educational phenomena has led to the development of EWS [8] to prevent their occurrence and provide decision-makers with strategic leeway to address the causes of these phenomena as early as possible, before they get out of hand. But it has also enabled advanced research into decision-support systems [9], given the growing need among players in the educational field for greater visibility of data and the most appropriate recommendations for each individual situation. Last but not least, we'd like to conclude with a more global topic which is the evaluation of education system performance [10]. Although this is a matter for official national or global institutions, EDM techniques can be of considerable help in projecting the results achieved against the objectives already set for education plans. In our work, we have given immense importance to this strategic level of educational planning, something we have not been able to find discussed in the field of EWS development research in education. This is motivated by the idea that to apprehend any educational phenomenon, we need to place it in a more global context, and necessarily resolve it with educational planning.

B. EDUCATIONAL PLANNING

Everyone recognizes that education, by its inference on individuals, is a societal affair of the first degree. And the level of education of the population, as well as the literacy rate recorded, undoubtedly reflects a country's level of development. But achieving any educational goal requires long-term, progressive planning, because the results of education take time to show up in behavior and achievement, hence the need for educational planning.

Educational planning is the logical and systematic process of putting in place an educational plan to achieve society's educational objectives, taking into account all available resources, whether human, financial or material. The right to education is an ultimate right supported by all international texts, and educational planning takes upon its shoulders the task of maintaining this right for all students with a view to equity and equal opportunity, with quality and efficiency.

Dropping out of school is a black mark in the process of implementing education plans, because it represents a flight of pupils from the education system, especially before they have completed the compulsory level of schooling. But it also represents a loss of resources deployed to keep students in school, and the most unpleasant thing is that students leave school before acquiring the skills they need to meet life's challenges. That's why we need educational planning that's flexible and renewed, and that takes advantage of all the new techniques for more rigorous and effective planning.

Our paper is organized as follows. In Section II, we present a literature review of similar projects that have developed EWS or EDM solutions for predicting school dropout. Section III, describes the materials used in our approach to build our EWS. And, in Section IV we explore the different methods we based on to elaborate our proposed solution. Section V then presents the results of our solution implemented to address the problem of students dropping out of

school. And finally in Section VI, the conclusion summarizes our ideas, and we stipulate our prospects for future work.

II. RELATED WORK

The application of EWS in education is not very common, especially in underdeveloped countries. But the need for a predictive system is paramount, given all the considerable advantages it offers.

Our contribution in this paper intersects with several initiatives and projects that have been working on the implementation of EWS in the education system. And EWS is a field that has prompted a great deal of production in recent years. Researchers in the field of education, whose aim is to develop teaching practice and minimize the risk of student attrition, have made a number of efforts on several fronts [11]. The various EWSs have acted on various factors, depending on the logic of analysis, the field of practice or the purpose of the system. These factors could be classified as follows:

- School factors [5], [8], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]: Represent the different variants that affect students' school performance. On the one hand, we start with the pupil's educational background, i.e.: everything to do with where he or she comes from, did he or she receive pre-school education or not? did he or she come from formal, non-formal or original education or something else? On the other hand, his or her learning experience throughout the years of schooling prior to the time of the study (grades and averages for courses and exams, as well as end-of-year averages). These factors represent a very important part of a student's heritage, especially if they are analyzed over time, in a cumulative way, and if possible within the frame of a cohort. In most cases, this time-series study is not possible, mainly due to a lack of data, which leads researchers to make projections on the basis of a school season. But in any case, school factors alone do not allow for an in-depth study of educational phenomena, perhaps only a statistical analysis of school performance.
- Human factors [5], [16], [23], [30], [31], [32]: All the elements that define the learner as a human being, by his gender, age, ethnicity or origin. At this level, we also talk about students' behavioural traits and habits, especially those that affect their learning and reduce their academic performance: absenteeism, procrastination, diligence…Body and mental aspects are also taken into consideration, since body health (disabled or not) and mental health (sick or healthy) can significantly affect a learner's performance. These factors are of major importance in the study of the learner's being and personality. Behavioural data can be collected through empirical studies or only through individual analysis of learners. Nevertheless, the lack of medical data is very much in evidence in the systems of underdeveloped countries, resulting in the marginalization of a very important effect that may explain the difficulty or deterioration of the learner's performance or retention. But added to this, the human being does not live in a personal context alone but in a society and environment, which likewise makes human indicators insufficient in the process of predicting educational phenomena.
- Environmental factors [15], [33]: The effect of the environment is undoubtedly a very significant one on the individual, which is why we speak in the education's sociology of a fortunate school belonging to a resource-rich environment, and an underprivileged school located in an underprivileged environment. The type of school is also categorized at this level, since a private school is often better in terms of academic results than its public counterpart, mainly due to disparities in resources. Another essential dimension of these factors is the family's internal situation, or as it is known in the general population census: household status. The latter highlights indicators of the family's economic capacity and social status, as well as the parents' level of education and the culture that reigns within the family and the aspirations of its members (parents, brothers and sisters). Finally, the school's internal atmosphere is also important in this context: the school's management model, facilities, the state of the teacher/learner relationship, internal peace and non-violence are maintained. These factors highlight the way in which learners are impacted by their school environment, as well as the actors operating in the school. On the other hand, we need to be careful about the way in which the environment is apprehended, as well as the type of variables used, for example, the degree of family impact, in relation to the school or the street. Also, perhaps there are other environmental factors that we couldn't capture but that are very important too, the effect of media and social networks, or that of older students in higher grades on new entrants, and others.
- Exceptional factors [6], [18], [20], [22], [24], [26], [32], [34], [35]: These include all exceptional contingencies and natural crisis situations such as natural disasters, wars and human conflicts that destabilize the educational process. One very exceptional situation we have witnessed over the last 3 years is the health pandemic caused by the Corona virus. This pandemic had an extraordinary effect on the whole learning mechanism, with schools closing their doors all over the world, and other learning methods taking over from face-to-face learning, such as distance learning, where e-learning and hybrid methods flourished. These circumstances usually have a radical and abrupt effect, and can only be studied in exceptional vision. But they are not taken into account in the development of a general model aimed at sustainability.

The important thing is that, in order to assess a learner's performance and make a sustained analysis, it is imperative that the analysis covers all aspects of his or her situation and activity [11]. This is very difficult to achieve most of the time, which is why we are content to study a set of thematic indicators and neglect the rest. This is the key observation for the projects listed before. The vast majority work in the field of higher education, and a large proportion focus on e-learning and distance learning platforms. This leaves marginalized empirical studies of students' diverse learning situations, and especially those concerning the elementary or secondary learning cycles where the dropout rate is highest and represents a dangerous large-scale socio-economic and

literacy indicator. We need to turn seriously to these levels of study in our research projects, since they represent the cognitive and formative basis of students' skills, and to approach them with a very rigorous and exhaustive scientific vision.

Another observation is that all the papers use a variety of methods and the latest technologies to achieve the best results, but the accuracy of the models is still far from precise, perhaps because they are applied to generic data sets and to specific countries or situations far removed from the researchers. In addition to the need to have knowledge of the educational field and educational planning to choose the parameters suitable for the study and a field that is multidisciplinary, involving: demography, economics, statistics, human sciences, psychology, management and others [36].

That's why our effort in this paper to develop an EWS to identify the dropout phenomenon starts with a professional knowledge of the educational planning's field, which means having a panoramic view of the educational process and all its actors and synergies. And from there to be able to control the inputs needed for each analysis to achieve the desired objectives, but also to be able to explain the resulting predictions and postulate the recommendations needed to make the appropriate decisions that go beyond changes in teaching practice, learning methods, upgrading curricula or the learning experience; towards structural, organizational or political actions.

III. MATERIALS

A. INFORMATION COLLECTION

Our study deals with the case of the Moroccan education system, taking it as a case study referring to the model of an education system in developing countries of the South, especially on the African continent.

The data collected empirically gives a snapshot of two different periods of the year: firstly, the data for the start of the school year, which contains all the pupils expected to continue their studies in the current year; then, in the middle of the year, comes the moment of the census, when managers check the accuracy of the data, especially concerning the situation of pupils. The aim of the census reference date is to increase the integrity and accuracy of the data stored in the information system, and their consistency with the situation on the ground. This is very useful for assessing the current situation, as well as for rigorous planning of the future school season.

The second information-gathering moment comes directly after the final results of the end-of-year exams are posted. This allows us to explore the academic achievement of the pupils present on the base at the start of the school year, and to get a normal idea of their academic performance.

B. DATA SET

The database used is built specifically for this project by aggregating data from several operational systems, as illustrated in Fig. 1. Each system provides part of the information useful for understanding and analyzing the dropout phenomenon:

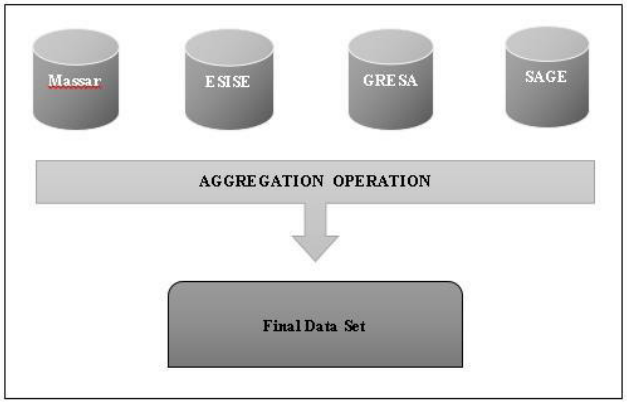

FIGURE 1. Aggregation of data from various operational educational information systems.

- *Massar: Provides information on the student and his/her school data during this school season, and other information on his/her history and background. Not forgetting school results.*
- *ESISE: This is used to extract data on educational delays calculated throughout the student's career and other indicators.*
- *GRESA: This is very useful for finding out which communes benefit from social assistance for pupils, which gives us an idea of the socio-economic level of the pupils concerned.*
- *SAGE: This is an examination management system, which gives us an idea of the pupils' level of performance in the final years of the cycle, in standardized examinations.*

It was decided to confine this research to the primary cycle, and perhaps to extend the results to the other school cycles. The number of pupils in the study population is 125354 including 59384 girls. Table 1 presents the statistics for the classes partitioning the database, and Fig. 2 gives an idea of the gender density in the data set.

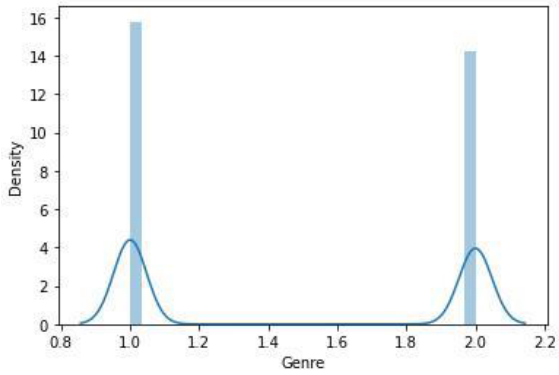

FIGURE 2. The density of male and female students in the experimental data set.

The dataset encompasses a considerable number of properties deemed highly important for defining a student who is likely to drop out or not. Each attribute has one or more binary

TABLE 1. Distribution of data set classes.

Class	Code	Total
Admitted	1	98543
Not Admitted	2	7838
Dropped	3	18973

values, as illustrated in Fig. 3, depending on the significant modalities of each attribute.

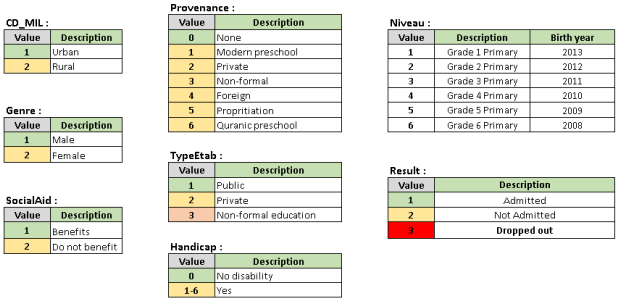

FIGURE 3. The description of features values.

Let's start with the binary attributes. The ''CD_MIL'' feature identifies the area to which the student belongs, whether urban or rural. ''Genre'' disaggregates the population studied by sex. And finally, the ''SocialAid'' feature indicates which pupils benefit from social aid. This aid is illustrated in the field mainly by financial donation to parents of schoolchildren, which is governed by the 'Tyssir' program for tutoring.

Multi-criteria features are very useful for adding various dimensions to the analysis of the student's status who is the subject of our model. Firstly, the ''Provenance'' feature can have a value from 0 to 6, with zero meaning that the student has never benefited from pre-school education, formal or non-formal, modern or traditional, national or foreign or other. This element is known for its considerable impact on learner access and ability to succeed in the primary cycle and avoid early drop-out. The ''TypeEtab'' feature is either public, private or non-formal education. Next, the ''Disability'' feature shows whether or not a student suffers from a disability, and we list 6 types of disability. Obviously, disability is a major factor affecting children's schooling, especially in a non-inclusive education system. The values of the ''Niveau'' feature are the grades of the primary cycle, and to be used afterwards in the academic delay calculation, we highlight the expected year of birth of the pupil at each school level. And finally, the target attribute ''Result'', which is explained in Table 1.

IV. METHODS

The implementation logic of an EWS requires the adoption of a generic data processing model, with the prediction mechanism at its heart. But given the difficulty of gathering information from a variety of sources, a great deal of pre-processing is required. The data arrive missing and inconsistent, so a solidification by join according to the student code is necessary, in addition to an imputation of empty and missing data to complete the database image. The processing model followed is shown in Fig. 3.

TABLE 2. List of transformed features.

Feature	Type	Encoding
Type Etab	Number	Integer 64
Genre	Number	Integer 64
Niveau	Number	Integer 64
RetardSco	Number	Integer 64
Provenance	Number	Integer 64
Handicap	Number	Integer 64
SocialAid	Number	Integer 64
CD_MIL	Number	Integer 64
Moy	Number	Float 64
Result	Number	Integer 64

A. DATA EXTRACTION

Gathering information from the various operational systems represented a real challenge, both in terms of understanding and exploitation. Especially as we are dealing with an architecture of distributed systems where information is dispersed and difficult to detect: a school management system (Massar), a school directory management system (GRESA), a census system (ESISE) and an examination management system (SAGE). This list is limited to the systems we had the opportunity to use in this project, although there are other systems related to other education management professions, such as: the human resources management system(MasiRH), the school mapping system(CarteSco), RAED, PSP and others.

Efforts are being made by the Moroccan Ministry of Education to split all these operational information systems into a single integrated system, but to date the multi-system case persists, and the constraints posed by this are enormous.

B. PRE-PROCESSING

The database gathered through the cross-referencing of the various Information System databases mentioned above is very extensive, but required a number of pre-processing operations. It was difficult to find a unique identifier to perform a natural join between all the databases, so we worked with the student's national code at one moment, or with the student identifier at another. A number of data items required type conversion, especially attributes that were supposed to be considered as numbers. Table 2 lists the attributes subject to transformation:

We've ensured that all features values are mapped to a binary number (1 and 2) whenever possible, as is the case for features such as: Type Etab, Genre, Handicap, SocialAid, CD_MIL. The other features ''Niveau'' and ''Provenance'' contain several possible values, so they are left as they are, as is the ''Moy'' attribute, which is a continuous variable. In addition, occurrences of the ''Genre'' attribute have undergone a regulation and coding/decoding operation, since some systems code males as 1 and females as 2, and others the opposite.

Finally, the "RetardSco" attribute which concerns academic delay is a field calculated through an equation specific to our 2020 study year and the age of the pupils they are supposed to have at each level of the primary cycle. The equation is as follows:

$$\begin{aligned}&\textbf{Academic Delay}\\&= IF\left(Niveau == 1; \left(2014 - Year_{of_{Birth}}\right);\right.\\&\quad IF\left(Niveau == 2; \left(2013 - Year_{of_{Birth}}\right);\right.\\&\quad IF\left(Niveau == 3; \left(2012 - Year_{of_{Birth}}\right);\right.\\&\quad IF\left(Niveau == 4; \left(2011 - Year_{of_{Birth}}\right);\right.\\&\quad IF\left(Niveau == 5; \left(2010 - Year_{of_{Birth}}\right);\right.\\&\quad IF\quad \left(Niveau == 6; \left(2009 - Year_{of_{Birth}}\right)\right)\right)\right)\right)\right)\end{aligned}$$

All the previous manipulations have finally resulted in a stable, consistent data set, and most importantly, it is ready for use. Fig. 5 shows, through the visual correlation matrix, that the data set's predictor attributes are highly independent of each other, meaning that they are all important and indispensable for defining the Target output.

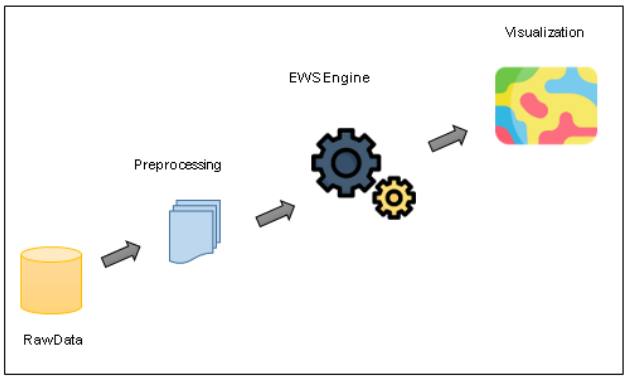

FIGURE 4. The proposed model of the predictive data processing system.

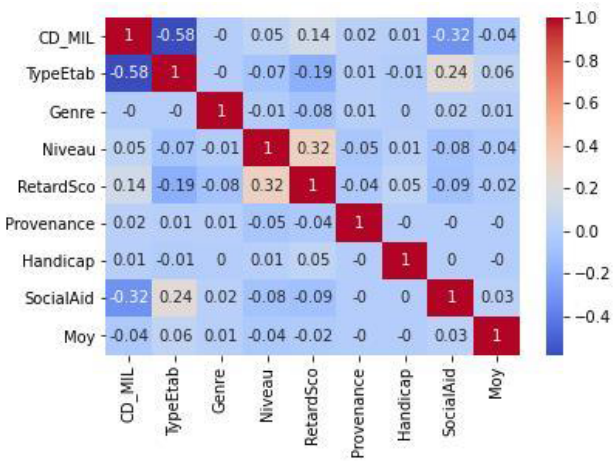

FIGURE 5. Correlation matrix of the experimental data set.

C. EARLY WARNING SYSTEM FOR DROPOUT

The system set up as part of this project has the primary objective of detecting anomalies in students' academic data that explain their dropping out of school at any point during the school year. The analysis system at the heart of the early warning system is organized sequentially, as shown in Fig. 6.

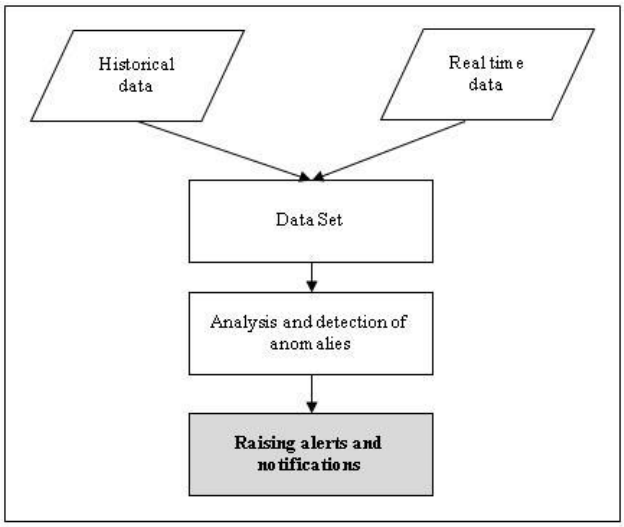

FIGURE 6. The EWS flow chart.

We used mainly machine learning analysis methods specialized in classification. Hence the use of the best-known and most efficient for this aspect of analysis: SVM, Random Forest, SGD and KNN. We have tried to calibrate the internal parameters of each algorithm in order to use the best possible version.

1) SVM

Supervised machine learning encompasses a range of algorithms, one of which is the support vector machine (SVM). This algorithm is specifically designed to handle classification and regression tasks. Its effectiveness is particularly pronounced when dealing with complex datasets containing many features. The main objective of SVM is to locate a hyperplane, which serves as a decision boundary, that optimally separates distinct classes of data points.

The equation of the hyperplane can be represented as follows:

For a linearly separable case:

$$\omega.x + b = 0$$

where:

- w is the weight vector perpendicular to the hyperplane.
- x is the input feature vector.
- b is the bias term (also known as the intercept).

For non-linearly separable cases, SVM uses the concept of a "soft margin" which allows for some misclassification of data points. This involves introducing slack variables and modifying the equation slightly:

$$\omega.x + b - \varepsilon = 0$$

where ξ is the slack variable that represents the margin of error for each data point. The goal becomes to minimize

both the misclassification error and the margin size while still finding the best separating hyperplane.

2) RANDOM FOREST

Random Forest is a popular ensemble learning algorithm used for classification and regression tasks. It aims to improve the accuracy and robustness of predictions by combining the output of several individual models (usually decision trees). The algorithm takes its name from the idea of creating a ''Forest'' of decision trees, each trained on a random subset of the data. The idea behind a Random Forest is to reduce overfitting and improve generalization by leveraging the diversity of multiple decision trees.

$$\text{Random Forest} = \sum \text{Decision Trees}$$

A *decision tree* represents a series of hierarchical decisions based on features that lead to a final prediction or decision. It's a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label (for classification) or a value (for regression).

3) SGD

It stands for Stochastic Gradient Descent, which is a popular optimization algorithm used in machine learning and deep learning for training models, particularly when large datasets are involved. SGD is a variant of the more general gradient descent optimization algorithm, but introduces randomness and speed improvements that make it suitable for large-scale data processing.

The basic equation for updating model parameters using SGD is as follows:

$$\theta_{t+1} = \theta_t - \alpha.\nabla L\left(\theta_t; x_i, y_i\right)$$

where:

- θ_t is the vector of model parameters at iteration t.
- α is the learning rate, a positive scalar that determines the step size of the update.
- $\nabla \boldsymbol{L}(\theta_t ; \boldsymbol{x_i}, y_i)$ is the gradient of the loss function L with respect to the model parameters θ_t, evaluated on the data point (x_i, y_i).

In each iteration, the algorithm selects a single data point (or a small batch of data points) randomly from the training set and computes the gradient of the loss function with respect to the model parameters using that data point. Then, the model parameters are updated in the opposite direction of the gradient to minimize the loss.

4) KNN

K-Nearest Neighbours (KNN) is a simple and intuitive machine learning algorithm for classification and regression tasks. It is instance-based learning that makes predictions based on the majority class or average of the k nearest data points in the feature space. KNN algorithm is calculated using various distance metrics. The two most commonly used distance metrics are Euclidean distance and Manhattan distance. These metrics provide a way to measure the similarity or dissimilarity between two data points in the feature space.

a: EUCLIDEAN DISTANCE

$$\textit{Euclidean Distance}\,(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

where x_i and y_i are the i-th features of data points x and y, respectively. *Euclidean distance* is the straight-line distance between two points in a Euclidean space.

b: MANHATTAN DISTANCE

$$\textit{Manhattan Distance}\,(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

It is calculated as the sum of the absolute differences between the corresponding coordinates of the points. Manhattan distance, also known as the *taxicab* or *city block* distance, is the distance between two points measured along the gridlines.

The predictive model resulting from this processing will be expected to be reused in the future as part of a long-term planning operation. This will give greater visibility to student drop-out problems, and greater scope for action by educational decision-makers.

The EWS will also support a process of data and model evaluation, and we will use a set of the most effective metrics to be confident of the reusability of our model. The danger of this maneuver is that any error in the prediction will automatically lead us to make a mistake about the future of one or more students, resulting in the wrong intervention and the wrong decision for the situation under study.

D. EVALUATION

The performance of a model lies in its ability to increase its degree of accuracy in predicting outputs ($\hat{y}$) by exactly resembling their counterparts in the original test data set (Y). On the other hand, the other facet of the operation is the minimization of the margin of error over the entire database analysis process. Studying the trend in the amount of error also gives an idea of which parameters are detrimental to the model's performance, enabling direct regulatory intervention. The metrics used are listed below:

1) MEAN ABSOLUTE ERROR

Gives the weighted and absolute average of the difference between the predicted and actual value of each element in the database.

$$\boldsymbol{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

2) ROOT MEAN SQUARED ERROR

A metric calculated on the mean square root of the residual standard deviation through the accumulation of the square root of the error.

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(y_i - \hat{y}\right)^2}$$

3) R-SQUARED OR COEFFICIENT OF DETERMINATION

It is a statistical measure that indicates the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model.

$$R^2 = 1 - \frac{sum\ squared\ regression\ \ (SSR)}{total\ sum\ of\ \ squares\ \ (SST)},$$
$$= 1 - \frac{\sum (y_{i-\hat{y}_i})^2}{\sum (y_{i-\bar{y}})^2}$$

4) ROC CURVE

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to assess the performance of a binary classification model at various thresholds. It illustrates the trade-off between the true positive rate (also called sensitivity or recall or TPR) and the false positive rate(FPR) as the classification threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis for different classification thresholds. Each point on the curve represents the model's performance at a specific threshold. The two terms are defined as follows:

- $\textbf{\textit{Sensitivity}} = TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$
- $\textbf{\textit{Specificity}} = FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$

All these metrics are widely used, mainly to evaluate linear models. But the most widely used are RMSE, R2 and the ROC Curve. RMSE is clearly better than MAE at detecting the net error of different regression models. And R2 is very effective for specifying the degree of strength with which the independent variables in the data set explain the target dependent variable. The ROC Curve, on the other hand, exposes the degree of sensitivity and specificity recorded at each stage of the prediction operation, each time the threshold varies proportionally. This should have a direct impact on the predictive consistency of the model's outputs, in addition to the possibility of studying these indicators for each class of the target attribute.

Consequently, the model is said to have high performance when the MAE and RMSE measures are low. On the other hand, the model is highly accurate when the R2 value is high and the AUC of the ROC Curve tends upwards and is close to or equal to 1.

E. VISUALIZATION

Visualization is considered a very important element in any data processing project. It presents a consistent picture of the analysis process's results, providing a basis for exploring the situation under analysis, and enabling decision-makers to reflect on the actions needed to remedy the problems experienced.

We have developed an MVC architecture application as the final product of our work. It is elaborated with the Django 4.2 Framework and using the Bootstrap MDB package. The application was coded using Visual Studio Code 1.81.1, running on top of the Anaconda 1.10 distribution environment. On the one hand, the application will include the task of collecting information, which we first present to the system in the form of a global CSV file for direct processing. On the other hand, the process of analysing the data set will be carried out at system level, along with visualization.

For the time being, we have decided to limit ourselves to three forms of result visualization: Tables, Maps and Charts. Tables present overall student information, as well as notifications of those likely to dropout. The geographical maps and graphs are thematic, and we start with maps that geo-locate the communes most suffering from the dropout phenomenon, and the graphs will demonstrate the trends of the phenomenon's variability between these communes.

TABLE 3. Performance measurements for the algorithms used.

	TRAINING SET			TEST SET		
	MAE	**RMSE**	**R2**	**MAE**	**RMSE**	**R2**
Random Forest	0.0057	0.0606	0.9931	0.0090	0.0868	0.9858
SVM	0.0095	0.1068	0.9913	0.0090	0.1051	0.9919
SGD	0.0123	0.1193	0.9885	0.0120	0.1183	0.9889
KNN	0.0056	0.0863	**0.9953**	0.0076	0.0988	**0.9933**

V. RESULTS AND DISCUSSION

In this work, we experimented with a range of methods and algorithms to arrive at the most important threshold of our results, especially concerning the accuracy of our model, for which we calculated the coefficient of determination R2. We also used the MAE and RMSE measures to draw up the loss curve and identify the model with the least error.

A. EWS' LINEAR MODEL

Since the problem is of a classification nature, we opted to explore the most famous algorithms in this field, citing: SVM, Random Forest, SGD and KNN. The results, as presented in Table 3, are highly satisfactory and allow the model to be reused with confidence.

In the end, the KNN algorithm performed the best, recording the least loss (MAE & RMSE) and automatically the best accuracy score (R2) at over 99.5% for the training set and over 99.3% for the test set.

B. FEATURE IMPORTANCE

The model built was based on a number of features that mainly guided the outcome of each prediction, and enabled a certain classification of the training set data. Table 4 shows the weight of each feature in the importance of its participation in the choice of model output (Target label).

Let's consider that the most important attribute in the definition of the Label is the grade point average, which represents a synthetic picture of the student's academic performance that merges all subject grades throughout the year as well as exam averages, especially for the certifying grade

TABLE 4. The weight of each feature in the target prediction.

Feature	Weight
Moy	0.9962324622
RetardSco	0.0010250153
Genre	0.0008447858
Provenance	0.0007557517
Niveau	0.0006246890
SocialAid	0.0002653846
CD_MIL	0.0001764032
TypeEtab	0.0000722366
Handicap	0.0000032718

6 level. The second attribute in order of importance is the academic delay, which by another expression means grade repetition, has an impact on the student's future academic progress.

The ''Genre'' attribute is also considered very important in the model, since in developing countries, gender remains one of the first factors to impact school dropout. Subsequently, the attributes ''Provenance'' and ''Niveau'' have a similar level of importance. Where a pupil comes from before entering primary school is part of his or her educational history, and could have an impact on whether or not he or she continues his or her studies. A pupil who has benefited from pre-school education, for example, has a cognitive base with which he or she can easily approach knowledge and courses in primary school, and also has a psychological advantage in that he or she is familiar with the school and classroom climate. Level also plays a part in the dropout equation, since the higher a pupil climbs, the more academic failures accumulate, and the idea of leaving school to seek other alternatives is strongly present.

There remain a number of attributes that can be considered to have a less significant effect on the prediction of a Label class. The first is social support, which the model did not mark as very important, even if a student who benefits from a type of social support, especially financial aid, is considered to come from an underprivileged and precarious environment. The area is the other facet of the notion of social support, since housing environment automatically refers to social level and the presence of material means. The model also ranks the Type of school attribute at the bottom of the list. The sociology of education speaks of the effect of the school, and imperatively considers that pupils educated in so-called unlucky schools (in poor or disadvantaged environments) are more likely to fail at school, and hence to dropout. The last attribute in the importance ranking is Disability, which can be interpreted as meaning that students with special needs have almost the same chance of continuing their school career or dropping out as normal students, perhaps because of the inclusive pedagogy and classrooms that cater for these students with special needs.

If we focus on the indicators for students who drop out of school, analysis of the standard deviation entities shows that the majority of students affected by this phenomenon can be modelled under the following data vector:

$$\begin{aligned} &\textbf{\textit{Dropped out}} \\ &= \{CD\_MIL == 2;\ TypeEtab == 1; \\ &\quad Genre == 2;\ Niveau == 4; \\ &\quad RetardSco == 2;\ Provenance == 1; \\ &\quad Handicap == 0;\ SocialAid == 2;\ Moyenne == 0\} \end{aligned}$$

This vector can be read as follows, students most likely to drop out of school are those: living in rural areas, students in public schools, female, in grade 4 and above, who are necessarily considerably behind in school, who come from preschool or not, disability is not an important factor, students who benefit from tutoring, and finally who have a low performance at school.

C. ROC CURVE

It is very instructive to study further the degree of correlation between the attributes independent of the predicted attribute, and the quantification of the ability to predict exactly the expected result or the gap between the set and predicted value. This is possible through the construction of the ROC Curve in Fig. 7, which in our case affirms that our model is indeed almost perfect in predicting the various results of the data set, especially for students at risk of dropping out, and which is considered to be the purpose of our analysis.

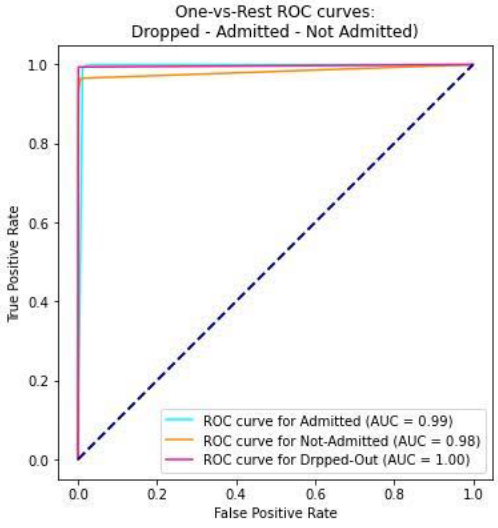

FIGURE 7. ROC Curve of Label class modalities.

D. VISUALIZATION

To show the results of applying machine learning techniques, we may need to use various visualization and reporting tools. For our case, in Fig. 8 we have created a table that displays the entire Data Set by pagination, mentioning the result predicted by the EWS. The table can be searched by free keywords

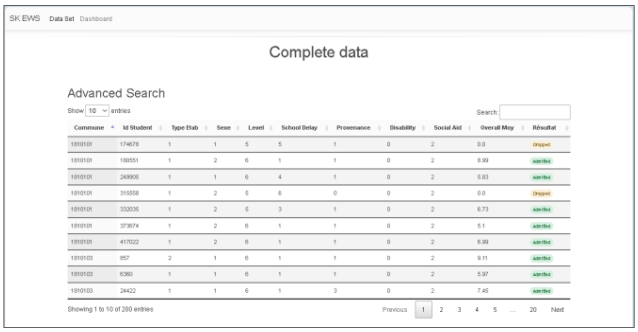

FIGURE 8. Data Set table with notification of prediction results.

for all rows containing it, and is sorted in ascending order of student IDs.

On the other hand, Fig. 9 presents a visualization of two kinds: a thematic map of communes with higher and lower dropout rates. In addition, a curve graph showing the level of the dropout indicator in relation to the communes.

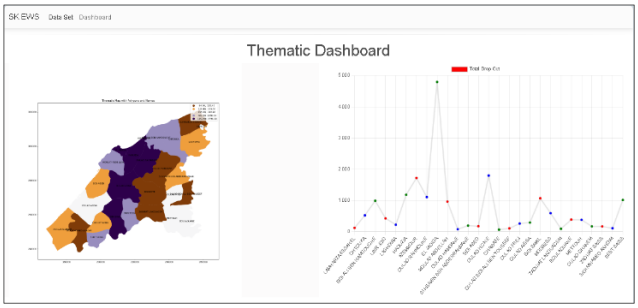

FIGURE 9. Visualization of the most affected areas by the drop-out phenomenon and their classification.

By reading the figures, we can pinpoint the areas most affected by the dropout phenomenon, and thus make appropriate decisions that will have an impact on the current situation. Depending on the most important explanatory factors of the predictive system, several types of decisions can be taken: interventions can target the construction of new schools in areas with high dropout rates, or plan the extension of existing schools. There will also be a need for tutoring, school transport and school access facilities. Interventions can also take the form of administrative, communication or structural regulations. In any case, the planning process will need to take urgent note of EWS alerts in order to remedy the situation as soon as possible.

E. DISCUSSION

It is considerably important to state that there have been various projects and initiatives that have attempted to build a comprehensive database bringing together all the factors influencing student drop-out. Educational data, like all data from scientific fields, presents a real difficulty in its collection, processing and explanation. The term ‘‘educational sciences’’ has recently been coined to describe the complex scientific dimensions of education.

Lee et al. [8] took a snapshot of the National Education Information System (NEIS) database in South Korea for the year 2014. They used 15 features to predict the dropout of high school students, spread over several behavioral, family, academic and other aspects. This yielded very significant results, with a prediction accuracy of 99% for the Target attribute, especially by applying the boosted decision tree (BDT) algorithm with minority oversampling techniques (SMOTE) to avoid the drawbacks of class-imbalanced datasets. On the one hand, the binary classification employed in this model makes the results more discriminating, even if it hides from another angle all the nuances that can nevertheless mark the students’ situation. On the other hand, the battery of factors used to predict drop-out neglected students’ personal and demographic properties, focusing instead on what might be described as external and environmental factors.

Bañeres et al. [26] have implemented a Gradual At-Risk (GAR) model based on online data from the Universitat Oberta de Catalunya (UOC) datamart. The aim of the project is to predict which university students are most likely to fail a specific course, and through the accumulation of these failures this may lead to dropping out of education. The data exploited concerns student activity on the UOC’s operational systems, in addition to course evaluation data over a long period aggregated by time. The GAR model built by this project consistently performed best with the K-Nearest Neighbors algorithm for all courses, exceeding 95% accuracy. Most efforts to model students on e-learning platforms are based on historical records of enrollee activity on one or more courses, or on a given curriculum. This information base is still insufficient to carry out a thorough and exhaustive study of the factors leading a registrant to drop out of a course. As in the case of the UOC, which caters for higher education students, other cycles and levels of education are masked and attract no interest, given their complexity in reaching this kind of learner, especially with the generalization of elementary education in every country in the world.

Realinho et al. [37] also developed a consistent dataset resulting from a fusion of information from several systems of a higher education institution Polytechnic Institute of Portalegre(PIP) in Portugal. They worked to enrich the database with several attributes that could be classified according to the following classes: academic path, demographics, macroeconomics, socioeconomic factors and student performance by grades. In order to effectively predict the dropout phenomenon for all students enrolled in the institute’s academic disciplines, they focused on determining the most important attributes weighing on the predictive model, and thereby raising the accuracy of the system. This database reflects very roughly the idea we’re trying to define through this project. The construction of our database has tried to integrate as many explanatory factors as possible for the dropout phenomenon. The academic database of Realinho et al. [37] gave researchers the chance to delve deeply into the logic of academic drop-out, and several manipulations served to increase the quality of this data set. There were several similarities to our test, in terms of the analysis factors used to predict and the problem processing logic. Table 5 reflects the parallelism

TABLE 5. Benchmarking of our project results.

	Analysis factors					
	Demographic	Socio economic	Macro economic	Academic	Environmental	Applied Method
Our data	*x*	*x*		*x*	*x*	**KNN** (99%)
PIP data	*x*	*x*	*x*	*x*		**RF** (90%)

between the two databases used in the quest to model the dropout phenomenon.

Furthermore, through our database, we have tried to be exhaustive in assembling the factors essential to an accurate and fair prediction of the school dropout phenomenon. On the other hand, we have tried in our project to turn to points that have been marginalized in educational research. Firstly, we have chosen to focus on the primary cycle, as we consider the importance and impact of the quality of a pupil's journey through this cycle on his or her retention in school, mainly during the period of obligatory schooling. On the other hand, we attach considerable importance to the impact of the EWS on the regulation of educational planning. This is motivated by the idea that, to deal with the dropout problem rigorously, we need to go beyond pedagogical changes in teaching or interventions in teaching practice, and develop a comprehensive educational planning strategy that targets all aspects of education: pedagogical, structural, organizational, social and human. A narrow perception of the problem can be detrimental to the solution, and raising the alarm about students at risk is only the first step in identifying the dropout phenomenon.

VI. CONCLUSION AND FUTURE WORK

Education is a highly sensitive area of strategic national security, since the development of any country has a direct bearing on the level of literacy and training of its citizens. That's why the phenomenon of school drop-outs represents a dilemma for those in charge of education, and is seen as a drain on financial resources and a waste of effort with no return on investment for those generations who don't continue their learning. This situation is seen as even more delicate when it comes to early school wastage, as in the primary cycle, which is why we have chosen to focus on data from this cycle.

Despite all the initiatives aimed at tackling this phenomenon, it still persists, especially in contexts where there is a lack of resources. Our EWS project aims to provide a panoramic view of the various factors that can cause school dropout, as well as a proactive vision of the areas most affected by this phenomenon, or the individuals likely to be victims of dropout. These indicators will need to be taken on board by educational planners in order to target areas where the phenomenon is most prevalent with appropriate policies and interventions.

Although the system is performing well, there is still considerable scope for development. A first dimension is the expansion of the Data Set, both in terms of the data quantity, but above all the addition of other attributes that give greater precision to the predictive model. For the Data Set we developed for this project, we did our best to gather as many indicators as possible, but this wasn't easy given the difficulty of accessing the information and its dispersion over several operational systems with diversified forms. Another dimension of evolution is the upgrading of our client application by adding further interfaces for administration and data manipulation in dynamic and interactive ways. Finally, we may work on a recommendation system that takes into account the outputs of the EWS in order to propose appropriate interventions in predicted situations.

DATA AVAILABILITY

Source codes and models used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors have no conflict of interest to disclose.

FUNDING STATEMENT

There is no funding to declare for this work.

REFERENCES

[1] Y. Han and S. Liu, "Construction and research of big data platform for party building in colleges and universities," in *Proc. 2nd Int. Conf. Internet, Educ. Inf. Technol.*, 2022, pp. 431–436, doi: 10.2991/978-94-6463-058-9_70.

[2] A. N. de Vasconcelos, L. A. Freires, G. D. L. Loureto, G. Fortes, J. C. A. da Costa, L. F. F. Torres, I. I. Bittencourt, T. D. Cordeiro, and S. Isotani, "Advancing school dropout early warning systems: The IAFREE relational model for identifying at-risk students," *Frontiers Psychol.*, vol. 14, Jul. 2023, Art. no. 1189283, doi: 10.3389/fpsyg.2023.1189283.

[3] B. M. Mcmahon and S. F. Sembiante, "Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention," *Rev. Educ.*, vol. 8, no. 1, pp. 266–301, Feb. 2020, doi: 10.1002/rev3.3183.

[4] Z. Alharbi, J. Cornford, L. Dolder, and B. De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," in *Proc. SAI Comput. Conf. (SAI)*, Jul. 2016, pp. 523–531, doi: 10.1109/SAI.2016.7556030.

[5] M. S. Ahmad, A. H. Asad, and A. Mohammed, "A machine learning based approach for student performance evaluation in educational data mining," in *Proc. Int. Mobile, Intell., Ubiquitous Comput. Conf. (MIUCC)*, May 2021, pp. 187–192, doi: 10.1109/MIUCC52538.2021.9447602.

[6] Y.-H. Hu, C.-L. Lo, and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," *Comput. Hum. Behav.*, vol. 36, pp. 469–478, Jul. 2014, doi: 10.1016/j.chb.2014.04.002.

[7] Y.-C. Chang, W.-Y. Kao, C.-P. Chu, and C.-H. Chiu, "A learning style classification mechanism for e-learning," *Comput. Educ.*, vol. 53, no. 2, pp. 273–285, Sep. 2009.

[8] S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Appl. Sci.*, vol. 9, no. 15, p. 3093, Jul. 2019, doi: 10.3390/app9153093.

[9] S. Bansal and N. Baliyan, "A study of recent recommender system techniques," *Int. J. Knowl. Syst. Sci.*, vol. 10, no. 2, pp. 13–41, Apr. 2019, doi: 10.4018/ijkss.2019040102.

[10] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea, and C. Heiner, "An educational data mining tool to browse tutor-student interactions: Time will tell," in *Proc. Workshop Educ. Data Mining, National Conf. Artif. Intell.*, 2005, pp. 15–22.

[11] M. A. M. Iver and D. J. M. Iver, "Beyond the indicators: An integrated school-level approach to dropout prevention," George Washington Univ. Center Equity Excellence Educ., Arlington, VA, USA, Tech. Rep. ED543512, 2009.

[12] U. B. Qushem, S. S. Oyelere, G. Akcapinar, R. Kaliisa, and M. J. Laakso, "Unleashing the power of predictive analytics to identify at-risk students in computer science," *Technol., Knowl. Learn.*, vol. 28, pp. 1–16, Jul. 2023, doi: 10.1007/s10758-023-09674-6.

[13] W. M. Ei Leen, N. A. Jalil, N. M. Salleh, and I. Idris, "Dropout early warning system (DEWS) in Malaysia's primary and secondary education: A conceptual paper," in *Proc. Int. Conf. Inf. Syst. Intell. Appl.* (Lecture Notes in Networks and Systems), vol. 550, 2023, pp. 427–434, doi: 10.1007/978-3-031-16865-9_33.

[14] M. Yagci, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environments*, vol. 9, no. 1, Dec. 2022, Art. no. 11, doi: 10.1186/s40561-022-00192-z.

[15] X. Sun, Y. Fu, W. Zheng, Y. Huang, and Y. Li, "Big educational data analytics, prediction and recommendation: A survey," *J. Circuits, Syst. Comput.*, vol. 31, no. 9, Jun. 2022, Art. no. 2230007, doi: 10.1142/S0218126622300070.

[16] T. A. Kustitskaya, A. A. Kytmanov, and M. V. Noskov, "Early student-at-risk detection by current learning performance and learning behavior indicators," *Cybern. Inf. Technol.*, vol. 22, no. 1, pp. 117–133, Mar. 2022, doi: 10.2478/cait-2022-0008.

[17] Y. Wu, "Design framework of early warning mechanism of student achievement with machine learning," in *Proc. IEEE 4th Eurasia Conf. IoT, Commun. Eng. (ECICE)*, Oct. 2022, pp. 239–241, doi: 10.1109/ECICE55674.2022.10042897.

[18] Y. Qu, Z. Sun, and L. Liu, "Research on the academic early warning model of distance education based on student behavior data in the context of COVID-19," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 633–642, 2022, doi: 10.14569/IJACSA.2022.0131074.

[19] S. Yi, "The development of early warning system for college English academic performance based by big data computation," in *Proc. Int. Conf. Comput., Big-Data Eng. (ICCBE)*, May 2022, pp. 226–229, doi: 10.1109/ICCBE56101.2022.9888191.

[20] A. B. Soussia, A. Roussanaly, and A. Boyer, "Toward an early risk alert in a distance learning context," in *Proc. Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2022, pp. 206–208, doi: 10.1109/ICALT55010.2022.00067.

[21] L. You, "Construction of early warning mechanism of university education network based on the Markov model," *Mobile Inf. Syst.*, vol. 2022, pp. 1–9, Jul. 2022, doi: 10.1155/2022/7302623.

[22] D. Baneres, A. E. Guerrero-Roldán, M. E. Rodríguez-González, and A. Karadeniz, "A predictive analytics infrastructure to support a trustworthy early warning system," *Appl. Sci.*, vol. 11, no. 13, p. 5781, Jun. 2021, doi: 10.3390/app11135781.

[23] A. K. Veerasamy, M. J. Laakso, D. D'Souza, and T. Salakoski, "Predictive models as early warning systems: A Bayesian classification model to identify at-risk students of programming," in *Intelligent Computing* (Lecture Notes in Networks and Systems), vol. 284. Cham, Switzerland: Springer, 2021, pp. 174–195, doi: 10.1007/978-3-030-80126-7_14.

[24] D. Baneres, A. Karadeniz, A. E. Guerrero-Roldan, and M. E. Rodriguez, "A predictive system for supporting at-risk students' identification," in *Proc. Future Technol. Conf.* (Advances in Intelligent Systems and Computing), vol. 1288, 2021, pp. 891–904, doi: 10.1007/978-3-030-63128-4_67.

[25] M. B. Ada and K. Turinicova, "Developing a dual dashboard early detection system," in *Proc. IEEE 20th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2020, pp. 155–157, doi: 10.1109/ICALT49669.2020.00052.

[26] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz, "An early warning system to detect at-risk students in online higher education," *Appl. Sci.*, vol. 10, no. 13, p. 4427, Jun. 2020, doi: 10.3390/app10134427.

[27] M. Chunqiao, Y. Niefang, and P. Xiaoning, "Method and system constructing for learning situation early warning based on data mining techniques," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 964–969, doi: 10.1109/ICCSE.2019.8845370.

[28] Z. Wang, C. Zhu, Z. Ying, Y. Zhang, B. Wang, X. Jin, and H. Yang, "Design and implementation of early warning system based on educational big data," in *Proc. 5th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2018, pp. 549–553, doi: 10.1109/ICSAI.2018.8599357.

[29] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Exp. Syst.*, vol. 33, no. 1, pp. 107–124, Feb. 2016, doi: 10.1111/exsy.12135.

[30] W. Wen, Y. Liu, Z. Zhu, and Y. Shi, "A study on the learning early warning prediction based on homework habits: Towards intelligent sustainable evaluation for higher education," *Sustainability*, vol. 15, no. 5, p. 4062, Feb. 2023, doi: 10.3390/su15054062.

[31] H. Wei, L. Wang, D. He, and Y. Li, "Research on the early warning system of college students' mental health based on big data analysis," in *Proc. Int. Conf. Comput. Netw., Electron. Autom. (ICCNEA)*, Sep. 2022, pp. 50–54, doi: 10.1109/ICCNEA57056.2022.00022.

[32] A. Rahmah, "Designing early warning system for course completion using learning management system," in *Proc. 6th Int. Conf. Informat. Comput. (ICIC)*, Nov. 2021, pp. 1–5, doi: 10.1109/ICIC54025.2021.9632939.

[33] H. Colak Oz, Ç. Güven, and G. Nápoles, "School dropout prediction and feature importance exploration in Malawi using household panel data: Machine learning approach," *J. Comput. Social Sci.*, vol. 6, no. 1, pp. 245–287, Apr. 2023, doi: 10.1007/s42001-022-00195-3.

[34] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, no. 1, Jan. 2023, doi: 10.1186/s41239-022-00371-5.

[35] V. Siafis and M. Rangoussi, "Educational data mining-based visualization and early prediction of student performance: A synergistic approach," in *Proc. 26th Pan-Hellenic Conf. Informat.*, Nov. 2022, pp. 246–253, doi: 10.1145/3575879.3576000.

[36] M. Skittou, M. Merrouchi, and T. Gadi, "A model of an integrated educational management information system to support educational planning and decision making: A Moroccan case," in *WITS* (Lecture Notes in Electrical Engineering), vol. 745, S. Bennani, Y. Lakhrissi, G. Khaissidi, A. Mansouri, and Y. Khamlichi, Eds. Singapore: Springer, 2022, doi: 10.1007/978-981-33-6893-4_16.

[37] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, Oct. 2022, doi: 10.3390/data7110146.

MUSTAPHA SKITTOU received the master's degree in computer science (option: engineering of conception and software development) from the Faculty of Sciences and Technics, Hassan First University of Settat, Settat, Morocco, in 2016, where he is currently pursuing the Ph.D. degree in computer science.

He is also an Educational Planner and the Head of Service for the Ministry of National Education in Morocco. His research interests include big data and artificial intelligence.

MOHAMED MERROUCHI received the M.S. degree in computer science (option: distributed information systems) from the University of Hassan II, Casablanca, Morocco, in 2017, where he is currently pursuing the Ph.D. degree.

He is also a member of the Research Laboratory in Mathematics, Computer Science and Engineering, Faculty of Science and Technology, University of Hassan II, Settat, Morocco. He is also a Teacher for the Ministry of National Education in Morocco. His research interests include machine learning, deep learning, and big data.

TAOUFIQ GADI received the M.S. degree in computer science and the Ph.D. degree from Sidi Mohamed Ben Abdellah, Fes, Morocco, in 1994 and 1999, respectively. He is currently a full Professor in computer science with the Faculty of Science and Technologies, Hassan First University of Settat, Morocco. He has conducted more than 20 Ph.D. theses and written a 70 of scientific articles in the domain of 3D models analysis, models driving architecture, datamining and database analysis, the modeling of complex systems, and machine learning.