STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, May 2026, pp 3401–3415.
Published online in International Academic Press (www.IAPress.org)

New Regression Model for Academic Achievement and New Classification Method for School Dropout Based on Artificial Bee Colony Algorithm

Hicham EL Yousfi Alaoui [1], Ziad Bousraraf [2], Amal Hjouji [3], Omar EL Ogri [1]
Jaouad EL-Mekkaoui [1,*]

[1]*SIG Laboratory, Sidi Mohamed Ben Abdellah University, Fes, Morocco*
[2]*Sidi Mohamed Ben Abdellah University, Fez, Morocco*
[3]*LISAC Laboratory, Sidi Mohamed Ben Abdellah University, Fez, Morocco*

Abstract This article focuses on the analysis of two major phenomena in the field of education: academic achievement and student dropout. Academic achievement corresponds to the academic performance of students, generally assessed through their grades, averages or the achievement of educational objectives. It is influenced by various factors such as personal abilities, motivation, family support and the quality of education. On the other hand, school dropout refers to the premature abandonment of studies, often caused by academic, social or economic difficulties. These two phenomena are among the greatest challenges facing educational institutions in most countries of the world, especially in developing countries. They have serious social and economic consequences for individuals and societies. To analyze the risks resulting from these two phenomena, it is necessary to use advanced forecasting techniques and methods, including statistical methods and artificial intelligence algorithms using available data. These methods allow us to understand the factors of each phenomenon individually and predict its negative risks. In order to improve the quality of predictions, we propose in this article a new regression model based on the multiple exponential regression model and the polynomial regression model. In order to identify the impact of social, economic and personal factors of the student and his environment on school dropout, we present an innovative classification method based on a generalization of the logistic regression model, replacing the linear term with a multiple polynomial term. To estimate the coefficients of the two proposed models, we use the ABC (Artificial Bee Colony) optimization algorithm. The two proposed approaches were applied to two different databases: the regression model was used to predict academic achievement and the classification method was used to predict the risks of school dropout. We carry out comparative studies with recent methods. The results obtained showed the reliability and superiority of the proposed approaches in terms of prediction and accuracy.

Keywords Academic achievement, Student dropout, Regression model, Classification method, Artificial bee colonies (ABC)

DOI: 10.19139/soic-2310-5070-2420

1. Introduction

*T*he analysis of two major phenomena in the field of education, academic achievement and school dropout, reveals essential issues. Academic achievement, measured by academic performance such as grades, averages or the achievement of educational objectives, is influenced by factors such as personal abilities, motivation, family support and the quality of teaching. At the same time, school dropout, which refers to premature abandonment of studies, is often linked to academic, social or economic difficulties.These two phenomena are interdependent: insufficient academic results significantly increase the risk of dropping out. Their impact is particularly pronounced in developing countries, where they amplify social and economic inequalities, limiting

*Correspondence to: Jaouad ELMekkaoui BP 1796 Fez-Atlas 30003, Fez, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

access to future opportunities. Understanding this complex relationship is essential to develop effective strategies to improve academic performance and reduce dropout rates. This requires innovative solutions, such as data analysis and the use of predictive models, but also educational policies that strengthen equity and inclusion, while mobilizing families and communities. These approaches offer a crucial opportunity to build education systems that can respond to current challenges and promote a more equitable society. To anticipate the risks associated with these phenomena, it is essential to use advanced prediction methods that exploit available data, in particular regression and classification techniques.

Regression is a statistical and machine learning technique used to analyze the relationship between a dependent variable and one or more independent variables. Its main objective is to model how changes in the dependent variable are associated with changes in the independent variables. This allows predicting or estimating a continuous variable based on other explanatory variables.

Regression can be linear or nonlinear, depending on the form of the database. It is widely used in various fields to analyze data, make predictions, and understand the relationships between different phenomena. Regression, a versatile technique, finds applications in various fields, including economics, health, social and natural sciences, marketing, engineering, among others. It aims to model and analyze relationships between variables, predict outcomes, and provide valuable information for decision-making based on existing data.

A regression model, in general, is constructed to explain (or predict, depending on the analytical perspective) the behaviour of a phenomenon (dependent variable) using a combination of explanatory factors (independent variables). The aim is to understand, explain, or forecast the behaviour of a dependent variable based on one or more independent variables, also known as explanatory factors or predictors. The objective varies depending on the analysis or study being conducted. Broadly, building a regression model involves identifying and quantifying the relationship between the dependent variable and the independent variables [1]. This relationship can be expressed as follows:

$$y = f(x_1, x_2, ..., x_k) + \epsilon \tag{1}$$

Where y is the dependent variable or the variable we aim to predict or explain, f is the regression function, x_1, x_2, . . . , x_k are the independent variables or predictors and ϵ represents the error term, the difference between the predicted and actual values of x.

Regression models find application in various fields such as social sciences, economics, natural sciences, medicine, engineering, and more. They serve to comprehend relationships between different variables and enable informed decision-making based on these relationships. The choice of regression model is closely linked to the function f, which we seek to represent and the nature of the data available. This relationship determines the type of regression model to use to obtain an accurate and meaningful representation of the data.

In the case of simple linear regression [2, 3], the model can be broken down more precisely as follows:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

It's used when the relationship between variables is linear. For example, if one variable is expected to increase or decrease in proportion to another. Where β_0 is the intercept, representing the value of x when x is zero and β_1 is the slope coefficient, indicating the change in y for a unit change in x. The primary goal of simple linear regression is to estimate the parameters β_0 and β_1 such that the linear equation best fits the observed data points. This estimation is commonly performed using the least squares method to minimize the sum of the squared differences between the observed and predicted values of y. Once the model parameters are estimated, they can be used to predict the value of the dependent variable y for new values of the independent variable x.

Linear regression is called multiple [4, 5, 6, 7] when the model is composed of k variables with $k \geq 2$. This model can be expressed as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k + \epsilon \tag{3}$$

Where β_0 is the intercept, representing the value of Y when all independent variables are zero and β_1, β_2, . . . , β_k are coefficients representing the change in x for a one-unit change in each independent variable while keeping other variables constant. Multiple linear regression is utilized when several factors can simultaneously influence the dependent variable. Similarly to simple linear regression, the least squares method is often used in multiple linear

regression to estimate the coefficients β_0,...,β_k to minimize the sum of squared differences between the predicted and actual values of the dependent variable.
Polynomial regression [8] is an extension of linear regression that allows modelling non-linear relationships between variables. Unlike linear regression, which assumes linear relationships, polynomial regression uses polynomials to fit more complex curves. Polynomial regression may be applied to a single regressor variable called Simple Polynomial Regression, that it is represented by an equation of the form:

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + ...\beta_n x^n + \epsilon \tag{4}$$

Where β_0, β_1, ... , β_n are coefficients representing the effect of each degree term on the dependent variable. When polynomial regression is applied to multiple predictor variables [8, 9], it's often referred to as "Multiple Polynomial Regression." This method allows for the utilization of polynomial functions with multiple independent variables to model complex relationships with the dependent variable.

$$y = P_n(x_1, x_2, ..., x_k) + \epsilon \tag{5}$$

Where P_n is a polynomial of degree less than or equal to n and of k variables defined as follows

$$P_n(x_1, x_2, ..., x_k) = \sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} x_1^{i_1} x_2^{i_2} ...x_k^{i_k} \tag{6}$$

The multiple exponential regression model [10, 11] is an extension of the classical exponential model, allowing to model a relationship between a dependent variable and several independent variables. It is useful for analyzing data where the effect of several factors explains an exponential behavior of the target variable. the general form of this model is:

$$y = a.e^{b_1.x_1+b_2.x_2+.....+b_k.x_k} + \epsilon \tag{7}$$

where a is initial constant (value of y when all x_i = 0) and b_1, b_2, ... , b_k are coefficients associated with the independent variables x_1, x_2,..., x_k.
To increase the prediction quality, we propose, in equation (8), a new regression model combining the multiple polynomial regression model and the exponential regression model.

$$y = \sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} x_1^{i_1} x_2^{i_2} ...x_k^{i_k} + a.e^{b_1.x_1+b_2.x_2+.....+b_k.x_k} + \epsilon \tag{8}$$

Traditional optimization methods used to estimate the parameters of a regression model can lead to overfitting, reducing the ability of the model to generalize to new data. Furthermore, the intensive computation and limited reliability of extrapolation, combined with the challenges of estimating coefficients for high degrees, make these methods less efficient and more resource-intensive. To address this problem, we use a novel parameter estimation technique based on the Artificial Bee Colony (ABC) optimization algorithm.
To study the phenomenon of school dropout, the majority of researchers have used the classification method based on logistic regression [12, 13, 14, 15, 16, 17, 18] based on the following logistic function:

$$f(x_1, x_2, ..., x_k) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+....\beta_k x_k)}} \tag{9}$$

Because of its effectiveness, we adopted it in this article while proposing a generalization. This generalization consists in replacing the linear term $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ by a polynomial term P_n (x_1, x_2, ... , x_k), where P_n (x_1, x_2, ... , x_k) = $\sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} x_1^{i_1} x_2^{i_2} ...x_k^{i_k}$ is a polynomial multiple of degree n. Therefore, our classification method based on the following logistic function:

$$f(x_1, x_2, ..., x_k) = \frac{1}{1 + e^{-P_n(x_1,x_2,...,x_k)}} \tag{10}$$

To optimize the coefficients of our classification model, we use the ABC (Artificial Bee Colony) optimization algorithm.
These approaches were implemented on two educational databases: the proposed regression model was used to predict academic achievement, while the classification model made it possible to identify the risks of dropping out of school.

2. The proposed regression model

Considering the effectiveness of multiple polynomial regression and multiple exponential regression, we propose a novel regression model that merges these two approaches with the aim of improving the accuracy of predictions: Our general regression model is presented by equation (8) as follows:

$$y = P_n(x_1, x_2, ..., x_k) + a.e^{b_1.x_1+b_2.x_2+.....+b_k.x_k} + \epsilon \tag{11}$$

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

Based on a training set of a database, we can estimate the coefficient vector $\overrightarrow{\beta} = (a, b_1, b_1, ...b_k, \beta_{i_1,i_2,...,i_k}; i_1 + i_2 + ...i_k \leq n)$ by solving the following optimization problem

$$Minimise\ l(\overrightarrow{\beta}) = \|\epsilon\|^2 \tag{12}$$

where

$$\|\epsilon\|^2 = \sum_{i=1}^{N} (P_n(x_1(i), x_2(i), ...x_k(i)) + a.e^{b_1x_1(i)+b_2x_2(i)+...b_kx_k(i)} - y(i))^2 \tag{13}$$

is the squared error between the actual values $y(i)$ and the predicted values $y_p(i) = P_n(x_1(i), x_2(i), ..., x_k(i)) + a.e^{b_1.x_1(i)+b_2.x_2(i)+.....+b_k.x_k(i)}$, $(x_1(i), x_2(i), ...x_k(i))$ is the vector of characteristics of individual i; $i = 1, ..., N$ and N is the size of the training set of the database.

To solve the problem (12), we tested several approaches, including the gradient method, the Newton-Raphson algorithm, and some meta-heuristic methods. Among them, we found that the artificial bee colony algorithm (ABC) is the most efficient method that can be used.

The ABC (Artificial Bee Colony) method is a meta-heuristic algorithm inspired by the behavior of bee colonies in search of food. This algorithm is particularly useful for solving optimization problems, especially those with a large search space or complex nonlinear functions. Here are the steps to estimate the coefficient vector $\overrightarrow{\beta} = (a, b_1, b_1, ...b_k, \beta_{i_1,i_2,...,i_k}; i_1 + i_2 + ...i_k \leq n)$ using the following ABC algorithm:

1. ***Initialization***
   Before starting the iterations of the algorithm, an initial population of solutions (food sources) is generated. This population is random, with each solution representing a possible configuration of the optimization problem. The initial steps are as follows:
   - ***Initialization of bee positions****: An initial population of solutions (usually denoted, $\overrightarrow{\beta}_i$, $(i = 1, 2....N)$ is generated. Each $\overrightarrow{\beta}_i$, is a potential solution to the problem. This population is distributed across the search space.*
   - ***Calculation of the quality of the solutions****: Each solution is evaluated using the objective function $l(\overrightarrow{\beta})$ which determines the quality of each solution $\overrightarrow{\beta}_i$.*
2. ***Employee Bees Phase***
   Employee bees are responsible for exploring the search space from the current solutions. They generate new solutions from their current positions and evaluate them. If a new solution is better than the previous one, it replaces it.
   - ***Generating a new solution****: Each employed bee $\overrightarrow{\beta}_i$ generates a new solution $\overrightarrow{v}_i$ in the neighborhood of $\overrightarrow{\beta}_i$. This new solution is calculated by applying a small random variation to the current solution: $\overrightarrow{v}_i = \overrightarrow{\beta}_i + \phi_i.(\overrightarrow{\beta}_i - \overrightarrow{\beta}_k)$ where ϕ_i is a factor chosen randomly in the interval $[-1, 1]$ and $\overrightarrow{\beta}_k$ is another solution chosen randomly in the initial population.*
   - ***Evaluation of the new solution****: The quality of the new solution $\overrightarrow{v}_i$ is calculated using the objective function $l(\overrightarrow{v_i})$.*

- ***Update of the solution****: If $l(\overrightarrow{v_i})$ is better than $l(\overrightarrow{\beta_i})$, the employed bee replaces $\overrightarrow{\beta}_i$ with $(\overrightarrow{v_i})$. Otherwise, it keeps $\overrightarrow{\beta}_i$*

3. ***Observer Bee Phase***
   - *Observer bees do not directly generate new solutions. They search for food sources based on probability, which is proportional to the quality of current solutions. An observer bee chooses a solution with a probability related to its quality:*
   - ***Solution selection****: Observer bees choose a solution $\overrightarrow{\beta}_i$ based on the quality of the solution $l(\overrightarrow{\beta_i})$, The higher the quality, the greater the probability of selecting $\overrightarrow{\beta}_i$.*
   - ***Neighborhood exploration****: Once the solution is selected, the observer bee explores neighboring solutions, applying the same formula as for the employed bees: $\overrightarrow{v}_i = \overrightarrow{\beta}_i + \phi_i.(\overrightarrow{\beta}_i - \overrightarrow{\beta}_k)$*
   - ***Update of the solution****: If $l(\overrightarrow{v_i})$ is better than $l(\overrightarrow{\beta_i})$, the employed bee replaces $\overrightarrow{\beta}_i$ with $\overrightarrow{v}_i$ Otherwise, it keeps $\overrightarrow{\beta}_i$.*
4. ***Scout Bee Phase***
   - *Scout bees are responsible for searching for new unexplored regions of the search space. Their role is crucial to prevent the algorithm from getting stuck in a local optimum. If a solution does not improve for a certain number of iterations (i.e., if a bee is in a local optimum), the scout bee abandons this solution and randomly generates a new one.*
   - ***Abandoning stagnant solutions****: If a solution has not improved for a certain number of iterations (a predefined threshold), it is abandoned.*
   - ***Generating new random solutions****: The scout bee generates new random solutions $\overrightarrow{\beta}_i$ in the search space, and evaluates them using the objective function.*
5. ***Stopping Criterion***
   - *The algorithm stops when a stopping criterion is reached. This can be.*
   - *A predefined number of iterations or optimization cycles.*
   - *A negligible improvement in the objective function (the change in the best solution is less than a given threshold).*
   - *A convergence to a stable solution, where the bees no longer find better solutions.*

Regarding the algorithm parameters, we set: the colony size to 100 individuals, the maximum number of iterations to 500, and the abandonment criterion (or "limit") to 50 cycles without improvement. These parameters were chosen after several preliminary tests, allowing good performance in terms of quality, precision, and computation time.

3. The proposed classification method

In this subsection, we present a new classification model based on generalizing the logistic regression classification model. This subsection is divided into two parts: in the first, we briefly recall the principles of the logistic regression classification method; in the second, we introduce our classification model, replacing the linear term $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$ by a multiple polynomial $P_n(x_1, x_2, ..., x_k)$ of degree n, this model extends the capabilities of logistic regression.

3.1. Brief reminder of the logistic regression classification method

The logistic regression model is a statistical and supervised learning tool used to solve classification problems. It is particularly suitable for cases where the target variable is categorical, often binary (e.g., success/failure or yes/no). This model is based on a logistic transformation applied to a linear combination of explanatory variables, allowing to estimate the probability that an observation belongs to a given class. The logistic function, or sigmoid, is defined

by:

$$f(z) = \frac{1}{1+e^{-z}} \tag{14}$$

where z represents the linear combination of predictors weighted by their coefficients, i.e.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p \tag{15}$$

The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by maximizing the likelihood of the observed data, a robust approach that ensures consistent probabilistic predictions.
The likelihood represents the probability of the observed data given the model parameters. In a binary classification problem, the probability predicts that the observation y_i belongs to a class is given by:

$$P(y_i|X_i, \beta) = f(z_i)^{y_i}(1 - f(z_i))^{1-y_i} \tag{16}$$

where $f(z) = \frac{1}{1+e^{-z_i}}$ is the sigmoid function, $z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip}$,$\beta = (\beta_0, \beta_1, ..., \beta_p)$ are the coefficients of the model and $y_i \in \{0, 1\}$ is the target variable (class).
The total likelihood for n observations is given by:

$$L(\beta) = \prod_{i=1}^{n} f(z_i)^{y_i}(1 - f(z_i))^{1-y_i} \tag{17}$$

To simplify the calculations, we often maximize the log-likelihood, which is the logarithmic equivalent of the likelihood:

$$L(\beta) = \sum_{i=1}^{n} [y_i log(f(z_i)) + (1 - y_i) log(1 - f(z_i))] \tag{18}$$

3.2. Our classification model

Similar to logistic regression, we follow the same steps to build our classification model. However, our model relies on a logistic transformation applied to a nonlinear function $z(x_1, x_2, ..., x_k)$ of the explanatory variables, such that:

$$z(x_1, x_2, ..., x_k) = \sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} x_1^{i_1} x_2^{i_2} ...x_k^{i_k} \tag{19}$$

The coefficients $\overrightarrow{\beta} = (\beta_{i_1,i_2,...,i_k}; i_1 + i_2 + ...i_k \leq n)$ must be estimated by minimizing the objective function (18), replacing z_i by $\sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} L_{i_1,i_2,...,i_k}(x_1(i), x_2(i), ..., x_k(i))$, where $(x_1(i), x_2(i), ..., x_k(i))$ represents the feature vector of individual i.
To estimate these coefficients, it is sufficient to solve the new optimization problem using the ABC algorithm:
The ABC algorithm combines global exploration (via scout bees) and local exploitation (via employee and observer bees) to search for the optimal solution in complex spaces. This balancing mechanism between exploration and exploitation allows it to efficiently adapt to various types of optimization problems.
After estimating the parameters $\overrightarrow{\beta}$ from a training set from the database, it becomes possible to classify a new individual, represented by a vector of characteristics $X = (x_1, x_2, \ldots, x_k)$. Using the generalized logistic function $f(z) = \frac{1}{1+e^{-z}}$, where $z = \sum_{i_1+i_2+.....i_k \leq n} \beta_{i_1,i_2,...,i_k} x_1^{i_1} x_2^{i_2} ...x_k^{i_k}$. we determine the class of the individual: If $f(z) > \frac{1}{2}$, the individual is assigned to class 1. If $f(z) \leq \frac{1}{2}$, the individual is assigned to class 0.

4. Application of our regression and classification model to the prediction of student dropout risk

To assess the impact of economic, sociological, and academic factors on academic achievement and school dropout, we began by preparing the data using a questionnaire distributed to a group of first- and second-year baccalaureate

$x_1 = AS$	the age of the student
$x_2 = GS$	the gender of the student
$x_3 = SS$	the social status of the child's family
x_4= CC	the community culture
x_5= SEF	the degree of support and encouragement from the family
x_6= EA	the degree of educational aspirations
x_7= MRBCI	the motivating reason behind the choice of the institution
x_8= IWS	the level of interaction with other students
x_9= ERASP	the expectations regarding academic and school problems
x_{10}= STR	the student-teacher relationship.
x_{11}= ELF	the economic level of the family
x_{12}= ELP	the educational level of the parents

Table 1. The independent variables addressed and mentioned in the questionnaire

students in Moroccan public institutions. This questionnaire included various questions addressing essential aspects of the students' lives. These questions were developed based on the concept proposed by Tinto et al. [19, 20], with the addition of specific variables adapted to the living conditions and particularities of students in Morocco. The main objective of the questionnaire was to identify and extract 12 key variables, presented in Table 1.

Through the questionnaire distributed to a group of secondary school students, first and second year of baccalaureate level, we collected the opinions of the participants on a set of variables according to their answers to the questions asked. These data reflect the attitudes, characteristics, opinions, as well as the social, economic and cultural level of the students. Continuous variables were assessed using a fivelevel ordinal scale, adding up the degrees of importance, satisfaction, agreement or contribution expressed by participants regarding these variables and their impact on school dropout. The higher the values obtained, the more pronounced the expression of the variables.

Independent variables: Based on the literature review and the country privacy, we defined and proposed the predictive variables as follows:

the age of the qualifying secondary school students (AS) ranges from 16 to 22 years, the gender of the student (GS) is a binary variable, 1 for "Boy", and 0 for "Girl". The influence of the social status of the student family (SS), the community culture (CC) and the support and encouragement from the family (SEF) were calculated and measured by a five-point scale, coded from 1 "low" to 5 points "very high". The educational aspirations (EA) degree was presented as a single item, asking broadcasting students to give their aspirations. This variable was presented as a binary variable: 1 for post-bachelor's studies (bachelor's, master's or doctorate) and 0 for any other type of degree. The motivating reason behind the choice of the institution (MRBCI) was measured by a binary variable: 1 for parents and 0 for any other reason. The degree of interaction with other students (IWS), the level of expectations regarding academic and school problems (ERASP) and the level of student-teacher relationship (STR) were measured by a five-point ascending scale, starting from 1 ("low") to 5 ("very high"). The relationship between the economic level of the family (ELF) and school dropout was measured by an increasing scale from 1 to 5, coded 1 (if there is no influence) to 5 (high influence) on school dropout. The educational level of the parents (ELP) was coded 1 if at least one parent has a degree above a bachelor's degree, and 0 if neither parent has one.

Dependent variable: We deal with two separate studies. In the first study, we seek to predict the impact of the extracted factors (the characteristics presented in Table 1) on dropping out or continuing studies (school dropout). In the second study, we analyze the effect of these same factors on academic achievement. To do this, we use the responses to the questionnaire distributed to participating students in order to extract two dependent variables:

First dependent variable: y (binary: 0 or 1), used to assess students' intention to continue (y=0) or drop out (y=1) their studies. To measure this intention, we simplified the five-item scale of Dresel and Grassinger [21] into a binary scale with two options. This adaptation is based on a direct question, for example: "I often think about dropping out of school." Possible answers are "yes" (1) or "no" (0).

Second dependent variable: y (continuous, between 0 and 20), used to assess students' academic achievement.

Students	x_1	x_2	.....	x_{12}	y (0 dropped or y=1 not abandoned)

Table 2. the structure of the first database

Students	x_1	x_2	.....	x_{12}	y: the average obtained during the first session of this year which is between 0 and 20

Table 3. the structure of the second database

In this context, we ask participants a direct question: "Indicate the average obtained during the first session of this year" (knowing that the questionnaire was distributed at the end of the previous year).
After processing the questionnaires and extracting the independent variables $(x_1, x_2, ..., x_{12})$, as well as the dependent variables y for each individual, we structured these data in two separate databases.
First database (Table 2): It contains the measurements of the independent variables $(x_1, x_2, ..., x_{12})$ and the dependent variable y, where y = 0 corresponds to a student who did not drop out, and y = 1 to a student who dropped out.

Second database (Table 3): It also contains the measurements of the independent variables $(x_1, x_2, ..., x_{12})$, but the dependent variable here represents the student's average over two years, namely the first and second years of the baccalaureate.

The first study: prediction of student academic performance: We use the first database to predict the impact of the extracted factors on the student's academic achievement. In this context, we apply our regression model defined in equation (11), $y = P_n(x_1, x_2, ..., x_k) + a.e^{b_1.x_1+b_2.x_2+.....+b_k.x_k} + \epsilon$
using the second database. We divide this database into two sets, 80% for training (for estimating the model coefficients) and 20% for testing (to assess the quality of our prediction model), where n is the degree of the polynomial $P_n(x_1, x_2, ..., x_k)$. In our study we carry out our experiments with 2 values of n = 2 and n = 3. To evaluate our prediction model and the other models tested, we used three evaluation criteria, the mean absolute percentage error (MAPE), the coefficient of determination R^2 and the root mean square error (RMSE):

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

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

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

Where N is the number of observations, y_j are the actual values, and p_j are the predicted values by the model for each observation j. $\bar{y}$ and $\bar{p}$ are the mean of the total real and estimated outputs. Figure 1 shows the three steps carried out to carry out this work, which begins with the learning step or the step of building our regression model based on the training set, the testing step based on the test set formed by 200 samples, and finally the evaluation step aims to evaluate the performance by calculating the three types of errors, MAPE, RMSE and R^2.
In this experiment we performed a comparative study of our polynomial regression approach of degree n, with other very well-known methods, such as multiple linear regression (MLR), multiple polynomial regression (MPR) of degree n, the regression presented by Singh et al. [12], the regression based on the Radial Bias Function Artificial Neural Network (RBF-ANN) method used by Olabanjo et al. [22]. Given the diversity of metaheuristic algorithms capable of solving the formulated optimization problem, we tested several of them and found that the artificial bee

Regression model	$RMSE$	$MAPE(\%)$	$R^2(-)$
MLR	25.19	36.19	0.561
MPR	19.15	11.21	0.863
Singh et al.[13]	18.12	12.29	0.871
RBF-ANN[24]	1.86	7.92	0.981
Our GA-based model	1.73	7.65	0.979
Our PSO-based model	0.94	1.14	0.978
Our ABC-based model	0.19	0.03	0.987

Table 4. the results of the criteria RMSE, MAPE and R^2 for our regression model of degree n = 2 and the other methods tested, RMSE, MAPE and R^2.

colony (ABC) algorithm offers the best performance. To this end, and in order to justify our choice, we conduct in this experiment a comparative study with some of the most well-known metaheuristic algorithms, such as the genetic algorithm (GA) and particle swarm optimization (PSO).The results obtained are represented in the tables 4 and 5.

We indicate that, the model has been better when the values of MAPE and RMSE are small and the value of R^2 is large. We notice that our regression model is very effective compared to the other methods tested. We also notice the quality of our model varies according to degree n of the polynomial. For n = 2 our regression model gives the values RMSE = 0.19, MAPE = 0.03 and R^2 = 0.987 and for n = 3 , we noticed these values were improved as follows RMSE = 0.11, MAPE = 0.02 andR^2 = 0.992. This shows the superiority of our regression model and the optimization method used regardless of n greater than 2.

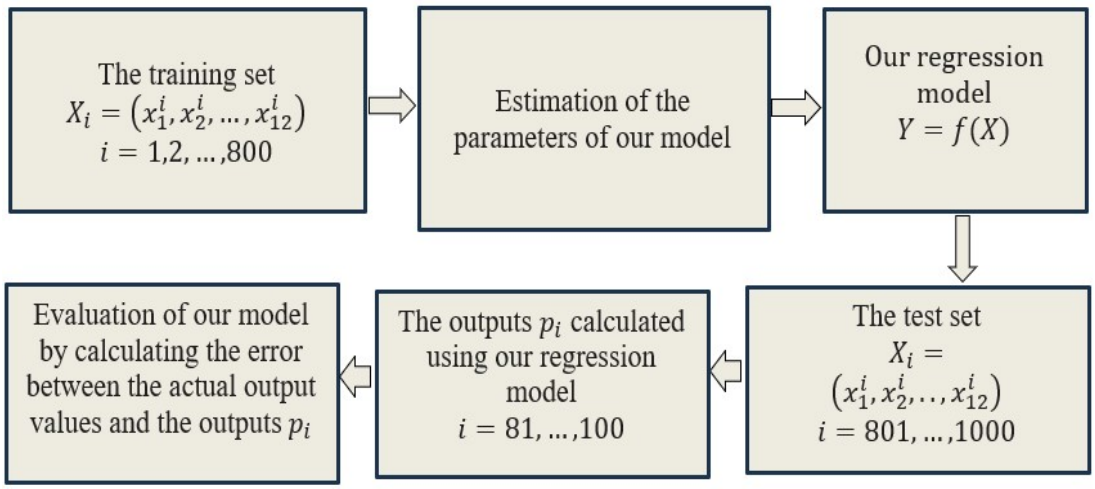

Figure 1. The learning, testing and evaluation processes of our regression model

In order to evaluate the performance of our regression model on a different population, we used another database, from the Portuguese population, called "UCI Student Performance Data [23]. Due to the specific nature of this database, the input variables differ for each individual. It was therefore necessary to adapt the structure of our model according to the type and number of characteristics available. The same comparative analysis was carried out this time on data from the Portuguese population. The results obtained are shown in Tables 6 and 7.

Based on the results presented in these two tables, we can make the same observation: our regression model based on the ABC optimization method proves to be more efficient, on the UCI Student Performance Data database, than the other models tested.

Regression model	$RMSE$	$MAPE(\%)$	R^2(–)
MLR	25.19	36.19	0.561
MPR	15.37	10.15	0.875
Singh et al.[13]	18.12	12.29	0.871
RBF-ANN[24]	1.86	7.92	0.981
Our GA-based model	1.45	7.42	0.977
Our PSO-based model	0.89	1.02	0.988
Our ABC-based model	0.11	0.02	0.992

Table 5. the results of the criteria RMSE, MAPE and R^2 for our regression model of degree n = 3 and the other methods tested, RMSE, MAPE and R^2.

Regression model	$RMSE$	$MAPE(\%)$	R^2(–)
MLR	29.01	40.12	0.659
MPR	22.17	14.10	0.873
Singh et al.[13]	20.58	13.90	0.945
RBF-ANN[24]	3.21	10.11	0.961
Our GA-based model	3.02	9.19	0.970
Our PSO-based model	2.18	1.92	0.978
Our ABC-based model	1.22	0.68	0.982

Table 6. The results obtained according to the RMSE, MAPE and R² criteria for our regression model of degree n = 2, as well as for the other tested methods, were evaluated from the UCI Student Performance Data database.

Regression model	$RMSE$	$MAPE(\%)$	R^2(–)
MLR	26.21	37.17	0.667
MPR	20.07	12.11	0.895
Singh et al.[13]	18.69	13.09	0.960
RBF-ANN[24]	2.01	8.12	0.974
Our GA-based model	2.14	8.21	0.972
Our PSO-based model	1.27	2.03	0.980
Our ABC-based model	0.42	0.12	0.985

Table 7. The results obtained according to the RMSE, MAPE and R² criteria for our regression model of degree n = 3, as well as for the other tested methods, were evaluated from the UCI Student Performance Data database.

From the tests we performed previously, we notice that when we increase the degree n, RMSE and MAPE errors decrease and the R^2 criterion increases. For this, we perform another test: we gradually increase the degree n of our polynomial, from 0 to 5, using the same training set and the same test set. At each step, we calculate the three error indicators: $RMSE$, $MAPS$ and R^2. The results obtained are shown in Figures 2, 3 and 4. From the three graphs presented in these figures, we can see that the prediction quality of our regression model improves when the degree n increases, but the calculation time becomes more complicated. For n = 2 and n = 4, The distribution of the values predicted by our model compared to the real data is illustrated in Figures 5 and 6. As shown in these figures, the proposed model predicts with high accuracy the values of academic achievement (measured by the average of the first semester of the school year). Moreover, the distribution of the predicted values is closer to the line of the bisector of the first quadrant, especially for n = 4.

The results obtained showed that our regression model, based on adapted multiple polynomials and the exponential term, is very promising in the field of prediction in the field of education. As we can see, it is a generalization of several regression models.

We carry out another experiment in order to evaluate the speed of our regression model at the level of the average

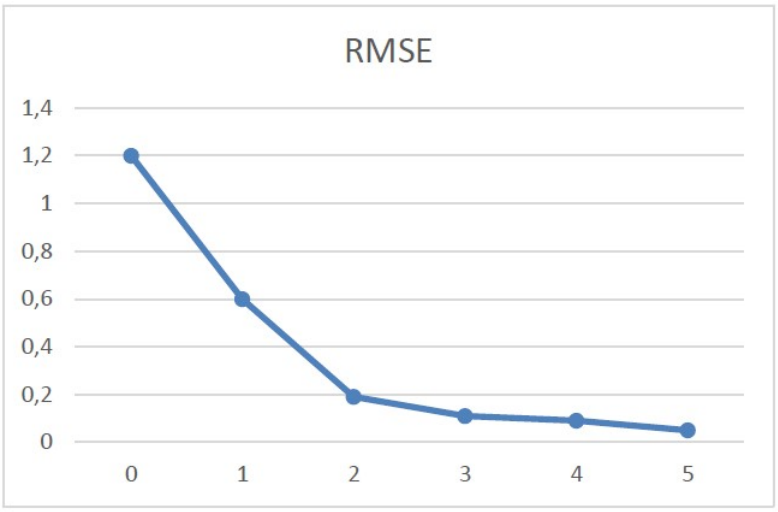

Figure 2. The RMSE criterion of our regression model depending on the degree n ($n = 0, \ldots, 5$).

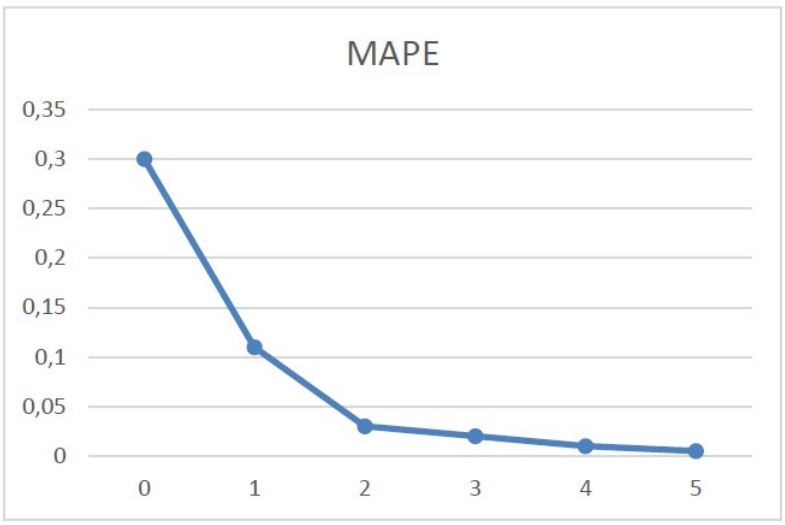

Figure 3. The MAPE criterion of our regression model depending on the degree n ($n = 0, \ldots, 5$).

Polynomial degree (n)	1	2	3	4	5
Average estimated training time (in seconds)	0.005	0.01	0.025	0.06	0.12

Table 8. The average estimated training time as a function of the degree of the polynomial

estimated training time as a function of the degree of the polynomial (n), on a data set of 10 individuals chosen from our Moroccan database, the results obtained are represented in table 8.

From the results presented in Table 8 as well as in Figures 2, 3 and 4, we observe that an increase in the degree of the polynomial leads to an increase in the calculation time, but also an improvement in the quality of the predictions.

The second study: prediction of school dropout

The study of school dropout prediction falls within the field of classification. In this context, we use our classification model presented in subsection 2.5.2, based on the second database (Table 3). This database is divided

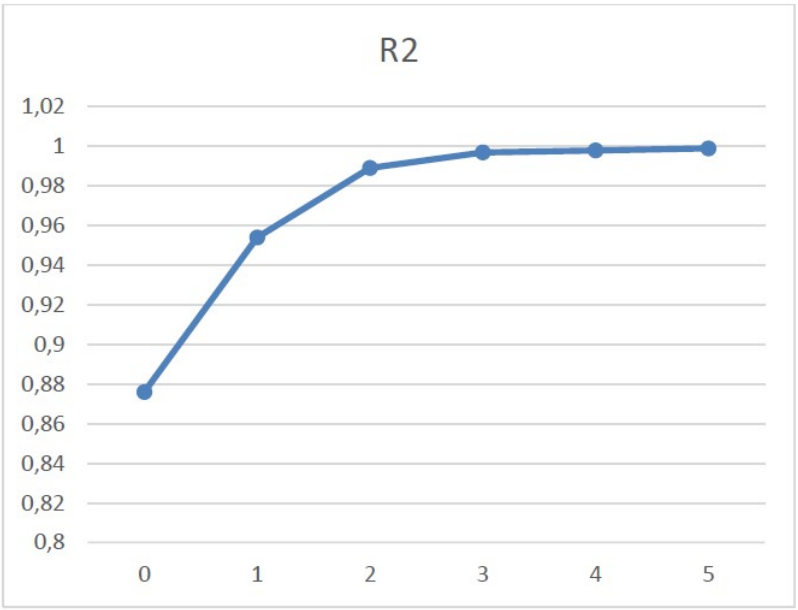

Figure 4. The R^2 criterion of our regression model depending on the degree n ($n = 0, \ldots, 5$).

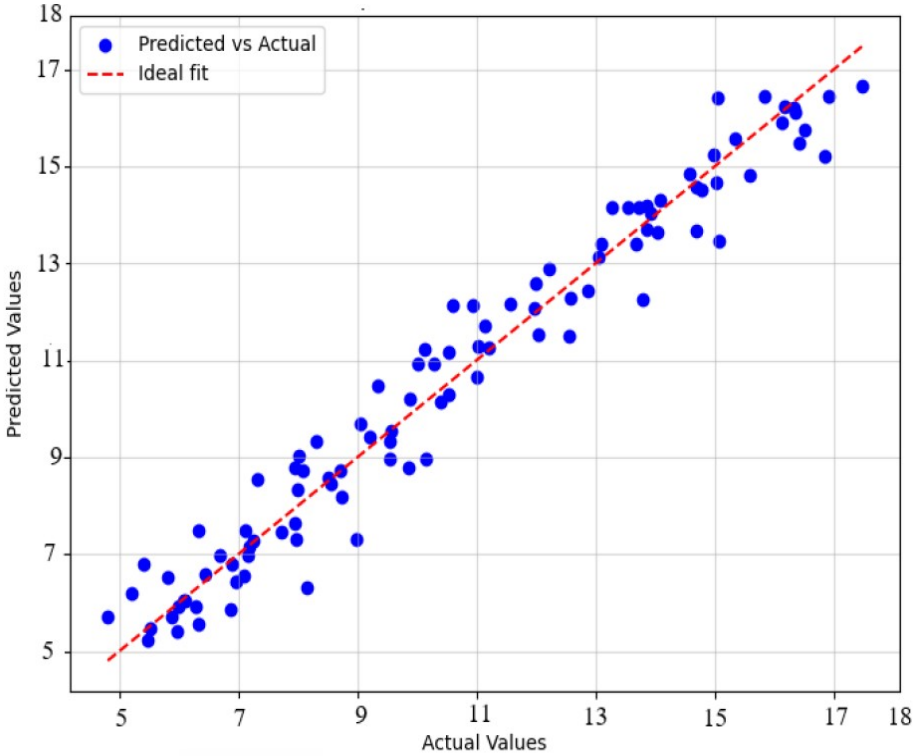

Figure 5. The results of the agreement between the actual and predicted values of the average score of the first semester with our model for $n = 2$.

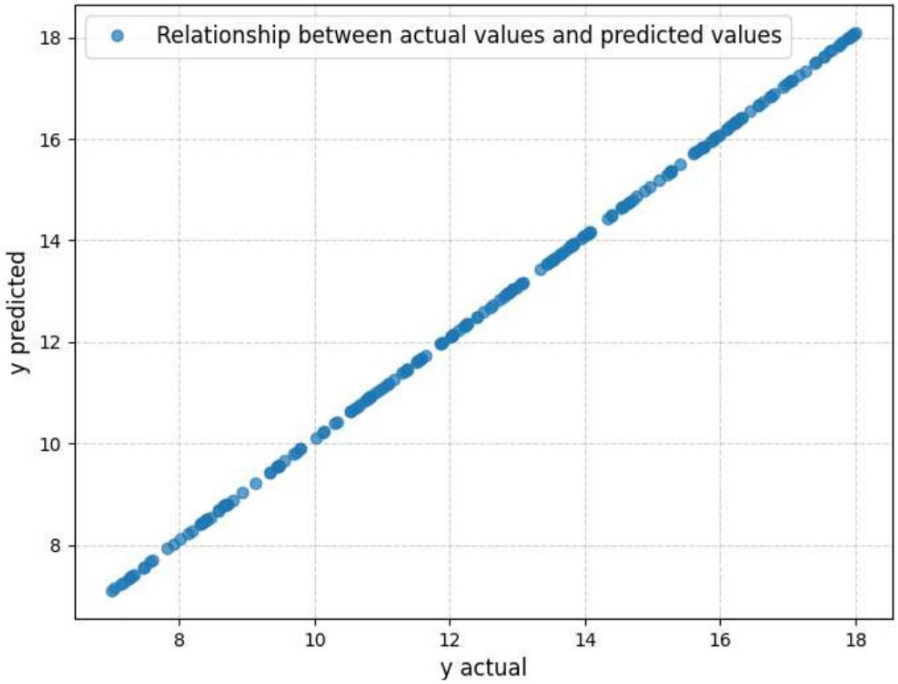

Figure 6. The results of the agreement between the actual and predicted values of the average score of the first semester with our model for n = 4.

Classification Method	CRL	SVM	RBF	MLP	Our method for $n = 2$
Accuracy	79.33%	78.14%	83.15%	89.13%	97.16%

Table 9. the results of the comparative study of our classification model with the CLR, SVM, RBF and MLP methods using the "accuracy" evaluation criterion

into two sets: 80% for training and 20% for testing, in order to evaluate the performance of our classification model. The accuracy, defined in equation (23), is used as an evaluation criterion.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \times 100\% \tag{23}$$

A comparative study of our classification model, with a maximum degree of the adapted Lagrange polynomials fixed at n=2, was also carried out by comparing it to recent classification models, such as classical logistic regression (CLR), the SVM method, as well as RBF and MLP neural networks. The results of this comparison are presented in Table 8.

According to the results obtained in Table 9, we can notice the superiority of our classification method for the prediction of school dropout.

In the second test, we increase the degree n from 2 to 6 and calculate the classification accuracy in order to evaluate the influence of the degree of the polynomials on the quality of the predictions of our classification model. The results obtained are illustrated in Figure 7.

From the results of this test, it can be noticed that the classification quality increases as the degree of the polynomial increases.

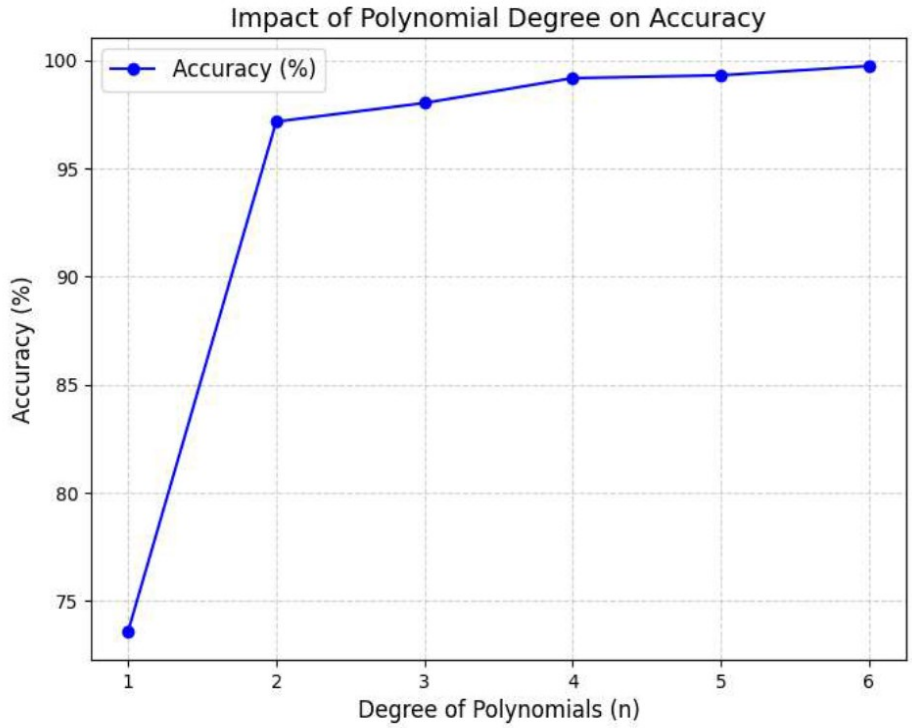

Figure 7. the accuracy of our classification model as a function of the degree of the polynomial.

5. Conclusion

To study the phenomenon of the academic performance of the Moroccan student and analyze and predict its degree we created a new regression model that generalizes the multiple polynomial regression and the multiple exponential models. To predict the degree of the risk of dropping out of school, we proposed a new classification method that generalizes the classification method by logistic regression. We collected data from a questionnaire distributed to students of Moroccan public schools at the first- and second-year baccalaureate level. To validate the two proposed models, we carried out several experimental tests. We carried out comparative studies with recent methods. The results obtained are very promising in the field of education and educational risk management. Although the study presented in this work provides significant results, we plan to explore in future research promising approaches for predicting academic performance and dropout based on machine learning methods, such as those presented in references [24, 25, 26, 27].

REFERENCES

1. WD. Berry, S. Feldman, *Multiple Regression in Practice (Quantitative Applications in the Social Sciences)*, SAGE Publications; Thousand Oaks. CA(1985)..
2. L.Yang, S. Liu, S. Tsoka, LG. Papageorgiou, *Mathematical programming for piecewise linear regression analysis*, Expert Systems with Applications, 44(2016), pp.156-167
3. BROOKOVER, Wilbur B. et SCHNEIDER, Jeffrey M. *Academic environments and elementary school achievement*, Journal of research Development in Education, 1975.
4. A. Hosseinzadeh, M. Baziar, H. Alidadi, J.L. Zhou, A. Altaee, A.A. Najafpoor, S. Jafarpour, *Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions*, Bioresource Technology 303 (2020), p. 122926
5. F.H.M. Salleh, S. Zainudin, S.M. Arif, *Multiple linear regression for reconstruction of gene regulatory networks in solving cascade error problems*, Adv Bioinformat (2017), pp. 1-15

6. M. Syazali, F. Putra, A. Rinaldi, L. Utami, W. Widayanti, R. Umam, K. Jermsittiparsert, *Partial correlation analysis using multiple linear regression: impact on business environment of digital marketing interest in the era of industrial revolution 4.0*, Manag Sci Lett, 9 (11) (2019), pp. 1875-1886
7. HASAN, Sabah Haseeb et IRHAIF, Narjis Hadi. *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, 2021, vol. 10, no 3, p. 1285-1299.
8. FEFER, Sarah A., OGG, Julia A., et DEDRICK, Robert F. *Use of polynomial regression to investigate biased self-perceptions and ADHD symptoms in young adolescents*. Journal of Attention Disorders, 2018, vol. 22, no 12, p. 1113-1122.
9. SAFFER, Boaz Y., MIKAMI, Amori Yee, QI, Hongyuan, et al. *Factors related to agreement between parent and teacher ratings of children's ADHD symptoms: an exploratory study using Polynomial Regression Analyses*. Journal of Psychopathology and Behavioral Assessment, 2021, vol. 43, no 4, p. 793- 807.
10. BEIRLANT, Jan, DIERCKX, Goedele, GOEGEBEUR, Yuri, et al. *Tail index estimation and an exponential regression model*. Extremes, 1999, vol. 2, p. 177-200.
11. MATTHYS, Gunther et BEIRLANT, Jan. *Estimating the extreme value index and high quantiles with exponential regression models*. Statistica Sinica, 2003, p. 853-880.
12. SINGH, Harman Preet et ALHULAIL, Hilal Nafil. *Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach*. IEEE Access, 2022, vol. 10, p. 6470-6482.
13. ARAQUE, Francisco, ROLDÁN, Concepción, et SALGUERO, Alberto. *Factors influencing university drop out rates*. Computers and Education, 2009, vol. 53, no 3, p. 563-574.
14. ALBAN, Mayra et MAURICIO, David. *Predicting university dropout through data mining: a systematic literature*. Indian Journal of Science and Technology, 2019, vol. 12, no 4, p. 1-12.
15. COSTA, Stella F. et DINIZ, Michael M. *Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course*. Education and Information Technologies, 2022, vol. 27, no 9, p. 12381-12397.
16. MASON, Cindi, TWOMEY, Janet, WRIGHT, David, et al. *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, 2018, vol. 59, p. 382-400.
17. CUJI CHACHA, Blanca Rocio, GAVILANES LÓPEZ, Wilma Lorena, VICENTE GUERRERO, Víctor Xavier, et al. *Student dropout model based on logistic regression. In : Applied Technologies: First International Conference, ICAT 2019, Quito, Ecuador, December 3–5, 2019, Proceedings, Part II 1*. Springer International Publishing, 2020. p. 321-333.
18. COUSSEMENT, Kristof, PHAN, Minh, DE CAIGNY, Arno, et al. *Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model*. Decision Support Systems, 2020, vol. 135, p. 113325.
19. TINTO, Vincent. *Dropout from higher education: A theoretical synthesis of recent research*. Review of educational research, 1975, vol. 45, no 1, p. 89-125.
20. TINTO, Vincent. *Rethinking the first year of college*. Higher Education Monograph Series, Syracuse University, 2001, vol. 9, no 2, p. 1-8.
21. DRESEL, Markus et GRASSINGER, Robert. *Changes in achievement motivation among university freshmen*. Journal of Education and Training Studies, 2013, vol. 1, no 2, p. 159-173.
22. OLABANJO, Olusola A., WUSU, Ashiribo S., et MANUEL, Mazzara. *A machine learning prediction of academic performance of secondary school students using radial basis function neural network*. Trends in Neuroscience and Education, 2022, vol. 29, p. 100190.
23. Cortez, P, Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance*.
24. Almasri, A., Alsaraireh, J., Salman, D., & Aburagaga, I. (2022, June). *Explainable Artificial Intelligence Models using Students' Academic Record Data, Tree Family Classifiers, and K-means Clustering to Predict Students' Performance*. In 2022 10th International Conference on Smart Grid (icSmartGrid) (pp. 46-51). IEEE.
25. ALMASRI, A. R., YAHAYA, N. A., & ABU-NASER, S. S. (2025). *PREDICTING INSTRUCTOR PERFORMANCE IN HIGHER EDUCATION USING STACKING AND VOTING ENSEMBLE TECHNIQUES*. Journal of Theoretical and Applied Information Technology, 103(2).
26. Zhou, Z., Ji, J., Wang, Y., Zhu, Z., & Chen, J. (2022). *Hybrid regression model via multivariate adaptive regression spline and online sequential extreme learning machine and its application in vision servo system*. International Journal of Advanced Robotic Systems, 19(3), 17298806221108603.
27. Shen, J., Qian, H., Zhang, W., & Zhou, A. (2024, March). *Symbolic cognitive diagnosis via hybrid optimization for intelligent education systems*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 13, pp. 14928-14936).