International Journal of Engineering &Technology, 13 (1) (2024) 102-111

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Proposal of new models for prediction of the cost of agricultural raw materials in a business intelligence and machine learning context

Kanga koffi [1] *, kamagaté beman Hamidja [2], BROU Aguié Pacôme Bertrand [3]

[1] Doctor in computer science specializing in software and database engineering: INPHB doctoral school Teacher – researcher at ESATIC (African Higher School of ICT: Republic of Ivory Coast)
[2] Doctor in computer science specializing in network and cybersecurity INPHB doctoral school Teacher – researcher at ESATIC (African Higher School of ICT: Republic of Ivory Coast)
[3] Doctor of Computer Science Teacher – researcher at ESATIC (African Higher School of ICT: Republic of Ivory Coast)
**Corresponding author E-mail: kanga.koffi@esatic.edu.ci*

Abstract

In this paper, we propose a data model for prediction of the cost of raw materials in a business intelligence context. Our contribution focuses initially on the implementation of a model with a star representation. This model highlights the fact (cost) to be predicted according to the axes linked to it. Secondly, from this basic model, our contribution is based on sub-models enabling us to carry out mono-dimensional anal-yses of the 'cost' fact. Thirdly, from these sub-models we establish associated mathematical models that allow us to deduce a global mathe-matical model from our basic model using linear regression and artificial neural network techniques. The implementation of these mono-dimensional sub-models in a machine learning database management system 'Minds DB', produces results that allow the prediction of raw material costs. Also, the predictions made by "Minds DB" are computationally validated by linear regression techniques which give better results than those of artificial neural networks.

Keywords: *Data Model; Business Intelligence; Prediction Model; Machine Learning; Database; Predictor.*

1. Introduction

The use of data models to predict the cost of agricultural raw materials requires a great deal of explanation. For decades, many countries have based their development on agriculture. These countries have set up organizations and governance systems to improve producers' incomes and their own various gross domestic products. At times, the various products produced by these agricultural organizations are experiencing difficulties in gaining visibility of production costs and other costs over time. The problem of a data model for predicting these different costs, therefore deserves to be addressed in order to find an effective solution. In the field of information science, a data model describes the way in which data is represented in an organization, an information system or a database. For some authors [1], a data model shows a structural foundation, represented in the form of a well-defined graphical characterization of a business information system. Business Intelligence [2] is the technological process of analyzing data and presenting information to help executives, managers and other end-users in businesses and organizations make informed decisions. It encompasses a wide variety of tools, applications and methodologies that enable organizations to collect data from internal systems and external sources. This data is then prepared for analysis to create reports, dashboards and other visualization tools to make analytical results available to decision-makers and operations. To achieve this, some companies have implemented tools that [1] and applications with storage systems based on data models that vary from one author to another. In a context of improving the income of part takers through effective prediction of their income and therefore, the components, that make up this income, it is therefore necessary to equip ourselves with models (data and mathematical models) for acceptable prediction in order to harmonize the results of this prediction.
The rest of our paper is organized as follows:

- In section 2, we will present the state of the art
- In section 3, we will set out our problematic
- In section 4, we will illustrate our contribution
- Section 5 is devoted to a discussion and we will end with a conclusion in section 6. In this section, we will outline a number of perspectives.

Copyright © Kanga koffi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. State of the art

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

2.1. Analysis of agricultural data using data mining techniques: application of big data [3]

In this paper, the authors focus on the analysis of agricultural data and the search for optimal parameters to maximize agricultural production using data mining techniques (clustering by partitioning algorithms) such as PAM, CLARA, DBSCAN and multiple linear regression. These clustering methods are used and compared using quality metrics.
According to the authors of these comparisons, DBSCAN gives better clustering quality than PAM and CLARA, and CLARA gives better clustering quality than PAM.
Base on the analysis, a very good job has been done, but they don't take into account the aspects linked to the analysis of sales costs in order to find a context for improvement.

2.2. Application of data science and materials science to preventive maintenance in an industrial context [4]

In this work, the authors started from the observation that data science has applications in all areas of life and can therefore be applied to industry. According to the authors, five industrial predictive maintenance processes are used by data scientists. These are:

- The maintenance planning and scheduling process,
- the decision-making process based on reliability and degradation,
- the joint optimization process,
- the process of estimating and optimizing maintenance costs and risks,
- the optimization process for multi-state and multi-component systems.

A number of techniques, scientific methods and algorithms are used to extract and apply treatments to improve industrial production. Some of these tools include supervised learning algorithms, unsupervised learning and reinforcement learning algorithms.
On analysis, this work only concerns the industrial sector. It does not deal with issues linked to agriculture and therefore to the various costs involved. However, the techniques used to predict maintenance work could be used and even optimized for issues relating to costs (sales, production, etc.) and quantities (production, sales and demand) .

2.3. Data mining and wireless sensor networks for disease and pest prediction in agriculture [5]

This work was carried out as part of the fight against plant pests, parasites and diseases. To achieve their objectives, the authors use wireless sensor networks to collect temperature and meteorological data. To these data, they apply data mining techniques supported by mathematical models based on multivariate linear regression to make predictions about diseases and pests.
They do not dwell on the financial and quantitative aspects of raw material production. These are all elements that can improve the living conditions of producers. Of course, a sick plant cannot produce a satisfactory yield, but when it is in good health, it would be ideal to predict its production and the financial value of that production over time. Another limitation of this excellent work concerns storage. The work does not specify the format in which the data collected is stored let alone the model underlying this data.

2.4. Machine learning tools (atom) and databases

In [14], the authors carried out a study on the design of queries for processing large volumes of time-varying data using machine learning techniques for distributed systems. These techniques are based on the information produced by users, the properties (response time, production of reliable information, processing periodicity, collection and use of processing history or log, independence of processing interfaces) of these systems and the various functions associated with them (learning, inference and updating). An API model for automatic data processing using machine learning techniques was also presented. This API [figure1] has the following functions:

- data-store management for data to be predicted,
- Query processor,
- management of algorithms (model store),
- estimation management: this includes setting the parameters for models, queries and data, as well as the various estimating functions.

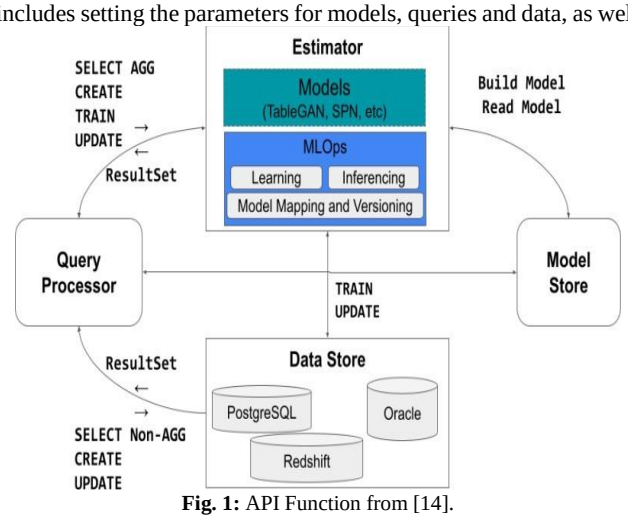


Fig. 1: API Function from [14].

After the analysis of the functionalities presented in this work, the aspects relating to prediction based on multidimensional star models are not taken into account.

2.5. Prediction model for time series [17]

A time series is a table showing a list of values or data that change over time. Time series are used in several fields. The best known to our knowledge are the ARIMA and SARIMA models.

2.5.1. The ARIMA model

This model consists of 2 parts:
The AR (AUTO REGRESSIVE) model, which aims to predict the value of a time series at a given date t by summing the data over a set of p previous instants. It represents a generalization of non- stationary series. It is composed of the AR (auto-regressive) models.

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t \tag{1}$$

With

- Xt the value at time t
- t ϵt he error at time t

the MA (MOVING AVERAGE) model, which aims to predict a value at time t based on the errors of the last q instants.

$$X_t = \epsilon_t + \sum_{i=1}^{q} \beta_i \epsilon_{t-i} \tag{2}$$

Where:
X_t the value at time t
ϵ_t the error at time t
β_i the coefficient associated with ϵ_{t-i}

So the ARMA process combines an AR process and an MA process. It is noted ARMA (p, q). with

- p = the number of lags to be considered for the auto-regressive model
- q = the order of the MA model.

The resulting ARMA formula is as follows:

$$X_t = \underbrace{\sum_{i=1}^{p} \alpha_i X_{t-i}}_{AR(p)} + \overbrace{\epsilon_t}^{Residual} + \underbrace{\sum_{i=1}^{q} \beta_i \epsilon_{t-i}}_{MA(q)} \tag{3}$$

$$X_t = \epsilon_t + \sum_{i=1}^{q} \beta_i \epsilon_{t-i} \tag{4}$$

Limit
Although this model is very simple andgives good results, it has a few limitations. In fact, it only gives good results only on so-called stationary timeseries, i.e. with constant means and variance. In its presentation, it does nottakes into account several parameters that determine the value to be predicted. Seasonality (periodicity) is not also taken into account.

2.5.2. The SARIMA model [18]

In this work, the authors made a comparison between forecasting models based on SARIMA and SARIMAX(SARIMA applied to exogenous data) for forecasting the electricity production of photovoltaic power plants. They assumed that electricity production in these plants isseasonal. All of which led them to choose SARIMA variants to carry out their prediction work. The daily updating of the data used by these models enabled the authors to observe differences in the accuracy of their results.

2.5.2 The SARIMA model [18]

In this work, the authors made a comparison between forecasting models based on SARIMA and SARIMAX(SARIMA applied to exogenous data) for forecasting the electricity production of photovoltaic power plants. They assumed that electricity production in these plants isseasonal. All of which led them to choose SARIMA variants to carry out their prediction work. The daily updating of the data used by these models enabled the authors to observe differences in the accuracy of their results.

3. Problem

From all of the above, it can be said that some excellent research work has been carried out. Some of this work has contributed to the prediction of diseasesand pests of plants producing rawmaterials, while others have involved the use of data mining techniques with their corollary processing tools. However, this work does not take into account the aspectslinked to storage, or the model that will enable standardized storage to be set up, sothat prediction work can be carried out taking time into account.

To make predictions based on data produced over time (year - half-year - quarter - month, etc), there is a need to harmonize and produce a data model incorporating this factor (time). This would enable governments to plan the sale of raw materials in order to improve producers' incomes based on predicted selling prices.

4. Contribution

In order to have a complete prediction system in the agricultural domain, our contribution is firstly to setup amultidimensional data model (integrating the notion of time) ensuring complete storage of the data to be used for predictions. An analysis is then carried out,leading to the formulation of mathematicalmodels associated with different data models. This contribution is implemented in a machine learning environment associating databases (under MinsDb) by combining linear regression and neural network algorithms in order to measure the effectiveness of our proposal.

4.1. Modelling

Firstly, we propose a storage model derived from static modelling techniques for information systems [6]. It presents entities linked together by relationships. The degree of connection between theseentities is represented by cardinalities (Figure 2).

Fig. 2: Cardinality and Association.

Our model takes into account the relevanceof the data (enabling a concrete improvement in producers' income). Thesedata are: price, year, quantity produced,quantity demanded, reserve andconsumption index. To do this, we proposeobjects that form the basis of our model. They have two (2) types of attributes:

- simple attributes and
- attributes qualified as identifiers (enabling others to be found uniquely without duplication).

This allows our model to avoid duplicate predictions. The objects we propose are :

- "YEAR", characterized by its identifier (year ID) and its value (Year), is the object used to identify the year in which the prediction is to be made.
- "COUT_PRODUIT " This object contains the "price" information that we wish to predict. This information could be a function of the quantity demanded (in a supply and demand context), the quantity produced (constituting the supply), the reserved quantity (quantity_reserved) and the consumption index (Val_index).
- "QUANTITY_DEM", representing market demand. It is used in the trade balance.
- "QUANTITY_RES", representing the reserve in the event of a product shortage. It could be used for speculation
- "QUANTITY_PROD" representing the producers' production
- "INDEX_CONSO" it Represents the ratio between the quantity consumed and the quantity produced. In other words, it represents the proportion of production consumed

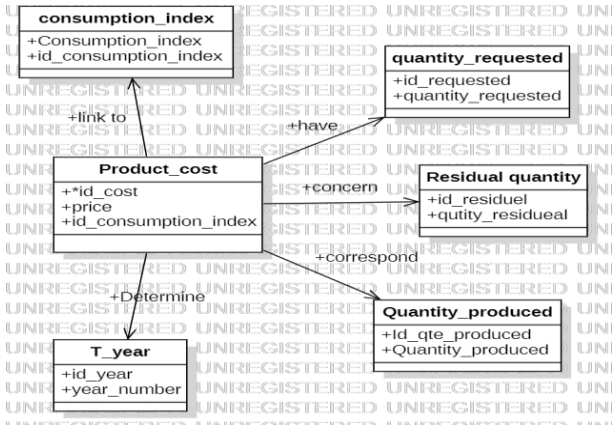


Fig. 3: Cost Forecast Raw Data Model.

4.2. Model transformation

In order to make an efficient prediction ofthe raw material price, we present a model(Figure 4) represented in the form of astar. It is obtained by transformation offigure 3. In this model we have 2 types ofobjects: dimension tables and fact tables.
In business intelligence, a fact-tablerepresents the fact that we want to analyze.In our case, the fact is represented by the'COST' table. The other tables representthe dimensions of our analyses.
The dimensions are Year, Quantity_Prod, Quantity_Res, Quantity_Dem And Index_Cons.
These are the axes along which the fact is analyzed or predicted.
Table 1 below shows each entity and its ibin the model (Figure 4).

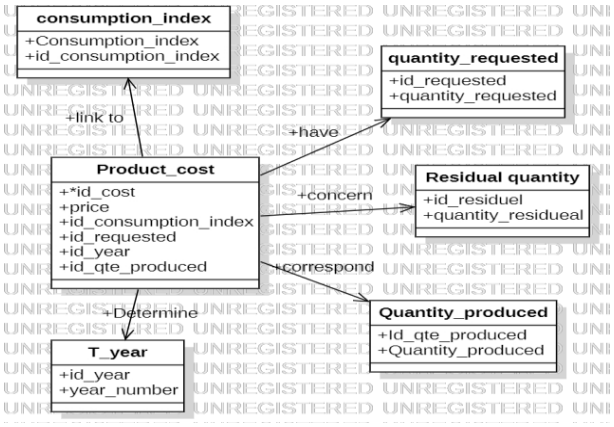


Fig. 4: Star Representation of Data Model.

Table 1 : Classification of Entities (Fact or Axis)

Denominationon	Roles	Fact	Axis
COST	Represents the item to be analyzed (price of sale)	yes	
YEAR	Represents the analysis period. So it represents the time analysis		yes
QUANTITY_PROD	Represents the quantity of material first produced		yes
QUANTITY_DEM	Represents the quantity of raw material requested onthe market		yes
QUANTITY_RES	Represents the product reserve. It could be used for the speculation		yes
CONSO_INDEX	It represents the proportion of the production		yes

Mathematical variables from the model
Analyzing the model (Figure 4) and the definitions in Table 1, we define the following variables:
cp = selling price
An = year of prediction,
qp= quantity produced,
qd =quantity demanded,
qr = reserve quantity ,
ic = consumption index

4.3. Prediction sub-models and mathematical modelling

In this part of our contribution, we presented the data sub-models, their associated mathematical models and the various relationships between these models.

- Sub-model 1 (Figure 5)

This sub-model links the 'COST' and'YEAR' entities, and can be used to store and carry out predictive analyses of the'COST' , 'YEAR' and selling price, taking into account theperiod .

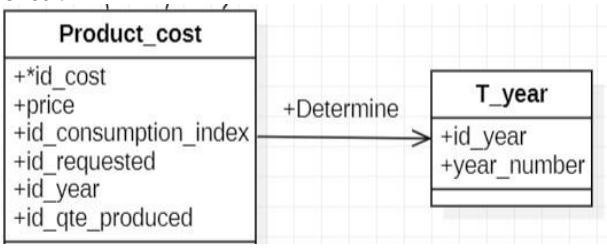


Fig. 5: Cost Per Year Prediction Sub- Model.

Figure 5 provides a prediction of the selling price by period. It could be refined by breaking down the year into months, weeks, days, etc.
Mathematically, saying that this sub-model1 (Figure 5) links the 'COST' and 'YEAR' entities is means that these entities are linked by a function that we denote fsm1, which is such that :

$$Cp = fsm1(an). \quad (4)$$

This function bases its argument on a series of discrete values that can be understood as a time series (where an represents a time variable).
Sub-model 2 (Figure 6)
Here, this sub-model links the 'COST' and 'QUANTITY_PROD' entities. It is used to store and predict the selling price (COST) as a function of the quantity produced.

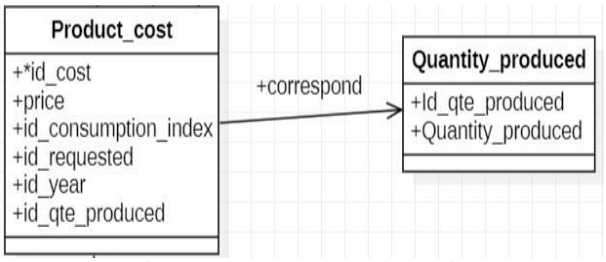


Fig. 6: Sub-Model for Predicting Cost as A Function of Quantity Produced.

Figure 6 shows a prediction of the evolution of the cost (selling price) giventhe evolution of the quantity produced.
Mathematically speaking, if we call cp thecost to predict, we can write: Mathematically, saying that this sub-model2 links the entities 'COST' and 'QUANTITY_PROD', means that these entities are linked by a function that we denote fsm2 which is such that :

$$Cp_2 = fsm2(an, qp) \quad (5)$$

- Sub-model 3 (Figure 7)

This sub-model links the 'COST' and'QUANTITY_RES' entities. It is used to store and predict the selling price (COST) as a function of the reserve quantity.

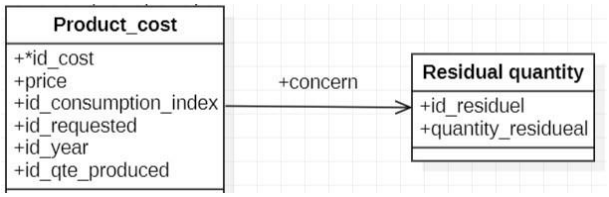


Fig. 7: Cost Prediction Sub-Model.

Mathematically, saying that this sub-model3 links the entities 'COST' and 'QUANTITY_RES', means that these entities are linked by a function that we denote fsm3 which is such that :

$$Cp_3 = fsm3(an, qr) \quad (6)$$

- Sub-model 4 (Figure 9)

This sub-model links the 'COST' and'CONS_INDEX' entities. It is used to store and predict sales prices (COST) as a function of the consumer index.

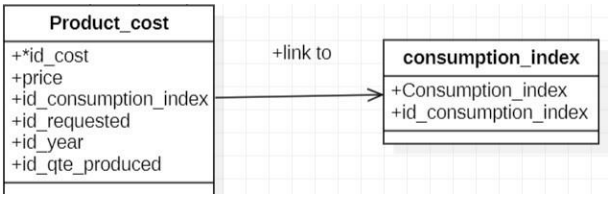


Fig. 8: Cost Sub-Model - Consumption Index.

Mathematically, this sub-model 4 linksthe 'COST' and 'CONS_QUANTITY' entities,
In other words, these entities are also linked by a function that we call fsm4 which is such that :

$$Cp_4 = fsm4(an, ic) \quad (7)$$

- Sub-model 5 (figure 9)

This sub-model links the 'COST' and'QUANTITY_DEM' entities. It is used to store and predict the selling price (COST) as a function of the quantity requested.

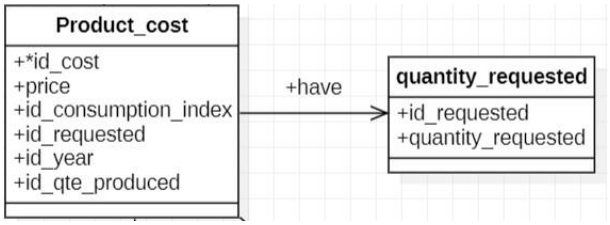


Fig. 9: Cost - Quantity Demanded Sub-Model.

Mathematically, saying that this sub-model 4 links the entities 'COST' and'QUANTITY_DEM' also means that these entities are also linked by a function that we denote fsm5 which is such that :

$$Cp_5 = fsm5(an, qd) \quad (8)$$

4.4. General mathematical model

Based on the Analysis of the data model shown inFigure 4 and the classification shown inTable1, taking into account the mathematical models derived from sub- models 4, 5, 6, 7 and 8, leads to the following conclusions:

Cp1 = fsm1(an). (4)
Cp2 = fsm2(an, qp) (5)
Cp3 = fsm3(an, qr) (6)
Cp4 = fsm4(an, ic) (7)
Cp5 = fsm5(an, qd) (8)

a) Application of linear regression techniques

By linear combination we obtain :

$$P = \sum Cp = fsm1(an) + fsm2(an, qp) + fsm3(an, qr) + fsm4(an, ic) + fsm5(an, qd) (9) + \varepsilon i \quad (9)$$

CP is written :

$$p_i = \beta 0 + \beta 1 offer_{i,1} + \beta 2 production_{i,2} + \beta 3 reservei_{,3} + \beta 4 consoi_index_{,4} + \beta 5 demandi_{,5} + \varepsilon i \quad (10)$$

With:
P_i is the i-th observation of the price variable ;
$offer_{i,1}$ is the i-th observation of the variable offer;
$production_{i,2=}$ the i-th observation of the production variable;
$reserve_{i,3}$ is the i-th observation of the reserve variable;
$index_conso_{i,4}$ is the i-th observation of the consumption index variable;
$demand_{i,5}$ is the i-th observation of the demand variable;
εi (the stochastic component) is the model error.
From equation (10), we can see that if we apply linkage regression, the value Pi is our endogenous variable and we have 5exogenous variables which are $offer_{i,1}$, $production_{i,2}$ + $reservei_{,3}$ + $consoi_index_{,4}$ + $demandi_{,5}$. Thus the difference generated by our model, which we call E, is
$E = C-\hat{C}$.
after calculating thecorrelation coefficient between the cost and each exogenous variable, we also calculate the RMSE (Root Mean Square Error), which is the "average" of the squares of the residuals . of our model . thermse is such that :
RMSE = SRes / (n -p - 1)
With :

- SRes = sum of Ri^2 (sum of residuals squared);
- n: number of individuals;
- p: number of explanatory variables;
- (n p-1): number of degrees of freedom.

NB: The smaller the RMSE, the better of the regression.

b) Application of technique neural network

For our model we have one endogenous variable which is the forecast cost P and 5 exogenous variables which are:
supply, production, reserve and consumption index and demand,

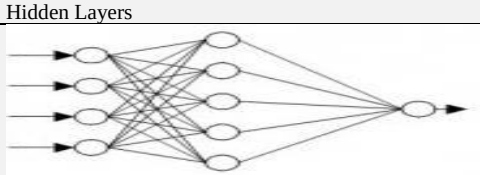
Input layers	Hidden Layers	Output
Offer Demand Consumption index Production Reserve		Price

Fig. 10: Schematic Diagram of A Multilayer Perceptron Illustrating Cost Prediction from the Observation of Criteria Involved in Determining Cost.

Cost ($pricei$) is the i-th observation of the price variable.
Offer ($offeri_{,1}$) is the i-th observation ofthe offer variable.
Demand ($demandi_{,2)\ is}$ the i-th observation of the variable demand.
Consumption index ($consoi_index_{,2}$) is the i-th observation of the consumption index variable.
Reserve (reservei,3) is the i-th observation of the reserve variable.

- Determination of parameters

Architecture: our network will be made upof 4 layers. The first layer will have 4 neurons, the second layer 5 neurons, the third layer 7 neurons and the last layer 1 neuron.
Sum function: the operator which combines the Wi weights is the sum.
Transfer function: the function to beapplied to the result of the summationfunction is the sigmoid.

$$g(x) = \frac{1}{1 + e^{-x}}$$

With g(0) = 0.5 $g(-\infty) = 0$ and $g(+\infty) = 1$

Learning phase Learning consists of iteratively descending the network, adjusting the weights at each pass according to the error calculation, untilthere are no more

Improvement. To achieve this, an errorback propagation algorithm is used.
Back propagation error algorithm
This algorithm is based on minimizing thesquared error between the calculated (actual) and desired (desired) outputs. Theterm gradient back-propagation comes from the fact that the error calculated at the output is transmitted in the opposite direction to the input.

- Test phase, performance evaluation the performance of our network can be evaluated on the basis of various parameters.

Firstly, it is possible to assess the error between the desired values and the values calculated by the output layer ofthe neural network. To do this, we calculate the RMS (Root Mean Square) parameter, which adds up the squares of the errors for each neuron in the output layer and takes the square root of this sum divided by the number of neurons in the output layer.

4.5. Implementing the model

To implement our proposal, we are using tools that include the possibilities of machine learning and databases. Machine Learning enables us to carry out our prediction work on the basis of the data contained in our model and its sub-models. In terms of data storage, we are implementing our data model using SQL (structured query language) combined withthe prediction functions of the 'MindsDb'tool. We are using the possibilities offered by mindsdb for the prediction of data based on time series, as our main data model includes information that allows us to make predictions according to time (year in our case).
The framework used, combining machine learning, databases for storage and prediction is "MINDSDB"[15] .
Our implementation procedure is as follows:

- Database creation
- Predictor creation and training
- Implementation of the mathematical model
- Model simulation and training with linear regression and neural network algorithms.

4.5.1. Creation of the storage database and its structure

The structure of our database contains the following elements (quantities requested, quantities produced, residual quantities, consumption indices and sales prices to be predicted)
Syntax for creating the database.
CREATE DATABASE data-source_name[WITH] [ENGINE [=] engine_name] [, PARAMETERS [=] { "key": "value",...}];
With :

- [data-source_name] = Identifier for the data source to be created
- [engine_string] = Engine to be selected depending on the database connection.
- PARAMETERS = {"key": "value"} object with the connection parameters specific for each engine.

Example:
CREATE DATABASE example_db WITH ENGINE = "postgres", PARAMETERS = {"user": "demo_user", "password": "demo_password","host": "3.220.66.106", "port": "5432", "database": "DB_PRED"};

4.5.2. Creation of our predictor and training

It uses the following syntax:
CREATE PREDICTOR mindsdb.[predictor_name] FROM [integration_name] (SELECT [sequential_column], [partition_column], [other_column],[target_column])
FROM [table_name])
*** Creation of the predictor PREDICT [target_column] ORDER BY [sequential_column]GROUP BY [partition_column] WINDOW [int] HORIZON [int].

Table 2: Explanations of Elements Contain in Sytax

Expression	Description
[Predictor_name]	Model name to create.
[integration_name]	Name of the integration created using the CREATE DATABASE statement or a data file
(SELECT [column_name, ...] FROM [table_name])	SELECT Declaration of selection of data to be used for training and validation.
PREDICT [target_column]	target_column is the column to be predicted.
ORDER BY [sequential_column]	The column by which time series is ordered. It can be a date or anything that defines the sequence of events.
WINDOW [int]	The number of rows to look back at when making a prediction. It comes after the rows are ordered by the column defined in ORDER BY and split into groups by the column(s) defined in GROUP BY. This could be interpreted as "Always use the previous 10 rows".
HORIZON [int]	It is optional. The number of future predictions (it is 1 by default).

This syntax contains the elements explained in the following table (table 1):
For example,
CREATE PREDICTOR mindsdb.sales_price FROM example_db
(SELECT * FROM cost.selling_price PREDICT selling_price ORDER BY year GROUP BY product_id WINDOW 20 HORIZON 7);
The model is trained automatically.

4.5.3. Implementation of our general mathematical model

According to economists, the price of a good depends on the law of supply and demand. To set up the model, we will opt for multiple regression at first. Next, we will use neural networks to enhance the efficiency of our forecasting model.
To do this, we consider price to be a dependent variable and demand, reserve, consumption index and supply to be the independent variables.

4.6. Results

After training our predictor, we make a prediction based on three algorithms evaluating 2 parameters each. Also, after creating the database for storing and training our models using two algorithms: linear regression and neural networks from which we calculate the correlation coefficients and RMS, we obtain the results in Table 2 below.
Comparison of the performance of our models
Several criteria are used to test performance of our different prediction models. among these we present the correlation coefficient and the RMS.

Table 1: Comparison of the Performance of Our Models

Algorithms	Results	
	Correlation coefficient	RMS
Linear regression	0.785	357.096
Neural networks	-0.239	856.529

Based on the analysis of these results, we can say that linear regression produces much better results than neural networks, as it has a lower RMS than neural networks.

5. Discussion

In our work, we call "prediction axis" any element that gives us an idea of the evolution of a parameter (the fact). Under these conditions, our model (figure 3) as presented can enable us to make monodimensional predictions (one dimension only) and/or to deduce multidimensional predictions (several dimensions). These different predictions are also free of duplicates, given the presence of atomic identifiers for the attributes present and used by the model. The use of modeling techniques derived from business intelligence enables us to overcome some of the rigid constraints imposed by conventional modeling techniques. Indeed, business intelligence modeling techniques require that data sources be well known, in order to better calibrate data storage. In our case, the data sources could be the databases of agricultural organizations (state structures, cooperatives, ministries of agriculture and other agricultural umbrella organizations such as the ICCO). The adoption of our models for the collection, analysis and prediction of raw material costs could reduce and eliminate any difficulties in improving the living conditions of producers in countries that have focused their development on agriculture. Our proposal would also contribute to improving their GDP (Gross Domestic Product) growth. On a technical level, our proposal could speed up cost prediction work in the agricultural sector.

6. Conclusion and future work

The aim of the present work is to set up models to improve the prediction of raw material costs. To do this, we reviewed the most representative works to our knowledge. Of these, we have identified the limitations and made our own contribution, which consists in adopting data models that enable both monodimensional and multidimensional analysis. To these models, which manage the storage of the data to be analyzed, we have associated mathematical models to ensure the computational and algorithmic aspect of our work. Future work on implementing these models could turn to the use of machine learning technologies (in terms of algorithms) and datamining (for storing the data to be analyzed). In fact, these technologies would enable several algorithmic implementations to be used to compare prediction results

References

[1] Baumann, P., Misev, D., Merticariu, V., & Huu, B. P. (2021). Array databases: concepts, standards, implementations. *Journal of Big Data*, *8*(1), 1-61 https://doi.org/10.1186/s40537-020-00399-2.

[2] Chee, T., Chan, L. K., Chuah, M. H., Tan, C. S., Wong, S. F., & Yeoh, W. (2009). Business intelligence systems: state-of-the-art review and contemporary applications. In *Symposium on progress in information & communication technology* (Vol. 2, No. 4, pp. 16-30).

[3] Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques : application of big data. *Journal of Big data*, *4*(1), 1-15. https://doi.org/10.1186/s40537-017-0077-4.

[4] Sajid, S., Haleem, A., Bahl, S., Javaid, M., Goyal, T., & Mittal, M. (2021). Data science applications for predictive maintenance and materials science in context to Industry 4.0. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.357.

[5] Tripathy, A. K., Adinarayana, J., Sudharsan, D., Merchant, S. N., Desai, U. B., Vijayalakshmi, K., ... & Tanaka, K. (2011, December). Data mining and wireless sensor network for agriculture pest/disease predictions. In *2011 World Congress on Information and Communication Technologies* (pp. 1229-1234). IEEE. https://doi.org/10.1109/WICT.2011.6141424.

[6] Baptiste, J. L. (2009). Merise Guide pratique : Modélisation des données et des traitements, langage SQL. Editions ENI.

[7] He, B., & Yin, L. (2021). Prediction Modelling of cold chain logistics demand based on data mining algorithm. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/3421478.

[8] Cui, Y. (2021). Intelligent recommendation system based on mathematical modeling in personalized data mining. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/6672036.

[9] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, *47*(4), 547-553. https://doi.org/10.1016/j.dss.2009.05.016.

[10] Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, *55*(1), 359-363. https://doi.org/10.1016/j.dss.2012.05.044.

[11] Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications*, *39*(12), 11303-11311. https://doi.org/10.1016/j.eswa.2012.02.063.

[12] Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management*, *54*(5), 758-790. https://doi.org/10.1016/j.ipm.2018.01.010.

[13] Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (2018). A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, *9*(3), 365-378. https://doi.org/10.1016/j.asej.2016.01.012.

[14] Nam, K., Kim, S. S., Park, C. S., Nam, T. Y., & Lee, T. Designing ML-based Approximate Query Processing Services on Time-Varying Large Dataset for Distributed Systems.

[15] https://mindsdb.com/integrations

[16] https://www.verteego.com/technologie accessed on 10/10/20212.
[17] https://ledatascientist.com/arima accessed on 10/10/2022
[18] Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K., & Bakirtzis, A. G. (2016, avril). Comparaison des modèles SARIMAX, SARIMA, SARIMA modifié et ANN pour la prévision à court terme de la production photovoltaïque. En2016, Conférence internationale sur l'énergie de l'IEEE (ENERGYCON) (pp. 1-6). IEEE.