

New prediction model for school dropout based on an improvement of SVM and ABC optimization algorithm

Hicham EL Yousfi Alaoui¹, Ziad Bousraraf², Amal Hjouji³, Omar EL Ogri⁴,
Jaouad EL-Mekkaoui¹, Zaynab Hjouji²

¹LSIG, Higher School of Technology of Fez, Sidi Mohamed Ben Abdellah University, Fez, Morocco

²LESSOR, Faculty of Legal, Economic and Social Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco

³LISAC Laboratory, Faculty of Sciences of Fez, Sidi Mohamed Ben Abdellah University, Fez, Morocco

⁴LSAM Laboratory Sidi Bennour Higher School of Technology, Chouaib Doukkali University, El Jadida, Morocco

Article Info

Article history:

Received Apr 20, 2025

Revised Jun 12, 2026

Accepted Jul 5, 2026

Keywords:

Artificial bee colonies
Classification methods
Management of academic
success
Regression model
Student dropout
Support vector machines

ABSTRACT

School dropout is a major dare for Moroccan education systems. It generates profound repercussions, both socially and economically, affecting both individuals and the entire community. To assess the risks associated with this phenomenon, it is essential to use innovative technologies, relying in particular on statistical methods and artificial intelligence models capable of providing effective results using available data. These methods offer the possibility of identifying the factors influencing the educational phenomenon and anticipating the associated risks. To understand the effect of scholars' social, economic, and personal influences on school dropout, we propose in this work a new classifier, based on an improvement of support vector machines (SVM) and on the ABC algorithm. We also present a regression model designed to estimate academic performance based on the average grade obtained. We also propose a new regression model to predict the degree of academic achievement. The proposed model was validated on a real dataset, compiled from a questionnaire distributed to Moroccan secondary school students. Furthermore, several comparative studies were conducted with recent approaches to assess the performance of the proposed methods. All the experimental results confirm the robustness and improved performance of the developed methods, particularly with respect to accuracy and predictive capability.

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



Corresponding Author:

Jaouad EL-Mekkaoui

LSIG, Higher School of Technology of Fez, Sidi Mohamed Ben Abdellah University

B.P. 2202, Fez, Morocco

Email: jaouad.elemekkaoui@usmba.ac.ma

1. INTRODUCTION

Dropout is one of the most prominent problems facing the Moroccan education system and poses a major challenge to the government. In the 2021-2022 academic year, more than 300,000 students dropped out of school. According to statistics, these students are divided into three categories: the first is those who dropped out of school during the academic year, representing 15% of the total number of dropouts. The second category represents 30.4% of students who did not register at the beginning of the academic year. The third category is students who were expelled from school due to repeated failure, representing 54.6% of the total number of dropouts. This situation raises many questions about the reasons that contribute to the aggravation of this problem. Poverty is clearly one of the main factors, but it is not the only one. Social and economic conditions also play a decisive role, in addition to negative school experiences, regional difficulties, and an unsuitable We can address the prediction of school dropout using a classification system designed to

determine whether a student is likely to succeed or fail. Thus, we are faced with two interdependent educational phenomena: the first concerns academic performance and the second, dropping out of school. Analyzing school performance requires the use of regression methods, while studying the phenomenon of school dropout relies on classification algorithms.

Generally, the development of a regression model [1], based on a dataset, relies on the construction of a relationship expressed by:

$$y = f(X) + \varepsilon \quad (1)$$

In this equation, y denotes the predicted value, f corresponds to the regression function, $X = (x_1, x_2, \dots, x_k)$ represent the explanatory variables, ε is the random error. According to choose of f , the literature proposes several types of regression models [2]-[10]. Multiple polynomial regression (PMR) can be used for multiple independent variables based on polynomial functions of multiple independent variables. The model of this regression method is defined by:

$$y = P_n(X) + \varepsilon \quad (2)$$

where P_n is a polynomial defined by,

$$P_n(X) = \sum_{i_1+i_2+\dots+i_k \leq n} \beta_{i_1, i_2, \dots, i_k} x_1^{i_1} x_2^{i_2} \dots x_k^{i_k} \quad (3)$$

the exponential regression model [11], [12] defined as:

$$y = a \cdot e^{b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_k \cdot x_k} + \varepsilon \quad (4)$$

where $b_1, b_2, \dots, b_k$ and a are the parameters of the model.

In this work, we combine the two previous models (2) and (4), and we then propose the following regression model.

$$y = \sum_{i_1+i_2+\dots+i_k \leq n} \beta_{i_1, i_2, \dots, i_k} x_1^{i_1} x_2^{i_2} \dots x_k^{i_k} + a \cdot e^{b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_k \cdot x_k} + \varepsilon \quad (5)$$

To estimate the coefficients of a regression model we need an optimization method. Traditional methods used, such as gradient descent method, conjugate gradient [13], [14], can cause overfitting and make these methods less efficient. We use the ABC optimization algorithm [15] to estimate the coefficients of our model.

Much research on school dropout relies on machine learning methods [16]-[22]. Among these methods, among the most commonly used is the support vector machine (SVM) method, defined by linear hyperplane in (6).

$$HYP(X) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (6)$$

We introduce a new SVM by replacing linear hyperplane HYP with a polynomial hyperplane of degree n , denoted HYP_n and defined as (7).

$$HYP_n(x_1, x_2, \dots, x_k) = \sum_{i_1+i_2+\dots+i_k \leq n} \beta_{i_1, i_2, \dots, i_k} x_1^{i_1} x_2^{i_2} \dots x_k^{i_k} \quad (7)$$

To estimate the parameters $\beta_{i_1, i_2, \dots, i_k}$ of our classification model, we use the ABC optimization method. The resulting model is used to predict school dropout rates.

2. THE PROPOSED REGRESSION MODEL

The estimation of the parameters $\vec{\beta} = (a, b_1, b_2, \dots, b_k, \beta_{i_1, i_2, \dots, i_k}; i_1 + i_2 + \dots + i_k \leq n)$ of our regression model presented in (5) is based on the following optimization problem:

$$\text{Minimize } l(\vec{\beta}) = \|\varepsilon(\vec{\beta})\|^2 \quad (8)$$

were,

$$\|\varepsilon(\vec{\beta})\|^2 = \sum_{i=1}^N (P_n(x_{1i}, x_{2i}, \dots, x_{ki}) + a \cdot e^{b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \dots + b_k \cdot x_{ki}} - y_i)^2 \quad (9)$$

is the difference between true values y_i and the output calculated by $y(x_i)$ and $[X_i; y_i]_{i=1, \dots, N}$ is a training set and $X_i = (x_{1i}, x_{2i}, \dots, x_{ki})$.

Due to the complexity and experimental nature of the objective function, we use ABC optimization method to solve the problem (9) in order to obtain the optimal parameter vector $\vec{\beta} = (a, b_1, b_2, \dots, b_k, \beta_{i_1, i_2, \dots, i_k}; i_1 + i_2 + \dots + i_k \leq n)$:

3. THE PROPOSED PREDICTION AND CLASSIFICATION MODEL

This section describes the new classification model proposed for assessing school dropout, based on an improvement of SVM classifier. The first subsection reviews the main steps of the SVM classifier. The second subsection then presents our generalized classification model, in which the linear hyperplane is replaced by a nonlinear hyperplane.

3.1. A reminder about the classical SVM classifier

The classical SVM classification technique is suitable for handling discrimination problems, that is, determining the category to which an individual belongs. To achieve this, it is necessary to construct a linear hyperplane *HYP* that associates an output y with an input x .

$$y = HYP(x) \quad (10)$$

For now, we limit ourselves to a two-class discrimination problem (binary discrimination), i.e., $y \in \{-1, 1\}$. The simple case is that of a linear function, obtained by linear combination of $X = (x_1, x_2, \dots, x_k)$, with the parameter vector $\beta = (\beta_0, \beta_1, \dots, \beta_k)$.

$$HYP(x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (11)$$

The decision boundary $HYP(x_1, x_2, \dots, x_k) = 0$ is a hyperplane, called the separating hyperplane, or separatrix. The sign of $HYP(x)$, which gives the side of the hyperplane in which x is located, then makes it possible to decide the class of x . It is then decided that x is of class 1 if $HYP(x_1, x_2, \dots, x_k) \geq 0$ and of class -1 otherwise. This is a linear classifier.

The goal of supervised classification is to construct the function $HYP(X)$ through a training set:

$$\{(X_1, l_1), (X_2, l_2), \dots, (X_p, l_p)\} \subset \mathbb{R}^k \times \{-1, 1\} \quad (12)$$

where the $X_i, i = 1, \dots, p$ are the points in an k -dimensional space, the $l_i, i = 1, \dots, p$ are the labels that are ± 1 , p is the number of elements in the training set, and k is the number of input variables. If the dataset is linearly separable, then the sign of $HYP(X_i)$ is equal to l_i , for all i . In other words:

$$l_i \cdot HYP(x_i) \geq 0 \quad 1 \leq i \leq p, \text{ in other words } l_i(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}) \geq 0 \quad 1 \leq i \leq p \quad (13)$$

The margin is the distance between the driving elements and the hyperplane that satisfies the maximum separability condition (i.e., $l_i(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}) \geq 0$) as explained previously). The distance of a sample $X_i = (x_{1i}, x_{2i}, \dots, x_{ki})$ to the hyperplane is determined by its orthogonal projection onto the parameter vector $(\beta_1, \beta_2, \dots, \beta_k)$.

$$\frac{l_i(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})}{\|(\beta_1, \beta_2, \dots, \beta_k)\|} \quad (14)$$

The coefficients β of the separating hyperplane with maximum margin is therefore given by (15).

$$\arg \max_{\beta} \left\{ \frac{1}{\|\beta\|} \min_{i=1, \dots, p} [l_i(HYP(x_{1i}, \dots, x_{ki}))] \right\} \quad (15)$$

In order to facilitate the optimization, we choose to normalize $(\beta_1, \beta_2, \dots, \beta_k)$ and β_0 , in such a way that the samples at the margin (x_{marge}^+ for the support vectors on the positive boundary, and x_{marge}^- for those located on the opposite boundary) satisfy:

$$\begin{cases} (\beta_1, \beta_2, \dots, \beta_k)^T x_{\text{marge}}^+ + \beta_0 = 1 \\ (\beta_1, \beta_2, \dots, \beta_k)^T x_{\text{marge}}^- + \beta_0 = -1 \end{cases} \quad (16)$$

hence for all samples, $i = 1, \dots, p$.

$$l_i((\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})) \geq 1 \quad (17)$$

This normalization is sometimes called the canonical form of the hyperplane, or canonical hyperplane. With this scaling, the margin is now worth $\frac{1}{\|(\beta_1, \beta_2, \dots, \beta_k)\|}$, the previous maximization problem is equivalent.

$$\begin{cases} \text{Minimize } \frac{1}{2} \|(\beta_1, \beta_2, \dots, \beta_k)\|^2 \\ \text{under constraints:} \\ l_i((\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})) \geq 1 \end{cases} \quad (18)$$

3.2. Our classification model

We present our SVM classification method, replacing the linear hyperplane *HYP* used in classical SVM classifier presented in by a polynomial hyperplane *HYP_n* of degree *n* defined in (7). The coefficients $\vec{\beta} = (\beta_{i_1 i_2 \dots i_k}; i_1 + i_2 + \dots + i_k \leq n)$ must be calculated by maximizing the function presented in (15), replacing *HYP*(*x*_{1i}, ..., *x*_{ki}) by *HYP_n*(*x*_{1i}, ..., *x*_{ki}) and β by $\vec{\beta}$. So, to determine these coefficients, we solve the optimization model (19) using the ABC optimization algorithm:

$$\begin{cases} \arg \max_{\vec{\beta}} \left\{ \frac{1}{\|\vec{\beta}\|} \min_{i=1, \dots, p} [l_i(HYP_n(x_{1i}, \dots, x_{ki}))] \right\} \\ \text{under constraints:} \\ l_i\left(\sum_{i_1+i_2+\dots+i_k \leq n} \beta_{i_1, i_2, \dots, i_k} x_{1i}^{i_1} x_{2i}^{i_2} \dots x_{ki}^{i_k}\right) \geq 1 \end{cases} \quad (19)$$

4. APPLICATION TO PREDICTION OF SCHOOL DROPOUT AND ACADEMIC PERFORMANCE

In order to analyze the influence of economic, educational, and social dimensions on academic performance and dropout rates, we constructed a dataset collected through a questionnaire administered to first-year and final-year high school students from Moroccan schools. This questionnaire comprised 12 questions designed to describe various aspects of the student. Its design was based on the work proposed by Tiinto [23] and Hiippala [24], enhanced by the integration of several variables that take into account the context and realities of learners in Morocco. These variables are explained in Table 1.

The data obtained are extracted through precise answers to questions posed to the students interviewed, reflecting the diversity of their characteristics and perceptions of school, family, programs, the teaching and learning process and the external environment, and their economic, social and cultural status. The evaluation of non-discreet factors was carried out using a five-dimensional scale that incorporates students' scores for importance, satisfaction, agreement, and collaboration. Higher values indicate a greater influence of these variables on school dropout. The variables obtained can be classified into two categories: Input variables and the output variable.

Input variables: using a thorough and detailed analysis of the literature on school dropout, while respecting the students' personal and private lives, we selected 12 relevant independent predictive variables: Student age (x1) is often between 16 and 22 years. Student gender (x2) is represented by a binary variable, coded 1 for "a boy" and 0 for "a girl." The student's family's socioeconomic status (x3), the household's cultural environment (x4) and the level of parental support (x5) were evaluated using a five-level scale ranging from 1 to 5. Educational aspirations (x6) were assessed through one question in which learners were asked to indicate their academic goals. This feature is binary, taking the value 1 when the qualification exceeds a bachelor's degree level, and 0 otherwise. The motivation for choosing the institution (x7) is also modelled as a binary variable (1 when this choice is determined by the family, 0 for other choices). In addition, peer interaction intensity (x8), expectations related to academic challenges (x9), and the quality of relationships between students and teachers (x10) were assessed using a five-level ordinal scale from 1 to 5. Similarly, impact the financial condition of the 5 household on school dropout (x11) was measured using a five-level scale between 1 and 5, where 1 indicates no impact and 5 represents a very strong impact. Finally, the parents' educational attainment (x12) is assigned the value 1 if the student's mother or father has a qualification higher than a two-year post-secondary diploma, and 0 otherwise.

Output variables: based on the answers to two questions at the end of the questionnaire, we can identify two types of output variables:

- Output variable for assessing school dropout:

We have only two categories: class 0 (students who left school before completion) and class 1 (students who have stayed enrolled in school). Therefore, the dependent variable takes one of these two values,

0 or 1. The value 0 means that the student has left school, and the value 1 means the opposite. To determine the corresponding binary value, we ask a very specific question in the questionnaire. From the student's response, we can know whether the student will drop out or not. The proposed prediction model:

- Output variable for measuring academic performance:

This is the average obtained by the student during the first session of the current year, between 0 and 20. We thus obtain two datasets, stored in two separate Excel files: one dedicated to school dropout and the other to academic performance. Examples of the structure of these datasets are listed in Tables 2 and 3.

Table 1. Explanatory variables examined in the questionnaire, along with their descriptions

Code	Variable name	Description
x ₁	The age of the student.	Student's age in years
x ₂	The gender of the student.	Student's gender: boy or girl
x ₃	The social situation of the student's family.	A child's social stability is measured using one of the following three indicators: death of a parent (0), divorce/separation (1) and stable status (2)
x ₄	The community culture.	Community culture is measured using four main dimensions: a sense of belonging to the school, engagement in school activities, solidarity among students, and respect for classroom norms. These dimensions are operationalized through specific and structured questions.
X ₅	The level of encouragement and support provided by the family.	The degree of family support and encouragement can be assessed through several indicators, including assistance with homework, regular monitoring of schooling, encouragement of success and parental participation in school activities.
X ₆	The degree of educational aspirations.	The degree of educational aspirations can be assessed through several indicators, including the level of studies envisaged, the clarity of the academic and professional project, academic ambition and the commitment to pursue higher education.
X ₇	The main factor that motivated the choice of the establishment.	The main reason for choosing the institution can be assessed through several explanatory factors, including the perceived quality of teaching, the academic reputation of the institution, its geographical accessibility, as well as recommendations or the influence of the family environment.
X ₈	The social interaction with friends	The level of interaction with other students can be assessed by the frequency of exchanges with peers, participation in group work, collaboration on school projects, and involvement in social or extracurricular activities.
X ₉	The anticipations regarding school and academic difficulties.	Expectations regarding school and academic difficulties can be assessed through several indicators, including expected forms of support, anticipated problem-solving strategies, perceived degree of self-efficacy, and prospects for improving academic performance.
X ₁₀	The expectations regarding academic and school problems,	Students' expectations regarding school and academic issues can be measured through several indicators, including the type of support expected from teachers and parents, the strategies considered for overcoming difficulties, confidence in their ability to resolve obstacles, and the academic results they hope to achieve.
X ₁₁	The financial condition of the household.	The household's financial status can be assessed through indicators such as monthly or yearly income, the occupations of family members, as well as housing conditions and owned assets.
X ₁₂	The parents' level of education.	The parents' level of education can be measured through several indicators, including the highest diploma obtained by the father and mother, the total number of years of schooling and the nature of the training undertaken (academic or vocational).

Table 2. The structure of the dataset used for school dropout

Variable 1	Variable 2	...	Variable 12	y takes the value 0 in case of abort and 1 otherwise
------------	------------	-----	-------------	--

Table 3. The structure of the dataset used for academic performance

Variable 1	Variable 2	...	Variable 12	The average obtained by the student during the first session, between 0 and 20
------------	------------	-----	-------------	--

The dataset obtained using the questionnaire reveals a significant imbalance between classes, with the number of students experiencing school dropout was substantially lower than the number of students who stayed in school. To address this, we employ the SMOTE oversampling technique to generate new synthetic observations of the minority class from its closest neighbours, thus rebalancing the data distribution without direct duplication.

To justify our choice of limiting ourselves to 12 factors to study the chances of leaving school prematurely, we begin with a statistical analysis of the variables to identify the factors most influential on this risk. Based on a training set of 200 samples selected from the obtained dataset, we calculate the correlation matrix to measure the relationship between each independent variable x_i (factors explained in Table 1) and the

dependent variable y . The correlation criterion assesses the degree of association between each predictor variable and the dependent variable y . Correlation coefficients approaching +1 or -1 indicate strong positive or negative relationships, respectively. Variables with the highest absolute correlation values are regarded as the most significant factors influencing school dropout. The corresponding results are reported in Table 4.

Table 4. Variable correlation matrix

Variable	Description	Correlation with dropout (Y)
x1	Age	Student's age
x2	Gender	Student's gender
x3	Family social situation	Family stability status
x4	Community culture	School belonging and engagement
x5	Family support	Level of encouragement from family
x6	Educational aspirations	Academic ambition and goals
x7	Choice of institution	Reason for choosing the school
x8	Social interaction	Interaction with peers
x9	Anticipation of difficulties	Expected academic challenges
x10	Academic expectations	Academic expectations
x11	Financial condition	Household economic status
x12	Parents' education level	Educational level of parents

Examination of the correlation matrix demonstrates that the 12 selected variables are all linked to school dropout risk (y) at different levels, validating their choice as explanatory variables. Some explanatory variables, such as educational aspirations (x_6), family support (x_5), and academic expectations (x_{10}), show very high absolute correlations, indicating a significant influence on school dropout. Furthermore, other variables, such as community culture (x_4), social interactions (x_8), and economic conditions (x_{11} , x_{12}), also contribute significantly to the dropout risk, although their effect is moderate. We note that the correlations of variables x_1 and x_2 are low because they are important in their combined interactions with the other explanatory variables. Thus, considering these 12 variables allows for a comprehensive characterization of the educational, social, familial, and psychological factors of school dropout, ensuring a more complete and robust description of the phenomenon under study.

To predict the student's academic performance, we apply our regression model proposed in (5) using the obtained dataset (the form of the dataset presented in Table 3). We partition the dataset into two parts: a training set representing 80%, the remainder being reserved for the testing phase. The parameter n denotes the degree of the polynomial $P_n(x_1, x_2, \dots, x_k)$. In this experiment, the tests are conducted for two values of this parameter, namely $n = 2$ and $n = 3$. To assess the quality of our regression model, three evaluation metrics were used: MAPE, RMSE, and R^2 , as defined in (20)-(22), respectively.

$$MAPE = \frac{1}{N} \sum_{j=1}^N \left| \frac{y_j - p_j}{y_j} \right| \times 100 \quad (20)$$

$$RMSE = \frac{1}{N} \sqrt{\frac{\sum_{j=1}^N (y_j - p_j)^2}{N}} \quad (21)$$

$$R^2 = \frac{\left[\sum_{j=1}^N (y_j - \bar{y})(p_j - \bar{p}) \right]^2}{\sum_{j=1}^N (y_j - \bar{y})^2 \sum_{j=1}^N (p_j - \bar{p})^2} \quad (22)$$

y_j denotes real values, while p_j represents the outputs obtained using the model for the element j . The notations $\hat{y}$ and $\hat{p}$ refer respectively to the means of the observed and predicted values. The methodology to be followed in this experimental part is organized into three essential phases. The first phase is dedicated to training, during which the model is derived from the training data. The second phase concerns testing, performed on a set of 200 samples. Finally, the last phase consists of evaluating the model's performance using three error measures: the MAPE, the RMSE, and the metric R^2 . We conducted a comparison between our PMR model and several reference approaches, including multiple linear regression (MLR), degree n PMR, the model used by Singh and Alhulail [16], and RBF method used by Olabanjo *et al.* [25]. The performance results are presented in Tables 5 and 6. It is generally accepted that a model performs better when MAPE and RMSE are low, while R^2 is high. The obtained results demonstrate that our regression model outperforms the other tested methods overall. Moreover, its performance varies depending on n the degree of the nonlinear polynomial hyperplane. If n is fixed at 2, the method yields $RMSE = 18\%$, $MAPE = 2\%$, and $R^2 = 0.988$. When $n = 3$, an improvement is observed, with values of $RMSE = 10\%$, $MAPE = 1.9\%$, and $R^2 = 0.996$.

The first two analyses showed a clear direction: as the parameter n increases, the RMSE and MAPE decrease, and the R^2 metric grows. To ensure the robustness of this result, further experiments were conducted by progressively varying the polynomial degree of the first term of our regression model (5), such that n varies from 0 to 5. At each step, the RMSE, MAPE, and R^2 evaluation metrics were recalculated for each value of n . The results obtained are shown in Figures 1-3. We also note that as n increases, the model becomes more predictive.

Table 5. The values obtained for the three indicators for our $n = 2$ degree model are presented in comparison with those of the other methods evaluated

Model	RMSE	MAPE	R^2
MLR	25.19%	36.19%	0.561
MPR	19.15%	11.21%	0.863
Singh and Alhulail [16]	18.12%	12.29%	0.871
RBF-ANN [25]	1.86%	7.92%	0.981
Our model	18%	2%	0.988

Table 6. The values obtained for the three indicators for our $n = 3$ degree model are presented in comparison with those of the other methods evaluated

Model	RMSE	MAPE	R^2
MLR	25.19%	36.19%	0.561
MPR	15.37%	10.15%	0.875
Singh and Alhulail [16]	18.12%	12.29%	0.871
RBF-ANN [25]	1.86%	7.92%	0.981
Our model	10%	1.9%	0.996

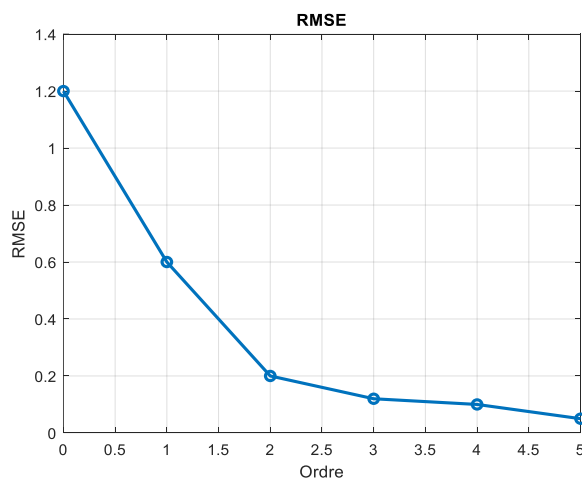


Figure 1. The RMSE value associated with our prediction method varies according to the degree n (from 0 to 5)

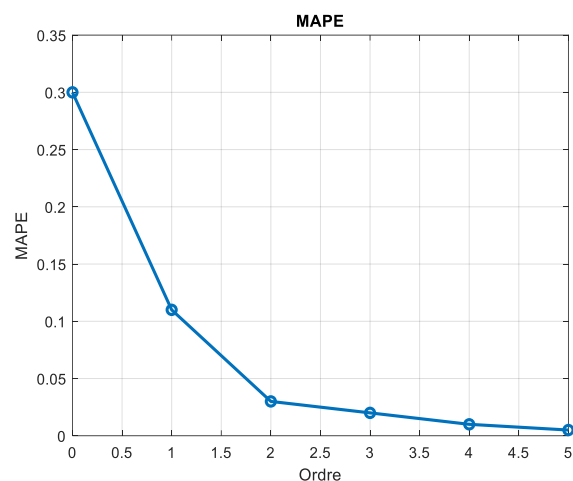


Figure 2. The MAPE value associated with our prediction method varies according to the degree n (from 0 to 5)

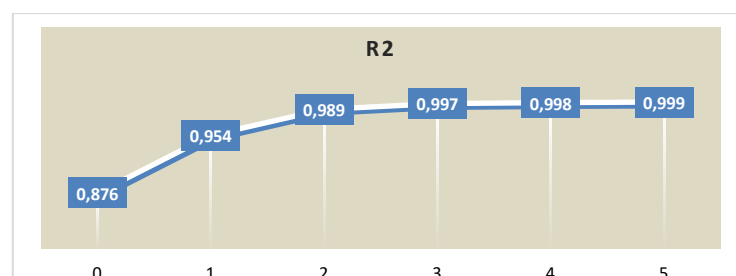


Figure 3. The R^2 value associated with our prediction method varies according to the degree n (from 0 to 5)

Figures 4 and 5 present a comparison of the results calculated by our model and the real values for $n = 2; 5$. Analysis of these graphs emphasizes the strong effectiveness of the proposed model to reliably reproduce levels of academic performance, assessed from the first-semester average. Furthermore, the predicted points are observed to approach the first bisector more closely, particularly when $n = 5$, indicating an improvement in the model's accuracy. The results obtained highlight the remarkable performance of the proposed regression model for prediction tasks in an educational context. They also show that this model represents a unified generalization of several existing regression approaches, which strengthens its relevance and flexibility.

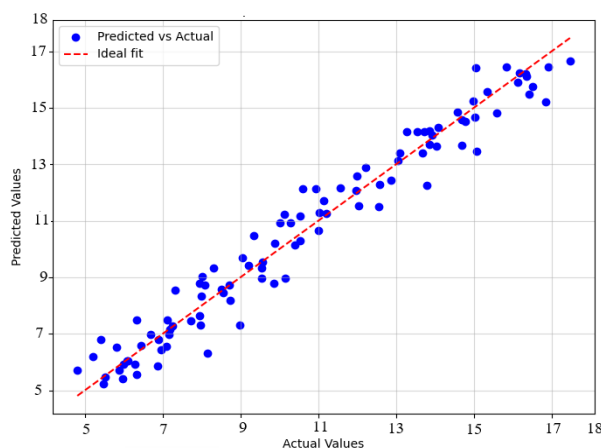


Figure 4. The correlation between real and estimated values of the first half-year average, obtained with the proposed model for $n = 2$

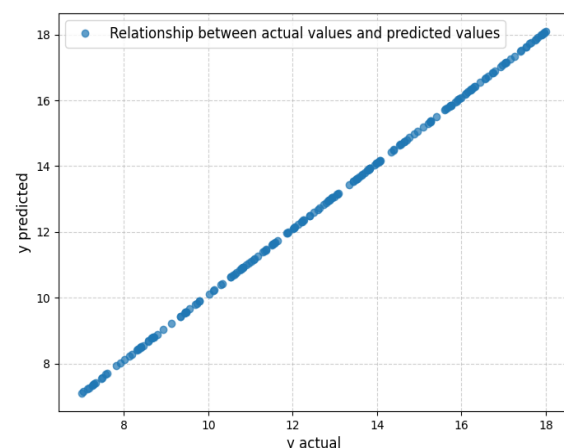


Figure 5. The correlation between real and estimated values of the first half-year average, obtained with the proposed model for $n = 5$

The Table 7 presents parameters of the ABC algorithm for solving the optimization problem (19) in order to estimate the parameters $\vec{\beta} = (\beta_{i_1 i_2 \dots i_k}; i_1 + i_2 + \dots + i_k \leq n)$ of our generalized SVM classifier. The study of school dropout risk prediction involves a classification problem. We apply the proposed classifier introduced in 3.2 using the second dataset (Table 3). This dataset is partitioned into two parts: we allocate 80% of the dataset for training and 20% for performance evaluation. The evaluation of the prediction method relies on measuring accuracy, as presented in formula (23).

$$\text{Accuracy} = \frac{\text{Number of right predictions}}{\text{Number of all predictions}} \times 100\% \quad (23)$$

A comparison was done to evaluate the performance of the proposed classifier, giving $n = 2$ as the degree of the nonlinear hyperplane h . To this end, our method was compared to several recent and widely used classifiers in the literature, including logistic regression (LR), the linear SVM classifier, and neural networks of the RBF and MLP types. This comparison aims to assess generalization ability and the prediction efficiency of the proposed classification model in relation to baseline methods. It also highlights any potential performance gains offered by our approach. The detailed results of this comparison are summarized in Table 8.

To evaluate the impact of degree on prediction performance and the influence of the optimization method used on classifier accuracy, we compared our generalized SVM based on the ABC optimization algorithm (ABC-SVM) with two other SVM classifiers: the first, which employs a genetic algorithm (GA-SVM) and the second on the PSO optimization method (PSO-SVM). The comparative study was conducted under the same conditions and with the same parameter values. The evaluation was performed according to the classification accuracy criterion. The experiments were carried out for different degrees of the polynomial hyperplane $n \in \{1, 2, 3, 4, 5\}$. The results obtained are illustrated in the graphs presented in Figure 6.

The results presented in Figure 6 show that the ABC-SVM approach outperforms the PSO-SVM and GA-SVM methods for all polynomial degrees considered. The superiority of our approach stems from the ABC optimization algorithm's ability to maintain a good balance between global exploration and local exploitation, thus avoiding local minima and improving the quality of the solutions. We also observe that accuracy increases

with the degree of the HYP_n hyperplane polynomial, demonstrating a better capacity to model the nonlinear polynomial relationship. However, the improvement becomes marginal at higher degrees, suggesting a trade-off between model complexity and predictive quality.

We performed a statistical comparison of our ABC-SVM method with the other methods PSO-SVM and GA-SVM using the Wilcoxon test. This test was performed on the accuracy values obtained with $n = 3$. It should be noted that the best method is the one whose performance is significantly superior to the others according to the Wilcoxon signed-rank test, i.e., when the p-value is below the significance level ($\alpha = 0.05$). The results obtained are presented in Table 9. The results shown demonstrate the superiority of our ABC-SVM classifier.

Table 7. Parameters of the ABC algorithm for solving the optimization problem (21) in order to estimate the parameters $\vec{\beta} = (\beta_{i_1 i_2 \dots i_k}; i_1 + i_2 + \dots + i_k \leq n)$ of our generalized SVM classifier

Category	Parameter	Value	Description
Problem	Optimization type	Constrained maximization	Margin-based criterion
	Constraints	presented in the second line of problem (21)	Ensures class separability
Dimension	Variables	$k = 12$ and n	Number of variables and polynomial degree
	Number of coefficients $\vec{\beta}$	$D = \binom{n+k}{n}$	Total number of parameters to optimize
Population	Food sources (SN)	50	Number of candidate solutions
	Employed bees	50	Local exploration
Exploration	Onlooker bees	50	Selection based on fitness
	Scout bees	Dynamic	Replace stagnated solutions
Iterations	Limit parameter (limit)	120	Max trials without improvement
Initialization	MaxCycle	1000	Total number of iterations
Optimization	Solution generation	Uniform within bounds	Ensures initial diversity
	Optimized vector	$\vec{\beta}$	Polynomial coefficients only

Table 8. The results of comparing our classifier with the other methods tested using the classification metric “accuracy”

Classifier	LR	SVM	RBF	MLP	Our model with $n = 2$
Accuracy	80.44%	79.25%	84.26%	90.24%	98.27%

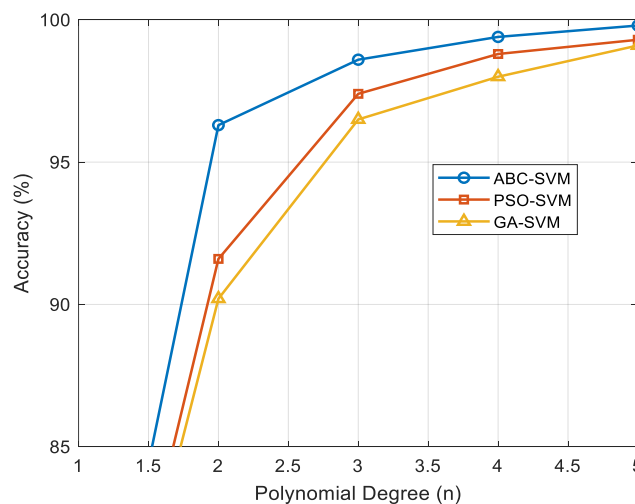


Figure 6. Comparison of accuracy for ABC-SVM, PSO-SVM, and GA-SVM with respect to polynomial degree (n)

To analyze the efficiency of our optimization approach, a convergence analysis of the ABC optimization algorithm was conducted by tracking the evolution of the best-fit value over iterations. The algorithm was initialized with a population of 50 food sources and run for up to 1,000 iterations, with a

limit parameter set at 120. For each iteration, the most favorable objective value obtained was recorded to analyze the convergence process. The resulting convergence curve (Fitness as a function of the number of iterations) shown in Figure 7 demonstrates a rapid decrease in the fit value during the initial stages, followed by a gradual stabilization as the algorithm converges to the optimal solution. This process demonstrates the ABC optimization algorithm's ability to efficiently explore the search space and exploit promising regions, leading to a stable and high-quality estimate of the parameter vector $\vec{\beta}$.

Table 9. Results of the Wilcoxon signed-rank test

Comparison	p-value	Significance ($\alpha = 0.05$)	The best method
ABC-SVM vs PSO-SVM	0.031	Significant	ABC-SVM
ABC-SVM vs GA-SVM	0.016	Significant	ABC-SVM

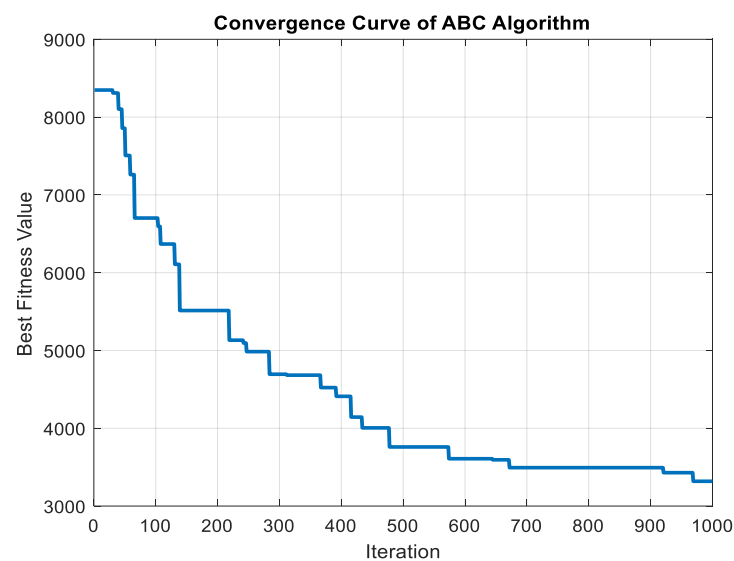


Figure 7. Convergence curve of ABC optimization algorithm

3. CONCLUSION

In this work, we focused on studying academic outcomes of students in Moroccan high schools, as well as analyzing and predicting their level. To this end, we proposed a new regression method, a generalization of polynomial and multiple exponential regression models, which allows for a better capture of the complex relationships between the various explanatory variables and academic performance.

Furthermore, to predict the risk of student dropout, we developed a classifier, derived from an extension of the SVM method by replacing the linear hyperplane with a nonlinear one. The generalized approach aims to refine the accuracy of dropout risk estimation and the ability to accurately distinguish between different classes. The data used in this study come from a questionnaire developed and disseminated to learners enrolled in Moroccan public high schools. This questionnaire provided a rich dataset, incorporating several economic, social, and educational factors essential for analyzing the phenomena under study.

To assess the relevance and efficacy of our approaches, several experiments were conducted using appropriate metrics. These were complemented by comparative studies with various recent reference methods, performing several experimental tests in regression and classification. The results clearly demonstrate the robustness and performance of the proposed nonlinear approaches, which are distinguished by their accuracy and generalizability. Finally, the conclusions of this study highlight the strong potential of generalized models and methods in the field of education. They offer promising prospects for the development and design of prediction and decision systems applied in educational fields, particularly in terms of academic performance and combating school dropout.

ACKNOWLEDGMENT

We thank Sidi Mohamed Ben Abdellah University for its support of this study.

FUNDING INFORMATION

The authors state that no funding was involved.

DATA AVAILABILITY

The data generated and analyzed in this study are real data, collected from 1000 Moroccan high school students using a questionnaire, and are available from the corresponding author.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CrediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hicham EL Yousfi Alaoui	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Ziad Bousraraf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Amal Hjouji	✓		✓	✓			✓			✓	✓		✓	✓
Omar EL Ogri	✓		✓									✓		
Jaouad EL-Mekkaoui					✓		✓			✓		✓	✓	✓
Zaynab Hjouji	✓			✓	✓					✓	✓	✓		

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest

DATA AVAILABILITY

Data availability does not apply to this paper as no new data were created or analyzed in this study.

REFERENCES

- [1] R. B. Darlington, "Multiple regression in psychological research and practice," *Psychological Bulletin*, vol. 69, no. 3, pp. 161–182, 1968, doi: 10.1037/h0025471.
- [2] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "Mathematical programming for piecewise linear regression analysis," *Expert Systems with Applications*, vol. 44, pp. 156–167, Feb. 2016, doi: 10.1016/j.eswa.2015.08.034.
- [3] B. S. Kweon, C. D. Ellis, J. Lee, and K. Jacobs, "The link between school environments and student academic performance," *Urban Forestry and Urban Greening*, vol. 23, pp. 35–43, Apr. 2017, doi: 10.1016/j.ufug.2017.02.002.
- [4] G. K. Uyanık and N. Güler, "A study on multiple linear regression analysis," *Procedia - Social and Behavioral Sciences*, vol. 106, pp. 234–240, Dec. 2013, doi: 10.1016/j.sbspro.2013.12.027.
- [5] F. H. M. Salleh, S. Zainudin, and S. M. Arif, "Multiple linear regression for reconstruction of gene regulatory networks in solving cascade error problems," *Advances in Bioinformatics*, vol. 2017, pp. 1–14, Jan. 2017, doi: 10.1155/2017/4827171.
- [6] J. D. Jobson, "Multiple Linear Regression," 1991, pp. 219–398, doi: 10.1007/978-1-4612-0955-3_4.
- [7] S. H. Hasan and N. H. Irhaif, "Use of regression analysis methods to determine factors related to school dropout (An applied study of a sample of primary schools in baghdad)," *Advances in Mathematics: Scientific Journal*, vol. 10, no. 3, pp. 1285–1299, Mar. 2021, doi: 10.37418/amsj.10.3.16.
- [8] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 59, no. 1, pp. 3–54, Jan. 1997, doi: 10.1111/1467-9868.00054.
- [9] X. Li, J. Ghosh, and G. Villarini, "A comparison of bayesian multivariate versus univariate normal regression models for prediction," *The American Statistician*, vol. 77, no. 3, pp. 304–312, Jul. 2023, doi: 10.1080/00031305.2022.2087735.
- [10] P. F. Smith, S. Ganesh, and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience," *Journal of Neuroscience Methods*, vol. 220, no. 1, pp. 85–91, Oct. 2013, doi: 10.1016/j.jneumeth.2013.08.024.
- [11] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys, "Tail index estimation and an exponential regression model," *Extremes*, vol. 2, no. 2, pp. 177–200, Jun. 1999, doi: 10.1023/a:1009975020370.
- [12] F. Pratavia et al., "The exponentiated power exponential regression model with different regression structures: application in nursing data," *Journal of Applied Statistics*, vol. 46, no. pp.10: 1792-1821, 2019.
- [13] H. Kim et al., "Variable three-term conjugate gradient method for training artificial neural networks," *Neural Networks*, vol. 159, pp. 125–136, Feb. 2023, doi: 10.1016/j.neunet.2022.12.001.

- [14] A. Teso-Fz-Betoño, E. Zulueta, M. Cabezas-Olivenza, U. Fernandez-Gamiz, and C. Botana-M-Ibarreta, "Modification of learning ratio and drop-out for stochastic gradient descendant algorithm," *Mathematics*, vol. 11, no. 5, p. 1183, Feb. 2023, doi: 10.3390/math11051183.
- [15] D. Karaboga and B. Basturk, "Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems," in *Foundations of Fuzzy Logic and Soft Computing*, vol. 4529 LNAI, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 789–798. doi: 10.1007/978-3-540-72950-1_77.
- [16] H. P. Singh and H. N. Alhulail, "Predicting student-teachers dropout risk and early identification: a four-step logistic regression approach," *IEEE Access*, vol. 10, pp. 6470–6482, 2022, doi: 10.1109/ACCESS.2022.3141992.
- [17] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Computers and Education*, vol. 53, no. 3, pp. 563–574, Nov. 2009, doi: 10.1016/j.compedu.2009.03.013.
- [18] M. Alban and D. Mauricio, "Predicting University dropout through data mining: a systematic literature," *Indian Journal of Science and Technology*, vol. 12, no. 4, pp. 1–12, Jan. 2019, doi: 10.17485/ijst/2019/v12i4/139729.
- [19] S. F. Costa and M. M. Diniz, "Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course," *Education and Information Technologies*, vol. 27, no. 9, pp. 12381–12397, Nov. 2022, doi: 10.1007/s10639-022-11117-1.
- [20] B. R. C. Chacha, W. L. G. López, V. X. V. Guerrero, and W. G. Villacis, "Student dropout model based on logistic regression," in *Research in Higher Education*, vol. 59, no. 3, 2020, pp. 321–333. doi: 10.1007/978-3-030-42520-3_26.
- [21] C. Mason, J. Twomey, D. Wright, and L. Whitman, "Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression," *Research in Higher Education*, vol. 59, no. 3, pp. 382–400, May 2018, doi: 10.1007/s11162-017-9473-z.
- [22] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, "Predicting student dropout in subscription-based online learning environments: the beneficial impact of the logit leaf model," *Decision Support Systems*, vol. 135, p. 113325, Aug. 2020, doi: 10.1016/j.dss.2020.113325.
- [23] V. Tinto, "Dropout from higher education: a theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, Mar. 1975, doi: 10.3102/00346543045001089.
- [24] T. Hiippala, "Aspects of multimodality in higher education monographs," in *Studies in Writing*, vol. 33, BRILL, 2016, pp. 53–78. doi: 10.1163/9789004312067_005.
- [25] O. A. Olabanjo, A. S. Wusu, and M. Manuel, "A machine learning prediction of academic performance of secondary school students using radial basis function neural network," *Trends in Neuroscience and Education*, vol. 29, p. 100190, Dec. 2022, doi: 10.1016/j.tine.2022.100190.

BIOGRAPHIES OF AUTHORS



Hicham EL Yousfi Alaoui    received obtained his master's degree in applied mathematics and statistics in 2014 from the Faculty of Sciences of Fez, Sidi Mohammed Ben Abdel lah University, currently enrolled is preparing a doctorate in applied mathematics at the Fez Higher School of Technology, Engineering Science and Management Laboratory. He can be contacted at email: alaoui.hicham77@yahoo.fr.



Ziad Bousraraf    he obtained his master's degree in statistics and management in 2018 and is currently pursuing a Ph.D. in Management and Statistics at Sidi Mohamed Ben Abdellah University, Morocco. His research interests include statistics, economics and management, and artificial intelligence. He can be contacted at email: jawad.mekkaou@gmail.com.



Prof. Amal Hjouji    received his B.Sc. and M.Sc. degrees in computer science from Sidi Mohamed Ben Abdellah University, in 2014 and 2016, respectively. She is currently pursuing a Ph.D. in computer science at Sultan Moulay Slimane University, Morocco. His research interests include pattern recognition, artificial intelligence, data science, and big data. She can be contacted at email: amal.hjouji@usmba.ac.ma.



Prof. Omar EL Ogrri    he obtained a B.Eng. in Mathematical and Computer Sciences and an M.S. in Engineering Science from Sidi Mohamed Ben Abdellah University, Fez, Morocco. He is currently a Ph.D. candidate in the Department of Physics at the same university. His research interests include signal security, artificial intelligence, and applied mathematics. He can be contacted at email: omar.elogri@gmail.com.



Prof. Jaouad El-Mekkaoui    he obtained his bachelor's, master's, and Ph.D. degrees in Mathematics from Sidi Mohamed Ben Abdellah University, Morocco, in 2003, 2006, and 2014, respectively. He is currently a professor in the Department of Computer Science at the Higher School of Technology of Fez, Sidi Mohamed Ben Abdellah University, Morocco. His research interests cover applied mathematics, artificial intelligence and data science. He can be contacted at email: jaouad.elemekkaoui@usmba.ac.ma.



Zaynab Hjouji    Ph.D. in Economics and management, professor at private university of Marrakech, Morocco. Affiliated with the Department of Economics and Management of Organizations (DEMO) and the Interdisciplinary Research Laboratory in Economics, Finance, and Management of Organizations (LIREFIMO). She can be contacted at email: amal.hjouji@usmba.ac.ma.