Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

# Artificial Intelligence Algorithms and their Role in Assessing the Financial Health of Municipalities in Algeria based on the Logistic Regression Model

**Charif, Aicha Salah**

Laboratory of globalization and economic policies, University of Algeria 3, Algeria.

**Article information**

***Article history:***
Received June 27,2024
Revised:November 15,2024
Accepted : November 20 ,2024
Available online December 1,2024

***Keywords***:
Wealth index
Classification models
Binomial logistic regression

***Correspondence:***
Charif, Aicha Salah
charif.aicha@univ-alger3.dz

**Abstract**

Using a binomial logistic regression model, through this paper we attempt to study the relationship between the state of financial health of municipalities in Algeria, based on the wealth index, in relation to a group of independent variables related to their revenues, the size of spending, and some variables that reflect aspects of this spending and their various specializations. Logistic regression is one of the classification models, and it is considered an alternative model to linear regression models, because this type of model has the property of predicting the probability of the occurrence or non-occurrence of the values of the nominal dependent variables based on a set of explanatory variables (independent variables), whether in their quantitative or qualitative type.

DOI: 10.33899/iqjoss.2024.185245 , ©Authors, 2024, College of Computer Science and Mathematics, University of Mosul.

## 1.Introduction

The aim of this research paper is to build a model to estimate the financial health of municipalities, where the analysis of financial ratios and financial balance indicators, which express a mathematical relationship between two values of quantitative financial variables, is the prevailing analysis, but some studies are interested in predicting the future status of local groups (the municipality or The state for Algeria) reduced the importance of financial ratios and quantitative indicators, From this standpoint, it was very appropriate to search for more accurate scientific methods and models that have a greater ability to predict the probability of faltering municipal financial health. In this context, we propose the logistic regression model due to its ability to predict the binary values of the dependent variable under study, unlike linear regression, as well as Achieving acceptable rating levels.

From this standpoint, the logistic regression model is considered an appropriate model in estimating the probability of municipalities faltering financially or not, as it depends on a set of explanatory variables that have a direct or indirect impact on the problem of municipalities faltering financially.

We attempted to analyze the finances of a sample of municipalities in Algeria, based on a basic problem, which is the suffering of municipalities, especially the largest ones, from the phenomenon of financial pressure, which even reaches the level of financial deficit, which has contributed to the reduction of services provided to citizens, and the government’s interference in many of its decisions. In addition to some spending practices that affect the general financial situation of municipalities in general.

Based on the aforementioned set of foundations, this research was divided into two parts, the first of which dealt with the theoretical analysis of the logistic regression analysis model, and with regard to the second part only it addressed the issue of predicting the financial position of municipalities based on the logistic regression model.

The research problem is represented in the following question:

To what extent can the logistical model be relied upon in classifying municipalities between financially distressed municipalities and financially sound ones?

To answer the main problem, the following sub-questions were asked:

- Are the variables related to revenues and spending considered key variables in the occurrence of the problem of financial distress for the municipalities under study?
- Are the variables related to population density and the powers granted to it considered key variables in the occurrence of the problem of financial distress?
- Is the logistic regression model considered an effective method in predicting the phenomenon of municipal financial distress?

To answer the problem at hand, we will try to test the following hypotheses:

- Variables related to revenues and spending can be considered as the only variables for the occurrence of the problem of financial distress.
- be considered as the only variables for the occurrence of the problem of financial distress.
- It can be considered that increasing the overall correct classification accuracy in predicting the binary values of the dependent variable of the logistic model is a process resulting from selecting independent variables, diversifying them between accounting and non-accounting, and integrating them into a single mathematical model.

The importance of research is demonstrated through explaining the importance of using logistical modeling as a means that can be proven effective in controlling the problem of financial distress of municipalities in Algeria, away from statistical methods that take financial analysis tools as their basis.

## 2.Materials and methods

### 2.1 Theoretical analysis of the logistic regression analysis model

This research aims to address the theoretical basics of the logistic regression analysis model, where an introduction to dependent dummy variables models will be addressed, then the logistic regression analysis model in general, its linear transformations, and its basic hypotheses will be presented.

First requirement: What are classification models?

Classification models are used to make decisions or assign items to categories. Unlike regressors, which produce continuous numbers, such as heights or weights, the output of classification models is logical—either true or false—or categorical decisions, such as “apple,” “banana,” or “cherry .”There are many types of classification models. Some work similarly to classic regression models, while others are fundamentally different. One of the best models that we will apply in our study is called logistic regression . In classification, the output may be a number and not a category. This number represents the probability that the input is a specific type, for example 30% that the fruit is an orange and 70% that it is an apple, and the classification is based on the greatest probability that the fruit is an apple. One example of this type of classification is the Naïve Bayes algorithm. .In short, we can say that the classification model attempts to draw some conclusions from the input values given to the training data in order to predict the outputs/categories for the data designated for testing the model ( Test set ).

The second requirement: An introduction to models of dependent formal (binary) variables

The nature of the dependent variables that follow the binomial distribution and respond to one of the two values (0 or 1) has made them studied and analyzed with statistical models of a special nature that are compatible with the nature of these variables and solve the problem of the shortcomings known to ordinary regression models of both simple and multiple types.

### 2.2 The theoretical presentation of the dependent (binary) variables models

Despite the expansion of the use of formal (binary) variables as explanatory variables, their use as dependent variables is still limited. Perhaps this is due to the many problems that arise when using these variables as dependent variables (1). The ordinary regression model estimated by the ordinary least squares method (OLS) , which takes the form $Y = a + bX$ It plays a good role if the dependent variables are qualitative variables, that is, variables coded with intersecting values such as zero and one. He will not be able to estimate the regression parameters efficiently, and therefore they will not be useful in predicting results or in analysis and forecasting. This type of regression, due to the nature of the dependent variable, will lead to... The lack of heterogeneity in the error variance and the correlation of the error values with the values of the independent variables, so the expected value of the dependent variable will not necessarily fall within the logical and actual range of that variable, that is, it will not fall between zero and one, that is, between no and yes if zero represents “no.” One represents “yes,” for example, since we only have yes or no in the voting process, and this problem remains regardless of the form of the estimated relationship, that is, in linear models and in nonlinear models as well. To address the above problems, we resorted to the linear probability model in LMP regression , which takes the form of the equation:

$$P = E(Y = 1/X) = a + bX \qquad (1)$$

where

$E(Y = 1/X)$ the expected value of the dependent variable.

$a$ : fixed.

$b$ : Coefficient of the independent variable.

$X$ : the independent variable.

It is a model that attempts to show the effect of the independent quantitative variable on the qualitative dependent variable, and the dependent variable here is a probabilistic variable of a special nature because it includes the values of zero and one, as we mentioned above, so its expected value will be between zero and one (2).

Among the most important models that are compatible with the nature of the formal dependent variables, we find the following models:

Logistic Model .
The probiotic model Probit Model .
Linear Probability Model (LPM) .
Tobit model Tobit Model .

**2.3 General introduction of the logistic regression analysis model**

The logistic regression analysis model is considered one of the most common models in terms of the degree of use in various fields in the current era, and this is due to the high degree of effectiveness it has achieved in studies interested in the prediction process for the binary dependent variable and multiple response, especially after the weakness of ordinary linear regression analysis models in estimating parameters. Explanatory variables for phenomena that have a binary and multiple character.

1- The emergence of the logistic model: Credit for developing this model goes to researcher Verhulst, as he was the first to use the logistic function to describe the growth of society and called it the growth function . Then in 1920, researchers Peal and Reed used the function to calculate population growth. They called it the logistic function instead of the growth function (3).

2- Definition of the logistic regression analysis model: The logistic regression analysis model is considered one of the non-linear probabilistic models that analyzes and explains the relationship that exists between a dependent variable (binary that follows the Bernoulli distribution), which is the variable that takes one of the two values (0,1), and several independent variables . It is considered a good method to determine how a number of factors affect the Binary view appears as a dependent variable (Response ) . By binary variable, we mean that this variable takes two possible values. Examples of this include (alive/dead), medical condition (healthy/sick), and so on (2). Logistic regression, sometimes called a logistic model or logistic model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of an event occurring by fitting the data to a logistic curve (4). There are two logistic regression models, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is binomial and the independent variables are either continuous or categorical. When the dependent variable is not binary and consists of more than two categories, multinomial logistic regression can be used (5).

3- Uses of logistic regression analysis: Logistic regression analysis is used to answer a number of questions, including the following:

- Statistical significance of prediction: that is, can the levels or groups of the dependent variable be predicted or predicted using a group of independent variables? This can be verified by comparing a model consisting of the fixed coefficient, in addition to the predictor variables, with the model that includes the fixed coefficient only, as the clear statistical discrepancy between the models indicates the relationship between the predictor variables and the levels or groups of the dependent variable.
- The significance of the predictor variables: that is, what variables predict the combinations of the dependent variable? Does a specific variable increase or decrease the probability of the outcomes or does it not affect the outcomes at all? This can be known in several ways through logistic regression, including the extent to which the model is affected by deleting the predictor variable, or evaluating the statistical significance of logistic regression parameters associated with the set of predictor variables.
- Effect size: What is the extent of the association or relationship between the sets of the dependent variable and the set of predictor variables in the specified model? What is the percentage of variance in outcomes associated with the set of predictor variables?

4- Logistic regression analysis model function: In the case of the Logit model , the relationship between probability and the explanatory variable is considered a non-linear relationship, and probability values range between zero and one. In the case of a single independent variable and a binary descriptive dependent variable, this relationship takes the following formula : Simple Logistic Regression (1):

$P_x = E(y_i / x_i) = 1/1+e^{-(b_0+b_1x_i)}; 0 \prec p \prec 1$ (2)

whereas:

$P_x$ : The dependent variable Y, $E(y_i / x_i)$ : The expected value of the probabilistic dependent variable, $e$ : It is the basis of the natural logarithm, $x$ : the independent variable, and $b_0, b_1$ : Coefficients estimated from the data.

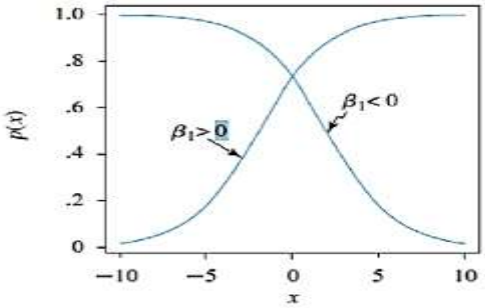
The previous formula can be written as follows:

$P_x = 1/1+e^{-zi}$ (3)

where

$Z_i = b_0 + b_1x_i$

The last equation represents the Cumulative Logistic Distribution Function . It is noted that when $Z_i$ it ranges $+\infty, -\infty$ then $P_x$ will be between 0 and 1, $Z_i$ and $b_i x_i$ it also has a non-linear relationship with $P_x$ . The graphical form of the logistic regression analysis model function can be illustrated in the following figure:



**Source:** R. Lyman Ott, Michael Longnecker, **An Introduction to statistical Methods and Data Analysis**, Broks/cole, Sixth Edition, Canada, 2010, p 702.

**Figure 1: The graphical form of the function of the logistic regression analysis model**

We note that the logistic regression function is S - shaped , meaning that a constant change in X has a smaller effect on the probability when it is near zero than when it is near the middle. Hence, it is a nonlinear response function. In both cases: If $\beta_1 \succ 0, \beta_1 \prec 0$ cannot be interpreted as the slope of the curve as in the case of simple regression, because the change that occurs in the direction of the logistic regression function (up or down) is related to the variable *X*. In other words, it is not a constant change as in ordinary regression. In the case of logistic regression, we can interpret $\beta_1$ as an effect on the odds ratio, meaning an increase *X* will affect in doubly $e^{\beta}$ on odds ratio (6). In the case of both the dependent variable and binary descriptive and there are several independent variables multiple Logistic Regression , the model takes the following form:

$P(x) = 1/1+e^{-z}$ (4)

where

$Z_i = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p$ (5)

Assuming that the random error (*e*) follows a normal distribution with a mean of 0 and a standard deviation $\sigma_{y/x}$ of $e$ □ $N(0, \sigma_{y/x})$ and the variable Y follows a normal distribution with a mean $\mu_{y/x}$ and standard deviation $\sigma_{y/x}$ , that is $y$ □ $N(\mu_{y/x}, \sigma_{y/x})$, for each value of the independent variable *X*.

Given that $E(e) = 0$ , the expected value of the dependent variable Y at a certain value of the independent variable X is as follows:

$E(y / x) = \beta_0^{\wedge} + \beta_1^{\wedge} X$ (6)

However, it is not possible to apply simple linear regression in the case of the binary dependent variable ( Y ), as a result of the following:

- The variance of the dependent variable ( Y ) changes as the values of the independent variable ( X ) change.
- The error variance is not distributed according to a normal distribution.
- The estimated values cannot be interpreted as probabilities because their values do not range between (0,1).

logit model is used , which addresses the previous problems, as it can be written in the case of the presence of one independent variable as follows:

$$\log_e (P/1-P) = \beta_0^{\wedge} + \beta_1^{\wedge} X \qquad (7)$$

In other words:

$$(P/1-P) = e^{\beta_0^{\wedge} + \beta_1^{\wedge} X} \rightarrow P = e^{\beta_0^{\wedge} + \beta_1^{\wedge} X} \,/\, (1 + e^{\beta_0^{\wedge} + \beta_1^{\wedge} X})$$

where:

P : It is the probability of the event of interest occurring, i.e. the probability of success.

P-1 : It is the probability of the event not of interest occurring, i.e. the probability of failure.

$(\frac{P}{1-P})$ : The odds ratio for the event of interest.

$\log_e$ : It is the natural logarithm, with e= 2.7182818284 .

$\log_e = (\frac{P}{1-P})$ : The natural logarithm of the odds ratio or logit .

The symbol log will be shortened from now on to express the natural logarithm $\log_e$ .

It is worth noting here that we use the expression "odds ratio" to the amount $(\frac{P}{1-P})$ .

It is clear from the last relationship that the weighting coefficient belongs to the field $[\infty+, 0]$ This means exceeding the problem of the upper limits of probability, so that $(P_i = 0)$ this means that the odds ratio is equal to zero, but if it is, $(P_i = 1)$ then the odds ratio tends towards $+\infty$ (7).

The goal of resorting to the logit function is the possibility of applying simple or multiple linear regression when analyzing relationships for data with a binary dependent variable, as the range of the logit ranges between $(\infty+, \infty-)$ .

Thus, the process of converting the model from the relationship of a dependent variable and the response rate to a straight line enables the use of simple and multiple linear regression methods to estimate the parameters of this model in terms of statistical and mathematical analysis. However, from a realistic perspective, the importance of this characteristic is known in order to move away from random results, as both Chen and Pounds : In cases where the data is not normally distributed and the appropriate model for this data is non-linear, here we must either fit an alternative model or use an appropriate transformation, which must be used to re-analyze the data (8, 9).

logistic function , as shown in Figure (1), is a continuous function whose range ranges between (0,1), as it approaches zero the closer the right side of the function approaches, $(-\infty)$ and it approaches one the closer this side approaches $(+\infty)$ .

Like other statistical predictive models, analyzing and interpreting the results of logistic regression analysis requires first estimating its parameters, including the Maximum Likelihood method and the Minimum Logit method. $\chi_2$) and the weighted least squares method, but the maximum likelihood method achieved the largest percentage of use in the process of estimating logistic regression parameters.

5- Indicators for evaluating the validity of models: Hosmezr and Lemshow (2000 ) believe that once we fit the logistic regression model (identifying the study variables), the process of evaluating the model begins, and among the statistical tests that can be applied to evaluate the validity of the model:

**Wald test**: It is used to test the significance of the model parameters as it tests the null hypothesis against the alternative hypothesis:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

It is based on the following statistics:

$$Z = \frac{\beta_j^{\wedge}}{S^{\wedge} \times E(\beta_j^{\wedge})} \qquad (8)$$

Since $S^{\wedge} \times E(\beta_j^{\wedge})$ it is the estimator of the standard error of the parameters $\beta_j^{\wedge}$ that follow (if correct $H_0$) a normal distribution, for this reason we compare their observed value with the critical values and do not reject them $H_0$ if they fall between them.

Hosmer-Lemshow test : The test is used to confirm whether the model represents the data well or not, in other words to ensure goodness of fit to evaluate the difference between the observed values and the expected values that were calculated from the estimates of the estimated model , by testing the following hypothesis. :

$H_0$: If the model matches the research data
$H_1$: If the model does not match the research data
The value of this test is compared with the tabular value of the Chi -square $\chi^2$ It is calculated $\chi^2$ from the intersection of the sums of the binary variable y with the sums of the estimated probabilities, and the Contingency table for Hosmer and Lemeshow is created from the intersection of the sums of the dependent variable Y with the sums of the estimated probabilities. The H statistic is used , which follows a distribution $\chi^2$ with a degree of freedom df = m-2 , when the segmentation method is defined by fixed points within the range [0,1], any number of segmentation points can be chosen, and often the segmentation points are m = 10. In this case, the group containing the pair evidence is $[y_i, p(x_i)]$ within the group K according to the following (10):

$$Jk = \left[ i \ : \ (K-1) \ / \ m \ \leq p(xi) \ \leq K / m \ \right] \tag{9}$$

The observed and expected frequencies are calculated , and thus the H&L test statistic is calculated according to the relationship:

$$H \,\&L = \sum_{s=1} \ \sum_{j=1}^{m} (h_{sj} - h_{sj}^{\wedge})^2 / h_{sj}^{\wedge}$$

Accordingly , the statistical hypotheses are formulated according to the following form:
_ Null hypothesis ($H_0$): It indicates that the observed cases are equal to the expected ( predicted ) cases, which means that the model represents the data well.
_ Alternative hypothesis ($H_1$): It indicates that the observed cases are not equal to the predicted cases , meaning that the model does not represent the data well. As for the decision, the null hypothesis is accepted if the statistical probability value is greater than the significance level specified by the researcher.
Confusion Matrix : In order to evaluate the model 's good classification ability , researchers agree that the confusion matrix is considered a good statistical indicator, as it shows actual belonging versus predicted belonging for each group, by matching the observed values of the dependent variable and those predicted . Its importance appears in that it allows determining the amount of error and clarifying its structure as well (11, 12).
If we assume that in the studied phenomenon the dependent variable is binary (1,0), the confusion matrix includes the following values:

- True Positives (TP ): The number of units that are classified + and are actually +.
- False Positives (FP) The number of units classified as + that are in fact -.
- True Negatives (TN) : The number of true negative samples that were correctly classified.
- False Negatives (FN) : The number of true positive samples that are incorrectly classified as negative.

Sensitivity SE: It is defined as the probability value that the expected classification will be positive (the test result is positive or the presence of the characteristic) for the case that is actually positive (knowing that the characteristic is present), and it is calculated according to the following equation:

$$SE = \frac{TP}{TP + FP}$$

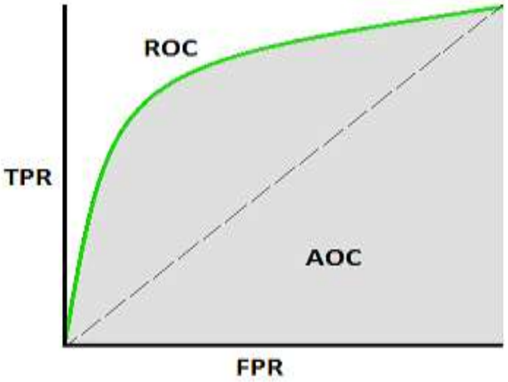
Specificity SP: It is defined as the probability value that the expected classification will be negative (the test result is negative or the property is absent) for the case that is actually negative (the property is not present), and it is calculated as follows:

$$SP = \frac{TN}{TN + FN}$$

Average Classification Rate : It is calculated according to the following relationship:

$$Average\_classification\_rate = \frac{SE + SP}{2}$$

Curve analysis ( ROC ): This analysis is based on studying the relationship between sensitivity and chance (1-accuracy), by representing sensitivity on an axis and in exchange for (1-accuracy) for all cut-off points on an axis, and thus indicates the receiver characteristic curve, which is known as the abbreviation curve ( ROC ) refers to the area under this curve, which ranges between zero and the correct one, as a measure of the model's ability to distinguish between cases that possess the characteristic or event, the subject of examination, and cases that do not possess that characteristic. The area above the diameter of the coincidence is equal to 0.5, and the greater the discriminatory ability of the model, the further away it is. The curve departs from the shell diameter towards the upper left corner as the area under the ROC curve increases until it reaches the value of one, which means complete discrimination of the cases you are interested in studying in the estimated model.



AUC - ROC Curve [Image 2] (Image courtesy: My Photoshopped Collection)

Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

**Figure 2: The graphical form of the function of the logistic regression analysis model**

AUC-ROC curve : It is a performance measure for classification problems at different threshold levels. It is a probability curve and AUC represents the degree or measure of separability. A high AUC value reflects the extent to which the model is able to distinguish between classes. That is, the model is better at predicting classes of 0 as class 0 and class 1 as class 1.

Observing the above curve, we conclude the following points:

If the ROC curve is equal to or less than 0.5 , this indicates the failure of the model to estimate the phenomenon to be studied and the failure of its predictive ability.

If the ROC curve is greater than or equal to the value of 0.7, and completely less than the value of 0.8, this indicates that the predictive ability of the logistic regression analysis model is acceptable and different from chance.

If the ROC curve is greater than or equal to the value of 0.8 and completely less than the value of 0.9, it indicates that the predictive ability of the logistic regression analysis model is excellent and different from chance.

If the ROC curve is greater than or equal to 0.9, it indicates that the predictive ability of the logistic regression analysis model is superior and different from chance.

Pseudo R$^2$ statistic: Pseudo R$^2$ value indicates the strength of the relationship between the qualitative variables used in the model, also called the R 2 McFadden statistic , is calculated according to the following relationship:

$$R^2 = \frac{LR(K)}{-2\log L(\beta_0)} \quad \text{or} \quad R^2 = 1 - \left[\frac{-2\log L(\beta_0, \beta_j)}{-2\log L(\beta_0)}\right]$$

$-2\log L(\beta_0)$ expresses the maximum value taken by the negative of twice the logarithm function in the case in which all model parameters are $\beta_j$ zero except for the constant, while $-2\log L(\beta_0, \beta_j)$ refers to the greatest value taken by the negative double of the reasonable logarithm function in the case where the model includes all parameters, $\beta_j$ including the constant $\beta_0$ . Given the specificity of the logistic regression model in estimating parameters, the coefficient of determination R2 is replaced It is used to determine the suitability of the proposed regression to the study data using the McFadden R2 and Cox & Snell R2 fit statistics They have the same statistical goal by determining the percentage of variance explained in the estimated model, and therefore they are estimated according to the following picture (13):

$$R^2 = 1 - \left[\frac{L_0}{L_1}\right]^{(2/n)} , \quad R^{\square 2} = \left(\frac{R^2}{R_Z^2}\right) , \quad R_Z^2 = 1 - (L_0)^{(n/2)}$$

Since:

$L_0$: The potential function if the model contains only the constant, $L_1$: The possibility function in the case of the model that includes all the explanatory variables, and n : sample size.

## 3. Results and Discussion

First requirement: Defining the population and variables of the study

The study population includes 1,541 municipalities during the fiscal year 2020, and these data are extracted from the information and exploitation system of the Ministry of Interior, Local Communities, and Territorial Development.

Determining the variables for the logistic regression analysis model is based on previous studies that dealt with the financial health of municipalities, as well as the variables proposed by the researcher, It is as follows:

Y : The response variable, which is a binary qualitative variable that takes only two values (value 1 if the financial health of the municipality is good, value 0 if the financial health of the municipality is bad) , and the wealth index was relied upon as the target variable .

The wealth index is the result of dividing the total municipal revenues achieved during year n by the total population for that year, and this index is relied upon to equally distribute the distribution subsidy allocations that municipalities benefit from annually to finance the expenses of the management department.

The explanatory variables are:

$X_1$ : Total self-revenue, a quantitative variable, $X_2$ : Total management expenses, a quantitative variable, $X_3$ : Total tax revenues , a quantitative variable, $X_4$ : The amount of raw saving, a quantitative variable, $X_5$ : The length of municipal roads, a quantitative variable, $X_6$ : Number of primary schools, quantitative variable, $X_7$ : Number of school canteens, a quantitative variable, $X_8$ : Number of treatment rooms, a quantitative variable, $X_9$ : Supply of potable water, a quantitative variable, $X_{10}$ : The rate of connection to potable water, a quantitative variable, $X_{11}$ : Connection to sewage, a quantitative variable, $X_{12}$ : Number of youth hostels, quantitative variable, $X_{13}$ : Number of municipal stadiums, a quantitative variable, $X_{14}$ : Number of cultural centers, quantitative variable, $X_{15}$ : Population density: a quantitative variable, and $X_{16}$ : Area: a quantitative variable.

Second requirement: building and validating the model

First : Descriptive statistics: From the table below the following was noted: No missing values, and there is a clear discrepancy between the various variables and they are in different units of measurement, which requires applying one of the transformations, which is scaling in this case, which represents one of the steps in the data processing and preparation phase in order to model the model as a later step, in order to increase the accuracy of the model.

**Table 1: Statistical indicators of the study variables**

| max | 75% | 50% | 25% | min | std | mean | count | Variable |
|---|---|---|---|---|---|---|---|---|
| 5.927769e+04 | 3.200800e+02 | 1.278000e+02 | 44.03 | 0.07 | 2.841000e+03 | 6.508500e+02 | 1541.0 | **ensitédelapopulation(Hbts/Km2)** |
| 1.487000e+03 | 5.900000e+01 | 3.780000e+01 | 19.60 | 0.00 | 6.649000e+01 | 4.743000e+01 | 1541.0 | **ngueurtotaldescheminscommunaux(Km)** |
| 1.860000e+02 | 1.500000e+01 | 9.000000e+00 | 6.00 | 0.00 | 1.326000e+01 | 1.253000e+01 | 1541.0 | **ombredeprimaires(Nbr)** |
| 6.800000e+01 | 1.100000e+01 | 7.000000e+00 | 4.00 | 0.00 | 7.740000e+00 | 9.150000e+00 | 1541.0 | **nombredecantines(Nbr)** |
| 4.900000e+01 | 5.000000e+00 | 4.000000e+00 | 2.00 | 0.00 | 3.190000e+00 | 4.090000e+00 | 1541.0 | **ombredesalledesoins(Nbr)** |
| 1.345000e+04 | 1.810000e+02 | 1.500000e+02 | 100.00 | 0.00 | 4.582500e+02 | 1.800500e+02 | 1541.0 | **otationjournalière(L/H/J)** |
| 1.000000e+02 | 9.800000e+01 | 9.499000e+01 | 78.58 | 0.00 | 2.504000e+01 | 8.275000e+01 | 1541.0 | **uxderaccordementàl'aep(%)** |
| 1.000000e+02 | 9.574000e+01 | 8.801000e+01 | 65.76 | 0.00 | 2.657000e+01 | 7.659000e+01 | 1541.0 | **uxderaccordementdel'assainissement(%)** |
| 9.500000e+01 | 1.000000e+00 | 1.000000e+00 | 0.00 | 0.00 | 2.590000e+00 | 8.200000e-01 | 1541.0 | **Maisonsdejeunes(Nbr)** |
| 6.100000e+01 | 1.000000e+00 | 1.000000e+00 | 1.00 | 0.00 | 1.750000e+00 | 9.100000e-01 | 1541.0 | **Stadescommunaux(Nbr)** |
| 1.300000e+01 | 0.000000e+00 | 0.000000e+00 | 0.00 | 0.00 | 6.300000e-01 | 2.700000e-01 | 1541.0 | **ntreculturelоumaisondelaculture(Nbr)** |
| 3.366176e+08 | 4.275680e+06 | 1.584500e+06 | 616400.00 | 0.00 | 2.184880e+07 | 6.019555e+06 | 1541.0 | **Recettepropre** |
| 5.076030e+09 | 2.024804e+08 | 1.268422e+08 | 91680356.55 | 0.00 | 4.046868e+08 | 2.324690e+08 | 1541.0 | **depensedefonct** |
| 4.767935e+09 | 4.424349e+07 | 1.059700e+07 | 3615694.04 | 0.00 | 3.060868e+08 | 9.373365e+07 | 1541.0 | **Recettefiscale** |
| 6.686060e+09 | 1.069583e+08 | 5.132119e+07 | 28279012.39 | 0.00 | 3.011729e+08 | 1.160157e+08 | 1541.0 | **Epargnebrut** |
| 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.00 | 0.00 | 4.100000e-01 | 2.100000e-01 | 1541.0 | **ratio_richesse** |

Source: Prepared by the researcher based on model estimation extracts via anaconda navigator

Second: Identifying the most important variables in the study: Machine learning algorithms allow us to deal with the correlation matrix as one of the mechanisms that can be relied upon, to study the correlation between various independent variables on the one hand, and on the other hand to study the correlation between the independent variables and the target variable. It is also considered a technique for selecting the most variables. independent influence on the target variable, and we chose a correlation coefficient threshold of 0.2 to determine the explanatory variables that we will rely on in the rest of the study such as figure 3. It turns out that these variables are population density, the number of primary schools, cultural centers, self-revenue, total management expenses, Tax revenues, raw savings.

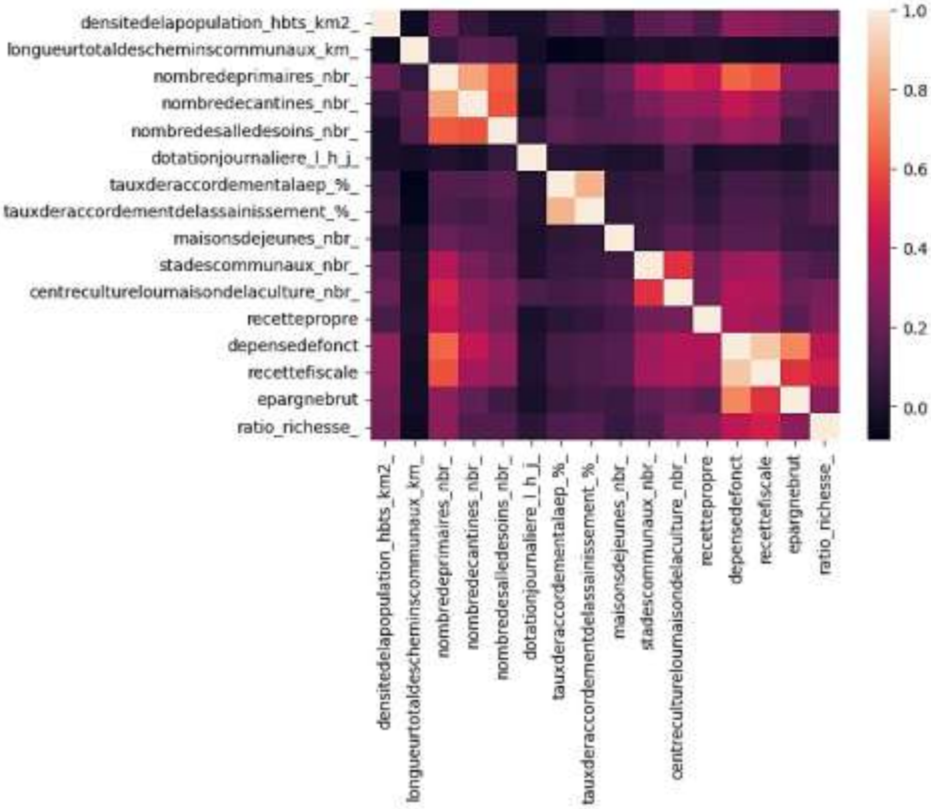


**Figure 3: The correlation matrix among study variables.**

On the other hand, we note that the target variable is unbalanced such as figure 4, as we recorded the presence of 1,219 municipalities whose wealth index is less than the national average, and the rest of the municipalities, which number 322 municipalities, are the only ones whose wealth index value exceeds the national average.

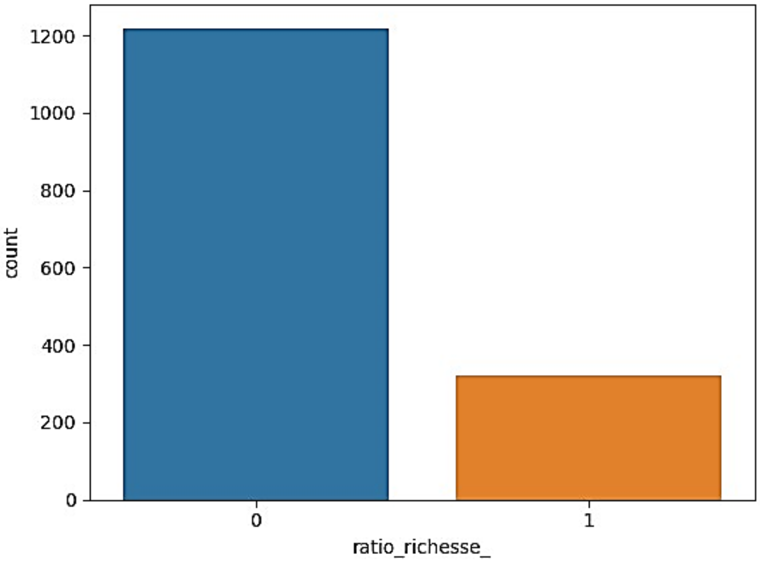


**Figure 4: The Distribution of the target variable values 'wealth index'.**

In order to apply the logistic regression algorithm and obtain has more results, as imbalance in the dependent variable can affect the results of the logistic regression algorithm in several ways (14, 15):

1 . Low prediction accuracy: Imbalance may cause the logistic regression model to decrease in accuracy, especially for the minority class. This happens because the logistic regression algorithm tends to favor the majority class during the training process, leading to the characteristics of the minority class being ignored.

2. Difficulty in interpreting the results: The results of the logistic regression model may become difficult to interpret in the event of imbalance, because the regression coefficients may not accurately reflect the true relationship between the independent variables and the dependent variable, especially for the minority class.

3. Increased likelihood of overfitting: A logistic regression model may be more susceptible to overfitting in the event of imbalance. This occurs because the model may over-preserve the characteristics of the majority class, leading to poor performance on new data.

We applied technology

4. Effects on evaluation metrics: The state of imbalance of the target variable may affect some evaluation metrics of the logistic regression model , such as prediction accuracy and AUC , because these metrics may not accurately reflect the model’s performance on the minority class, so we resorted to applying the RandomUnderSampler technique , a technique from Random sampling techniques, the principle of which is based on taking random samples from the majority class in the data, so that they become equal in size or smaller to the minority class. One of its disadvantages is that random samples from the majority class are deleted without any consideration of their characteristics or values (16). After applying this technique, the number of observations reached 644, divided equally between the number of municipalities whose wealth index exceeds the national average and those whose wealth index is less than the latter such as figure 5.

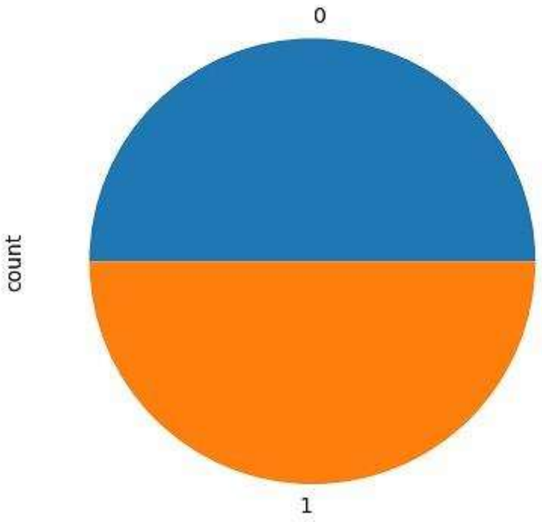


**Figure 5: The Distribution of the target variable values 'wealth index' after rebalancing.**

As we mentioned previously, applying StandardScaler method, which is one of the scaling methods used in data science and machine learning, and equating it as follows:

$$X-{}^{\mu}/_{\sigma}=X^{'}$$

where:

X are the original values of the data.

μ : is the arithmetic mean of the data .

σ : It is the standard deviation of the data. Applying this technique to the study data has accelerated the model training process and reduced the importance of features with large values in building the model, so that all features contribute equally.

Third: Building the model and evaluating it

The final form of the model after building it based on 75% of the data is as follows:

$$\mathbf{Y}=2.08+0.12\,X_{12}+2.14\,X_{1}+1.23\,X_{2}+5.66\,X_{3}-0.13X_{4}-1\,.21X_{6}-0.05\,X_{15}$$

Fourth: Measuring the performance of the logistic regression model

Accuracy : The model's accuracy rate was 83.85%, which reflects the percentage of samples that were correctly classified by the model, and this percentage is largely acceptable.

Recall : 67.47% Recall measures the proportion of true positive samples that are correctly classified as positive by the model, which is fairly acceptable.

Precision : 94.85%, which represents the percentage of samples classified as positive by the model that are actually positive .F1 Score : It is considered a good measure of the balance between the model's accuracy and its ability to detect positive cases, and in this model it reached 78.85%.

Confusion Matrix : As we mentioned previously, the confusion matrix provides a detailed display of the number of samples that were classified correctly and incorrectly by the model for each category, and based on the results obtained from the model, we find such as figure 6.

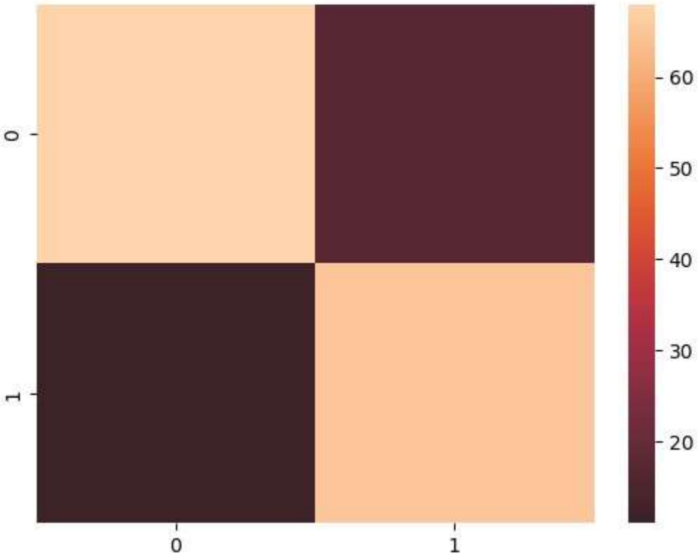


**Figure 6: The Confusion Matrix.**

_True Positives (TP ): The number of units that were rated + and are actually +, estimated at 72 observations.

_False Positives (FP): The number of units classified as + that are in fact -. Estimated 13 observations.

_True Negatives (TN) : The number of true negative samples that were correctly classified. Estimated 63 observations.

_False Negatives (FN) : The number of true positive samples that are incorrectly classified as negative. Estimated 13 observations.

AUC-ROC (Area Under the ROC Curve) : From the figure 7 we find that the ROC curve is greater than or equal to the value of 0.9, and this indicates that the predictive ability of the logistic regression analysis model is superior and different from chance.

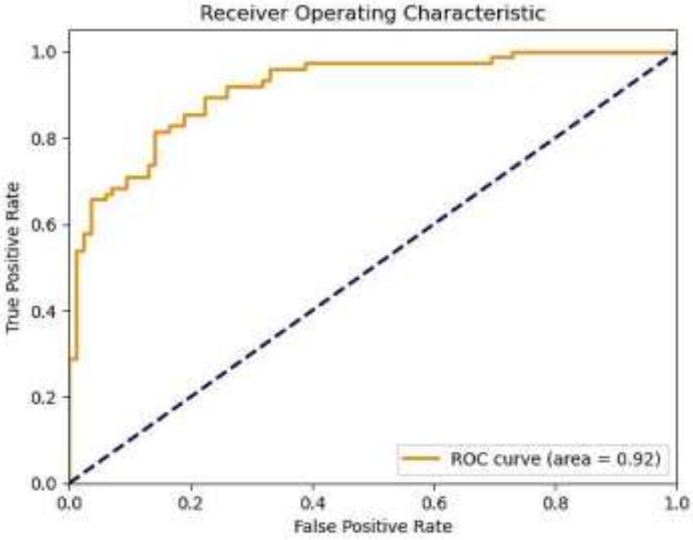


**Figure 7: AUC-ROC curve.**

## 4. Conclusion:

Adopting advanced models to estimate and analyze the financial health of local communities in general and municipalities in particular is considered a solution and a requirement to understand their situation and find solutions to their faltering, bearing in mind that the budget deficit is a phenomenon that most municipalities in Algeria suffer from.

The results of the model adopted in the study show that the total explanatory variables with which the study began, including only population density, number of primary schools, cultural centers, personal revenues, total management expenses, tax revenues, and crude savings, were the most affected and significant in the model, which proves the validity of the hypothesis. The second is that financial variables are key variables in the occurrence of the problem of financial default, in addition to a number of variables from the first hypothesis.

On the other hand, performance measures of the logistic regression model based on artificial intelligence algorithms showed that it is considered an effective tool for predicting the phenomenon of financial distress. The importance of adopting such a model appears in light of the growing financial challenges facing municipalities.

The phenomenon of financial distress is considered one of the most serious challenges facing municipalities, as it affects their ability to provide basic services to citizens and implement development projects. Therefore, predicting the occurrence of this phenomenon is crucial to enabling municipalities to take the necessary preventive measures to avoid it or mitigate its effects .

Prediction models, such as logistic regression based on artificial intelligence algorithms, play an important role in this field, because they... It analyzes large amounts of financial, economic, and other data related to municipalities, and extracts the main indicators that increase the probability of financial default . Artificial intelligence algorithms are characterized by the ability to analyze complex data more accurately and efficiently than traditional methods, which leads to improving the accuracy of predicting the occurrence of financial default . In addition to early prediction, these models can detect early warning signs of financial distress, allowing municipalities to take preventive measures early .

Overall, using artificial intelligence models to predict financial distress in municipalities provides a powerful tool that can help municipalities improve their financial stability and achieve their development goals.

## References

1. Kassens, A.L., *Theory vs. practice: Teaching undergraduate econometrics.* The Journal of Economic Education, 2019. **50**(4): p. 367-370.
2. Costa e Silva, E., et al., *A logistic regression model for consumer default risk.* Journal of Applied Statistics, 2020. **47**(13-15): p. 2879-2894.
3. Ani, C., U. Hassan, and M. Maiwando, *Prediction of Banks Stock Performance in Nigeria Stock Exchange Market: An Application of Logistic Regression Model.* Abacus (Mathematics Science Series), 2020. **44**(1).
4. Sagarra, M., C. Mar-Molinero, and M. García-Cestona, *Spanish savings banks in the credit crunch: could distress have been predicted before the crisis? A multivariate statistical analysis.* The European Journal of Finance, 2015. **21**(3): p. 195-214.
5. Park, H.-A., *An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain.* Journal of Korean academy of nursing, 2013. **43**(2): p. 154-164.
6. Fernando, R., *Logit, probit and tobit: Models for categorical and limited dependent variables.* PLCS/RDC Statistics and Data Series at the West, 2011.
7. Swamynathan, M., *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. 2017: Springer.
8. ALbazzaz, Z.M. and O.B. Shukur. *Using LSTM Network Based on Logistic Regression Model for Classifying Solar Radiation Time Series*. in *International Conference on Explainable Artificial Intelligence in the Digital Sustainability*. 2024. Springer.
9. Shukur, O.B., *Using hybrid procedure in generalized additive model and classification tree to classify agriculture evaporation in Mosul city.* Int. J. Agricult. Stat. Sci. Vol, 2021. **17**(1): p. 1509-1514.
10. Shukur, O.B. and O.A. Malaa, *Comparison of Logistic regression, Convolution Neural Network, and Kernel Approaches for Classifying the Caenorhabditis Elegans Motion.* Iraqi Journal of Statistical Sciences, 2023. **20**(2): p. 175-187.
11. Tareq, W.K. and O.B. Shukur, *Using Cart Approach for Classifying Climatic Status of Mosul City.* Journal Of Agricultural And Statistical Sciences, 2021. **17**: p. 2325-2331.
12. Rakotomalala, R., *Pratique de la Régression Logistique. Régression Logistique Binaire et Polytomique. Version 2.0.* Lyon, Université Lumière Lyon-2, multigr, 2014.

13.Baima, G., et al., *Periodontitis prevalence and severity in inflammatory bowel disease: A case–control study.* Journal of periodontology, 2023. **94**(3): p. 313-322.
14.Osisanwo, F., et al., *Supervised machine learning algorithms: classification and comparison.* International Journal of Computer Trends and Technology (IJCTT), 2017. **48**(3): p. 128-138.
15.Yuvalı, M., B. Yaman, and Ö. Tosun, *Classification comparison of machine learning algorithms using two independent CAD datasets.* Mathematics, 2022. **10**(3): p. 311.
16.*Random Under Sampler*, in *API reference*. https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

# خوارزميات الذكاء الاصطناعي ودورها في تقدير الصحة المالية للبلديات في الجزائر بالاعتماد على نموذج الانحدار اللوجستي

عائشة شريف صالح
مخبر العولمة والسياسات الاقتصادية ، جامعة الجزائر 3، الجزائر .

**الخلاصة:** باستخدام نموذج الانحدار اللوجستي ثنائي الحدين، من خلال هذه الورقة نحاول دراسة العلاقة بين حالة الصحة المالية للبلديات بالجزائر بالاعتماد على مؤشر الثروة بالنسبة لمجموعة من المتغيرات المستقلة المتعلقة بإيراداتها وحجم الإنفاق وبعض المتغيرات التي تعكس أوجه هذا الانفاق واختصاصاتها المختلفة.

الانحدار اللوجستي من نماذج التصنيف، ويعتبر كنموذج بديل عن نماذج الانحدار الخطي، وذلك لامتلاك هذا النوع من النماذج خاصية التنبؤ باحتمالات حدوث أو عدم حدوث قيم المتغيرات التابعة الاسمية انطلاقا من مجموعة من المتغيرات التفسيرية (المتغيرات المستقلة) سواء كانت في شكلها الكمي أو النوعي.

**الكلمات المفتاحية:** مؤشر الثروة، نماذج التصنيف، الانحدار اللوجستي ثنائي الحدين.