International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I9P214

Volume 71 Issue 9, 157-161, September 2023

Original Article

Multiple Regression Model for Predicting New COVID-19 Cases in Ivory Coast

Assie Brou Ida[1], N'Guessan Béhou Gérard[2]

[1]*UFRMI, Félix Houphouët-Boigny University, Abidjan, Ivory Coast.*
[2]*Data Base, Ivory Coast Virtual University, Abidjan, Ivory Coast.*

[1]*Corresponding Author : ida_as09@yahoo.fr*

Received: 21 June 2023 Revised: 03 August 2023 Accepted: 16 August 2023 Published: 03 September 2023

Abstract - *In recent years, the COVID-19 pandemic has shaken the whole world, particularly our country, Côte d'Ivoire, which has had certain barrier measures imposed by the state in its response to this disease. The appearance of new cases of contamination justifying these measures was an indicator of whether the government had stopped (abolished) them. The prediction of new contamination cases has, therefore, contributed to a framework of barrier measures to allow the population to go about their daily activities more easily. In this study, we describe the approach to implementing a model for predicting new contamination cases using machine learning algorithms. The involvement of expertise and its importance in the interpretation of the results obtained are highlighted.*

Keywords - *Machine learning, COVID-19, Multiple regression, Decision support.*

1. Introduction

In recent years, the COVID-19 pandemic has shaken the whole world, especially our country, Côte d'Ivoire, where the state has imposed certain barrier measures in response to this disease. The appearance of new cases of contamination justifying these measures represented an indicator of the cessation (abolition) of these by the state. Therefore, predicting new contamination cases has contributed to a framework of barrier measures to allow the population to go about their daily lives more easily. As a result, several collaborative studies have been conducted between scientific researchers and health workers. Some studies have been oriented towards understanding the disease and its evolution [1, 2], and others to help in decision-making [3, 4]. One of the issues surrounding this assessment indicator is the prediction of the number of contamination cases. This high number of contamination cases negatively influences the country's economic development in most cases. This makes it difficult to master the proposed measures. In view of all the above, it is important to consider the factors behind this problem. This work proposes to train a predictive model based on Machine Learning to approximate the number of contamination cases and understand the factors behind this indicator. It is a question of predicting the number of cases of contamination in order to help in decision-making. This model uses multiple regression to establish a relationship between the number of contamination cases (dependent variable) and several independent variables (factors).

To do this, it will first be necessary to present state of the art on machine learning to show their contribution to the production of knowledge; Then, to make a modeling of the situation; and finally, the simulation and discussion from a case study.

2. State of the Art on Machine Learning

In 2020, at the very beginning of the global coronavirus pandemic, several barrier measures were taken, including confinement, distancing, wearing a mask, etc. According to the Ministry of Health, Public Hygiene and Universal Health Coverage, these measures taken at that time were urgent and necessary for an effective response to the disease (official website of the ministry). Although this has yielded conclusive results, there is also some caution to observe. Hence the intensification of vaccination campaigns. Also, it should be remembered that this pandemic has favored producing a lot of COVID-19 data. Harnessing knowledge from this data can be used to create intelligent applications and help both in decision-making and their understanding of the Ivorian people. For example, in estimating the length of hospital stay at the time of patient admission, relevant data have been used for improved organization of care and better planning of activities [5-8]. In the field of energy and the environment, the authors used the data to predict the emergence of unsafe modes of operation from the point of view of security of supply before they manifest themselves [9, 10].

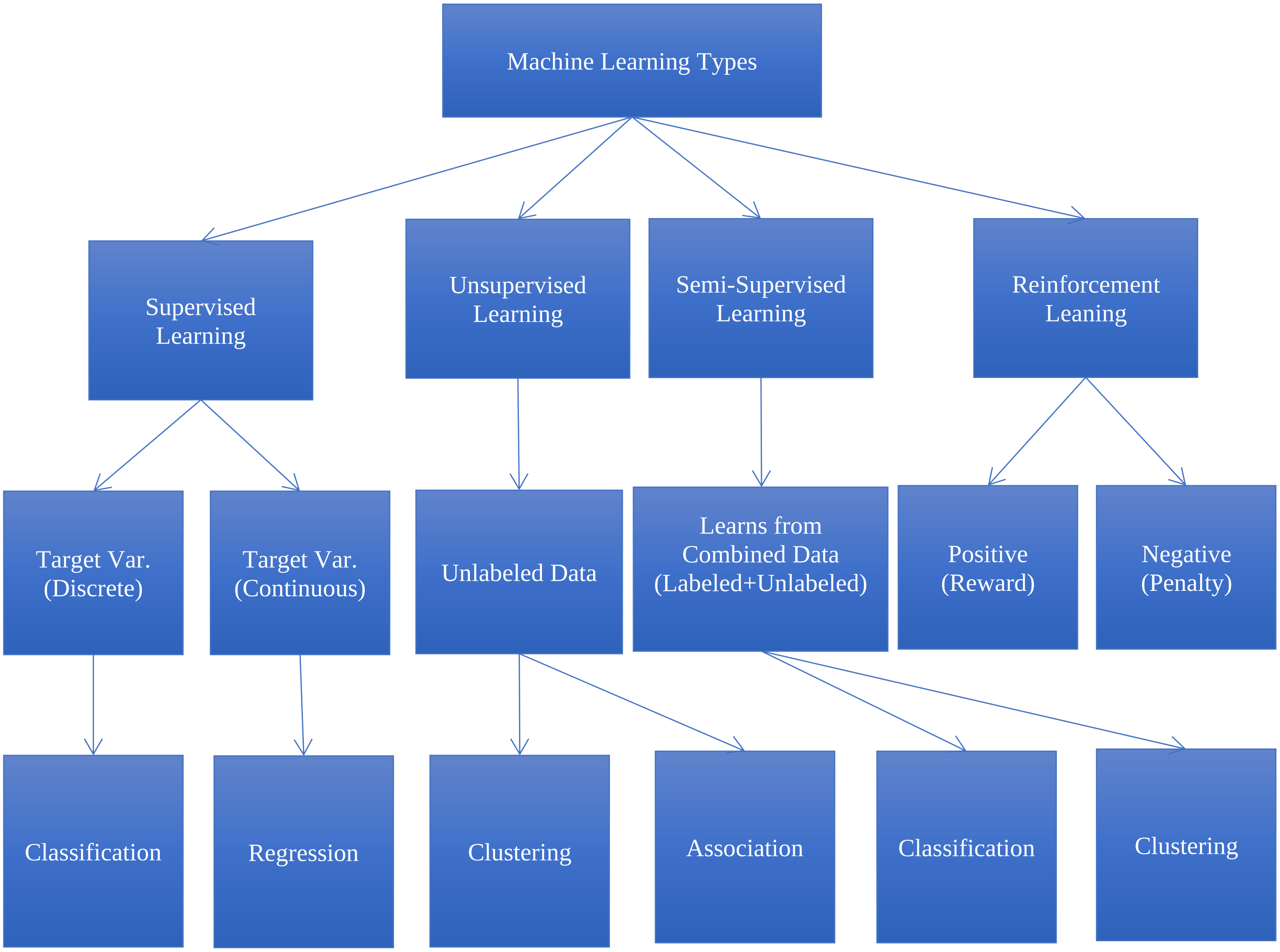

Fig. 1 Different types of machine learning techniques [13]

The authors in [11, 12] did a comparative study of linear and multiple regression models. The multiple regression model, according to these authors, represents a solid prediction model. In addition, several categories of learning algorithms can be classified into four groups, as shown in Figure 1. Each of these techniques is applied to a specific data type (structured, semi-structured, or unstructured data). The decision-making process proposed by these ML algorithms requires following an approach. It consists of data collection, data preprocessing, model construction, choice of algorithm for training and accuracy measurement of the model by an expert evaluation in the field.

3. Modeling

In this article, the data we worked on came from data on COVID-19 (extracted from the period from 2020-03-11 to 2021-05-29) from the official website of the Ivorian government. This is a file that contains six fields relating to the number of cases (total cases, retests, new cured, new deaths, new cases) per day. We will propose a model that predicts the number of new one-day cases from its numbers (new analyses, new cured, new deaths).

Let's consider m the number of lines in the data file and E_m the data set in the file. E_m can be written as:

$$E_m = \{ (x^1, y^1), (x^2, y^2), (x^3, y^3), \dots, (x^m, y^m)\}$$

$$\text{with } x^i = \begin{pmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_{n-1}^i \\ x_n^i \end{pmatrix}, \forall\, 0 < i < m. \qquad (x^i, y^i),$$

represents the i^{th} line of the data set E_m where, x^i represents the i^{th} input of X and y^i,the i^{th} output of Y.

Due to the nature of the variable y to be predicted from n (explanatory) variables $x^i_{j,\ \forall\, 0<i<m, \forall\, 0<j<n}$ of X, the prediction model is a multiple regression model.

The set of inputs X of E_m can, therefore, be written in the following matrix form:

$$X = \begin{pmatrix} 1 & x_1^1 & x_2^1 \dots & x_n^1 \\ \vdots & \vdots & \vdots \dots & \vdots \\ 1 & x_1^m & x_2^m \dots & x_n^m \end{pmatrix}$$

where,

$$x^i = \begin{pmatrix} x_0^i \\ x_1^i \\ \vdots \\ x_{n-1}^i \\ x_n^i \end{pmatrix} \text{ et } x_0^i = 1, \forall\, 0 < i < m.$$

The prediction model leads us to construct a regression hypothesis function which associates each x^i of X, 0<i<m, with a unique value y^i, i.e.:

$$\forall (x^i, y^i)_{0<i<m} \in E_m, f(x^i) \approx y^i$$

In general, the hypothesis function of a multiple regression is defined by :

$$h_\theta(x^{(i)}) = \theta_0 + \theta^1 x_1^{(i)} + \theta^2 x_2^{(i)} + \dots + \theta^n x_n^{(i)}$$

Table 1. Model algorithm

Step	
0	***Start***
1	• Load the E_m data file.
2	• Prepare the data by separating the explanatory data from the dependent data (data to be predicted) (x^i, y^i), $\forall\, 0 < i < m$.
3	• If the data is not significant, then the data is rejected
4	• Build the regression model
5	• Calculate the various model parameters
6	• Teach the model to the prediction machine • Prediction of new cases
	End

In our study, this function, denoted f, is given by :

$$f(x^i) = \sum_{k=0}^{n} a_k\, x_k^i + \varepsilon,$$

$\forall\, 0 < i < m$ and a_k, the coefficients or parameters of f and ε, the error.

We can rewrite f in vector form as follows:

$$f(x^i) = A^T x^i + \varepsilon, \text{ with } A = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} \text{ and } x^i = \begin{pmatrix} x_0^i \\ x_1^i \\ \vdots \\ x_{n-1}^i \\ x_n^i \end{pmatrix},$$

$\forall\, 0 < i < m.$

The table below shows the model algorithm, i.e. the algorithm for predicting the number of new cases based on new analyses, new cured cases and new deaths.

4. Results and Discussion

The numerical simulation of our model was based on COVID-19 data (extracted from the period 2020-03-11 to 2021-05-29). The dependent variable we want to predict is "New cases" from the explanatory variables ("new totals", "new analyses", "new cured", "new deaths").

With these explanatory variables obtained, we construct our model by determining its parameters or coefficients. Using the sklearn library of the Python software, we obtain the following result for the model coefficients from the training and test data sets:

```
Intercept:
[1.4909648]
Coefficients:
[[ 0.04430757  0.2598412  19.782908 ]]
```

Fig. 2 Datasets

The model can be written as

$$f(x^i) = 1.4909 + 0.0443\, x_1^i + 0.2598\, x_2^i + 19.7829\, x_3^i + \varepsilon,$$

$\forall\, 0 < i < m.$

The results highlighted in Figure "Fig. 3" are important elements in the conclusions of the proposed model. Indeed, subject to certain conditions met by specialists in the field, our model provides an accurate result at 46%.

These results can be interpreted as follows: A new analysis carried out and a new case cured cannot lead to a new case of contamination of COVID-19. This is not the case for a new case of death, which could lead to 19 new cases of contamination of COVID-19.

A test of the model with new entries for the explanatory variables gives the result shown in Table 2.

The increase in the number of death cases accentuates the number of cases of contamination with COVID-19. Hence, the restrictive measures taken by the state in the event of the loss (death) of a close relative due to COVID-19 to avoid contamination and the intensification of the vaccination campaign to protect oneself.

Table 2. Prediction of new cases

	Variables Explicatives			Dependent Variables
N Test	<< New analyzes >>	<< New healed >>	<< New deaths >>	<< New cases >>
1	0	0	0	1
2	50	0	0	3
3	5000	500	0	352
4	5000	500	10	550
5	5000	0	10	420
6	0	0	10	199
7	0	500	30	724
8	100	100	100	210
9	100	10	10	206
10	100	0	100	1984

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Nouveaux cas   R-squared:                       0.462
Model:                             OLS   Adj. R-squared:                  0.459
Method:                  Least Squares   F-statistic:                     124.7
Date:                 Fri, 10 Jun 2022   Prob (F-statistic):           2.78e-58
Time:                         14:30:05   Log-Likelihood:                -2595.8
No. Observations:                  439   AIC:                             5200.
Df Residuals:                      435   BIC:                             5216.
Df Model:                            3
Covariance Type:             nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  1.4910      7.042      0.212      0.832     -12.349      15.331
Nouvelles analyses     0.0443      0.004     10.070      0.000       0.036       0.053
Nouveaux guéris        0.2598      0.043      6.072      0.000       0.176       0.344
Nouveaux décès        19.7829      4.260      4.644      0.000      11.410      28.156
==============================================================================
```

Fig. 3 Model learning result

5. Conclusion

In this work, we have proposed a decision-making approach for predicting new cases of COVID-19 contamination. This approach was modelled using a multiple regression model.

The model was machine-learned using the sklearn module of the Python language. Our future work aims to extend the possibilities of data analysis or data processing of data sources to optimise the decisions taken in a specific domain.

References

[1] Rob Wallace et al., "Le Covid-19 et les Circuits du Capital," *HAL Open Science*, 2020. [Google Scholar] [Publisher Link]

[2] Benjamin Rader et al., "Mask-Wearing and Control of SARS-CoV-2 Transmission in the USA: A Cross-Sectional Study," *The Lancet Digit Health*, vol. 3, no. 3, pp. E148-157, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Li Yan et al., "An Interpretable Mortality Prediction Model for COVID-19 Patients," *Nature Machine Intelligence*, vol. 2, pp. 283–288, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Alaamujallad, and Haithamkhoj, "Is there Hope for the Hajj? Using the SIR Model to Forecast COVID-19 Progression in the City of Makkah," *SSRG International Journal of Economics and Management Studies*, vol. 9, no. 8, pp. 31-36, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Rachda Naila Mekhaldi et al., "Machine Learning in Predicting Hospital Lengths of Stay," *10th Francophone Conference in Management and Engineering of Hospital Systems, GISEH2020*, 2020. [Google Scholar] [Publisher Link]

[6] Pauline Alkhoury, MalekHejazie, and Firas Hussain, "The Prognostic Value of (NLR) Ratio in Patients with COVID-19," *SSRG International Journal of Medical Science*, vol. 9, no. 6, pp. 1-10, 2022. [CrossRef] [Publisher Link]

[7] Franck Jaotombo, "*Contributions of Machine Learning and Deep Learning Methods in the Prediction of Lengths of Hospital Stays and Re-Hospitalizations*," PhD Theses, Life Sciences [q-bio], Aix Marseille University, 2022. [Google Scholar] [Publisher Link]

[8] Arshiya Moin, "Artificial Intelligence Vs Covid19," *SSRG International Journal of Computer Science and Engineering*, vol. 7, no. 5, pp. 5-7, 2020. [CrossRef] [Publisher Link]

[9] Sanchez Celine, "*Specification and Implementation of the Multi Criteria Decision Support System for Preventive Maintenance and Asset Management of the ESCOTA Motorway Company: The SINERGIE Project*," PhD Theses, National School of Mines of Paris, 2007. [Google Scholar] [Publisher Link]

[10] Rizka Malia, "Forecasting Indonesia's Life Expectancy During the Covid-19 Period 2021-2026," *SSRG International Journal of Economics and Management Studies*, vol. 8, no. 2, pp. 105-109, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Suvendra Kumar Jayasingh, Jibendu Kumar Mantri, and P. Gahan, "Comparison between J48 Decision Tree, SVM and MLP in Weather Forecasting," *SSRG International Journal of Computer Science and Engineering*, vol. 3, no. 11, pp. 39-44, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[12] Iqbal H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 160, pp. 1-21, 2021.[CrossRef] [Google Scholar] [Publisher Link]

[13] Smita Rath, Alakananda Tripathy, and Alok Ranjan Tripathy, "Prediction of New Active Cases of Coronavirus Disease (Covid-19) Pandemic Using Multiple Linear Regression Model," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1467-1474, 2020. [CrossRef] [Google Scholar] [Publisher Link]