African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)
www.abjournals.org

PROGNOSIS METHOD ON THE OUTCOME OF COVID-19 PATIENTS IN SENEGAL

Cheikh Tidiane Seck[1*], Ibrahima Faye[1], Aba Diop[1], Mouhamed Amine Niang[1], Seydou Nourou Sylla[2], Abdourahmane Ndao[1] and Idrissa Sy[3]

[1]Department of Mathematics, Alioune Diop University, Bambey, Senegal

[2]Department of Information Technology and Communication, Alioune Diop University, Bambey, Senegal

[3]Department of Research (Biostatistics), Le Dantec Hospital, Dakar, Senegal

[*]Corresponding author: cheikhtidiane.seck@uadb.edu.sn

Cite this article:

Seck C.T., Faye I., Diop A., Niang M.A., Sylla S.N., Ndao A., Idrissa S. (2023), Prognosis Method on the Outcome of Covid-19 Patients in Senegal. African Journal of Mathematics and Statistics Studies 6(3), 93-103. DOI: 10.52589/AJMSS-VGAF69PO

Manuscript History

Received: 10 May 2023

Accepted: 27 July 2023

Published: 1 Aug 2023

ABSTRACT: *There have been disturbing waves of Covid-19 deaths in many countries due to a lack of adequate treatment in the early stages of the pandemic but also to the presence of co-morbidities in many hospitalised patients. This work aims to determine the discriminatory features between the surviving patients and the deceased ones after their hospitalisation to propose a new method of prognosis on the outcome of a new patient under treatment. To this end, we use three supervised classification methods, each allowing us to select the most significant features associated with patient death. These are binary logistic regression (BLR), random forests (RF), and support vector machines (SVM). The data comes from the Ministry of Health and Social Action of Senegal and covers the period from March 2020 to December 2022. Age emerged as the most discriminatory factor between the two patient groups: survivors and deceased. The study found that patients 60 and older are more likely to die of Covid-19.*

KEYWORDS: Covid-19 Patients, Supervised Classification, Outcome Prognosis, Patient's Features.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)

www.abjournals.org

INTRODUCTION

Covid-19 is a major public health issue that has affected all of humanity. Its rapid spread around the world has disturbed the health systems of all countries, developed as well as not developed. The coronavirus (SARS-CoV-2) continues to circulate, despite the existence of Vaccines. In Senegal, during the period from March 2, 2020, to December 31, 2022, the Centre of Health Emergency Operations (COUS) recorded 14,984 patients hospitalised for Covid-19. Among them, 138 patients died, leading to a mortality rate of 0.92%.

Since the first wave, the follow-up of Covid-19 patients in Senegal has been carried out at the epidemic treatment centres (CTE) implanted nationwide. During the treatment, different clinical variables were observed on patients and the patient's age, sex, and outcome (alive or deceased) after the hospitalisation. Specifically, the following variables were observed for each patient: *age, sex, province, medical history (fever, chill, sore throat, cough, rhinorrhea, dyspnea), pre-existing medical history (Comorbidity), travel abroad in the 14 days preceding the onset of symptoms (Travel14), visit a health facility in the 14 days preceding the onset of symptoms (Visit14), and outcome.*

The dataset is divided into two groups: survivors and deceased patients. Our aim is to identify the characteristics (or features) of the patients who survived and the characteristics of those who died of Covid-19 to establish a reliable prognosis method on the outcome of a new patient under treatment.

Various statistical and Machine Learning methods have been used to predict Covid-19 mortality. For example, Zhou et al. (2020) used multivariate logistic regression to identify factors associated with Covid-19 mortality. Bonanad et al. (2020) highlighted in a meta-analysis of 611,583 Covid-19 patients that those aged 60 to 69 had a higher mortality risk. Abdulhameed et al. (2020) used three Machine Learning techniques to predict the outcome (Alive/Deceased) of hospitalised Covid-19 patients based on demographic, clinical and epidemiological characteristics. Ngomas et al. (2022) used binary logistic regression to describe epidemiological aspects and determine factors of poor prognosis (mortality) in Covid-19 patients admitted to the emergency room at Libreville's CHU.

There has been a lot of work on predictive factors for COVID-19 deaths. Molka et al. (2021) proposed a specific and detailed literature review. To summarise, age and co-morbidities have often been identified as significant risk factors associated with Covid-19 mortality; see, for instance, Asfahan et al. (2020), Tian et al. (2020), Yu et al. (2020), Shi et al. (2020) and references therein.

In this paper, we use three supervised classification techniques: binary logistic regression, random forests and support vector machines to determine the clinical and demographic characteristics that are most significantly associated with Covid-19 mortality. This allows us to characterise the hospitalised patients that are likely to die, and by dichotomy, those that are likely to survive, enabling thus to make a reliable prognosis on the patient's outcome.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)

www.abjournals.org

METHODOLOGY

Data Preparation

Our global database contains 14,948 patients, of whom only 138 died. This means that there is an imbalance between the two groups: « survivors » and « deceased » in the database. Thus the application of predictive methods for classification into these two groups requires resampling to balance the training data. Here, we use the *undersampling* method, which reduces the number of data items in the majority class (i.e. the survivors' group) to balance the ratio in the training sample. We experimented with three ratios: 50-50, 45-55 and 40-60. For each ratio, the training sample was constituted by keeping all 138 deceased patients present in the global database and then by adding the necessary number of survivors to achieve this ratio. Based on the misclassification rate, the three tables below show that the 40-60 ratio gives the best performance for all of the three classification methods used. This is why, in the sequel, we deal with this 40-60 ratio, which led to a sample size of 345 patients.

Ratio: 50-50 Sample size=276

Method	BLR	RF	SVM
Misclassification rate (%)	18.47	23.55	22.82

Ratio: 45-55 Sample size=307

Method	BLR	RF	SVM
Misclassification rate (%)	19.54	22.8	21.82

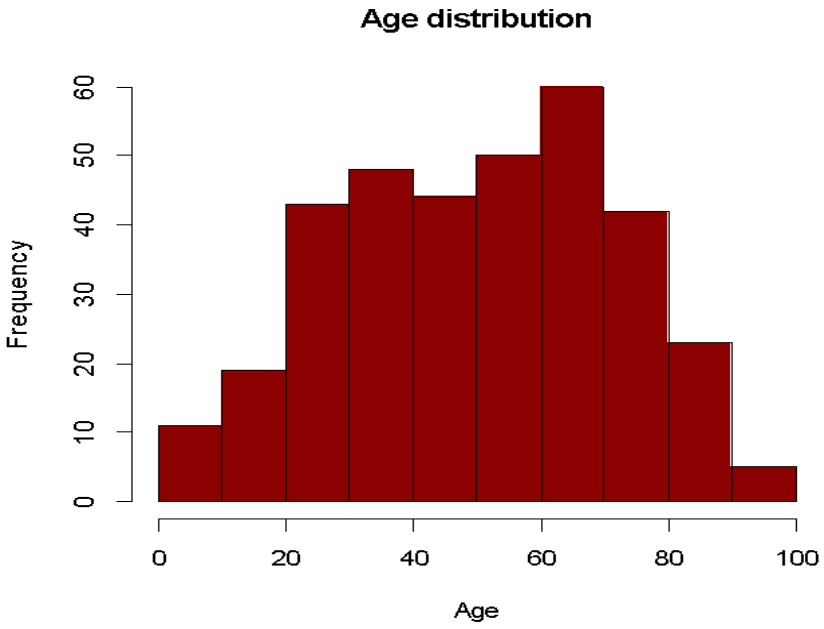
Ratio: 40-60 Sample size=345

Method	BLR	RF	SVM
Misclassification rate (%)	17.39	20.58	20.28

The target variable is the patient’s Outcome, which takes two categories: « Alive » and « Deceased »; the predictors are the eleven (11) other qualitative variables; age being considered here as a categorical variable because it is divided into classes (see histogram below). The variable « Province » was grouped into two modalities: « RM-Dakar » and « Other », because initially, it had many categories with very small counts.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)
www.abjournals.org



Description of the classification techniques

- ***Binary logistic regression***

Binary logistic regression is a statistical technique used to predict the occurrence or non-occurrence of an event using quantitative and/or qualitative explanatory variables $X = (X1, \ldots, Xp)$. It is based on the fundamental assumption that the probability $\pi(x)$ that the event occurs, given an observed value x of X, is a logistic function of the score $S(x)$, which is itself a linear combination of the explanatory variables. For a given observation x, the logit of $\pi(x)$ is defined by

$$ln\ ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = S(x).$$

Assuming that all the individuals for whom the event occurs form a group G1 and that all the individuals for whom the event do not occur form another group G2, then this observation x will be assigned to group G1, if $S(x) > 0$, i.e. $\frac{\pi(x)}{1-\pi(x)} > 1$ or $\pi(x) > 0.5$.

- ***Random forests***

They were introduced by Breiman (2001) to overcome the instability of « simple » decision trees. The random forest approach involves constructing a very large number of « simple » decision trees, and then aggregating them by averaging. Each decision tree is built on a bootstrap sample by randomly choosing, for each node, a subset of predictor variables to optimally divide this node. This method relies on two important parameters: the total number of trees in the forest (ntree) and the number of predictor variables (mtry) to be randomly selected at each stage of the construction of the trees (splitting a node).

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)

www.abjournals.org

- ***Support Vector Machines***

Support Vector Machines (SVM) are a Machine Learning technique that is used in both regression and classification problems. The SVM approach consists of finding a hyperplane in the space of characteristics (explanatory variables) that best separates the data into two groups in such a way that the distance of the hyperplane with the closest vectors is maximum. These closest vectors are called support vectors, and their distance from the hyperplane is the optimal margin. This method uses a kernel function to separate the two groups optimally. This kernel function can take several forms: polynomial, radial, etc., and its choice is crucial for the prediction results.

Performance measures

The performance of the classification techniques presented above can be assessed using measures derived from the confusion matrix, such as: *Accuracy*, *Precision*, *Sensitivity*, and *Specificity*, or using the ROC curve to calculate the Area Under the Curve (AUC).

Confusion matrix

Observed / Predicted	Deceased	Alive
Deceased	*TP*	*FP*
Alive	*FN*	*TN*

TP: All deceased patients that are predicted to have died;

TN: All surviving patients that are predicted to be alive;

FP: All surviving patients that are predicted to have died;

FN: All deceased patients that are predicted to be alive.

The following measurements are obtained by combining these four elements:

- *Accuracy* gives the percentage of well-classified patients

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

- *Precision* gives the percentage of « true positives » among all the patients predicted positives

$$Precision = \frac{TP}{TP + FP}.$$

- *Sensitivity* gives the percentage of « true positives » among all the patients observed positives

$$Sensitivity = \frac{TP}{TP + FN}.$$

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)

www.abjournals.org

- *Specificity* gives the percentage of « true negatives » among all the patients observed negatives

$$Specificity = \frac{TN}{TN + FP}.$$

RESULTS

We use data from COUS (Centre of Heath Emergency Operations), the agency responsible for collecting Covid-19 data from all around the country. Our training sample (obtained using the procedure presented in section 2.1.) comprises 345 patients whose cross-distribution according to the variable response *Outcome*, and each of the predictors is given in Table 1. The p-value of the Chi-square test of independence is also provided to assess the relationship between each of the predictors and the variable response *Outcome*.

Table 1: Cross-distribution of patients according to the outcome and different predictors

		Outcome			
Variables	Categories	Deceased(138)	Survivors(207)	Percentage	P-value
Age	0-20 20-40 40-60 60-80 80+	2 (1.4%) 5 (3.6%) 27 (19.7%) 76 (55.1%) 28 (20.3%)	18 (8.7%) 82 (39.6%) 61 (29.5%) 44 (21.3%) 2 (1%)	5.8% 25.2% 25.5% 34.8% 8.7%	2.2 e-16
Sex	M F	104 (75.4%) 34 (24.6%)	114 (55.1%) 93 (49.93%)	63.2% 36.8%	0.0392
Historical.fever	OUI NON	75 (54.3%) 63 (45.7%)	109 (52.6%) 98 (47.4%)	53.3% 46.7%	0.8428
Throat	OUI NON	6 (4.3%) 132 (95.7%)	53 (25.6%) 154 (74.4%)	17.1% 82.8%	6 e-7
Cough	OUI NON	57 (41.3%) 81 (58.7%)	99 (47.8%) 108 (52.2%)	45.2% 54.8%	0.4131
Rhinorrhea	OUI NON	9 (6.5%) 129 (93.5%)	59 (28.5%) 148 (71.5%)	19.7% 80.3%	1 e-6
Dyspnea	OUI NON	43 (31.1%) 95 (69.9%)	21 (10.1%) 186 (89.9%)	18.5% 81.5%	1.77 e-6
Comorbidity	OUI NON	14 (10.1%) 124 (89.9%)	13 (6.3%) 194 (93.7%)	7.8% 92.2%	0.1379
Travel14	OUI NON	1 (0.7%) 137 (99.3%)	4 (2%) 203 (98%)	1.4% 98.6%	0.6457
Visit14	OUI NON	11 (8%) 127 (92%)	4 (2%) 203 (98%)	4.3% 95.7%	0.4445
Province	RM-Dakar Other	133 (96.4%) 5 (3.6%)	203 (98%) 4 (2%)	97.4% 2.6%	0.5349

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)
www.abjournals.org

The majority of patients in the sample studied were men (63.2%), compared with women (36.8%). Patients aged between 20 and 40 represented 25.5% of the sample and accounted for 3.6% of deaths. Patients aged 80 and over represented only 8.7% of the sample but accounted for 20.6% of deaths. All the patients in this age category died except 2.

The 60-80 age group accounted for 55.1% of deaths and 21.3% of survivors, while the 40-60 age group accounted for 17.9% of deaths and 29.5% of survivors.

Most patients (82.8%) had no throat problems. Those with a sore throat accounted for only 4.3% of the total number of deaths.

45.2% of patients had a cough. Of these, 57 died, representing 41.3% of the total number of deaths in the sample.

Similarly, 19.7% of the total number of patients presented with rhinorrhea, but only 6.5% of deaths were related to this clinical sign.

Only 18.5% of the total number of patients presented with dyspnea, but 31.1% of deaths were associated with this symptom.

Binary logistic regression (BLR)

Table 2 gives the results of a backward stepwise logistic regression procedure, explaining the patient outcome variable, which led to a model where the variables significant at the 5% level were: Age, Throat, Cough, Rhinorrhea, Dyspnea and Comorbidity. However, Age and Dyspnea are also significant at the 0.1% level. This shows that these two variables are very strongly associated with the outcome variable. Hence, age older than 60 and the presence of dyspnea are highly significant risk factors for Covid-19 mortality. This is confirmed by the ORs (Odds ratios), which largely exceed 1 for these two variables.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)
www.abjournals.org

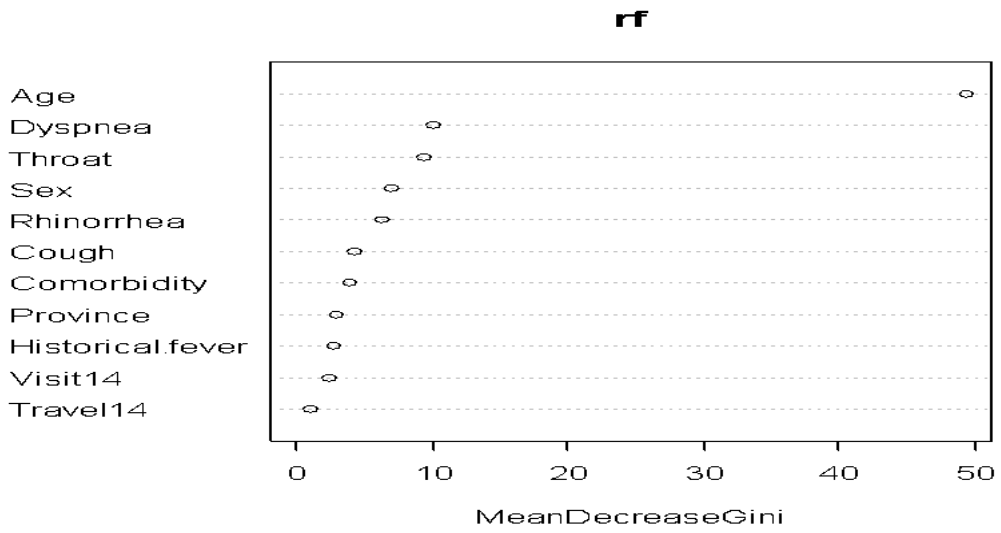
Table 2: Results of the logistic model (only significant characteristics have been retained)

Variables	Coefficients	OR (Odds ratio)	P-value
Age: 40-60	1.55	4.71	0.0540.
Age: 60-80	3.12	22.64	0.00017***
Age: 80+	3.47	32.13	0.00014***
Throat _YES	-1.49	0.22	0.0049**
Cough _YES	-0.89	0.41	0.0048**
Rhinorrhea _YES	-1.10	0.33	0.0207*
Dyspnea_YES	1.36	3.97	0.00043***
Comorbidity_YES	1.26	3.52	0.0243*

Random forests (RF)

Figure 1 shows the order of importance of the different predictors in the random forest model according to the "MeanDecreaseGini" measure. The most important variables for determining a patient's Outcome are: Age, Dyspnea, Throat, Sex, Rhinorrhea, and Cough.

Figure 1: Importance of variables in the random forest model



Support Vector Machines (SVM)

Here, the selection of the most important features for predicting patient outcome can be performed with the RFE-SVM algorithm, which uses cross-validation to estimate model performance with a subset of features. Based on the Accuracy and Kappa indices, Table 3 shows that the variables: Age, Dyspnea, Throat, Comorbidity, and Rhinorrhea are the most relevant features for predicting Patient Outcome with the SVM model.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 3, 2023 (pp. 93-103)

www.abjournals.org

Table 3: Importance of variables in the SVM model

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
1	0.7596	0.5035	0.06274	0.1268	
2	0.7399	0.4670	0.06620	0.1299	
3	0.7515	0.4829	0.06949	0.1402	
4	0.7609	0.5038	0.06986	0.1406	
5	0.7632	0.5075	0.07411	0.1491	
6	0.7677	0.5161	0.06826	0.1342	
7	0.7700	0.5166	0.07323	0.1504	
8	0.7752	0.5275	0.07079	0.1444	
9	0.7712	0.5196	0.07505	0.1539	
10	0.7718	0.5211	0.07175	0.1473	
11	0.7775	0.5337	0.06823	0.1391	*

The top 5 variables (out of 11): Age, Dyspnea, Throat, Comorbidity, Rhinorrhea.

Comparison of the performance of the three methods

Table 3 gives the performance measures (in per cent) of the three studied models and their Area Under the ROC Curve (AUC).

Table 3: Performance measures estimated from the sample

Method	*Accuracy*	*Precision*	*Sensitivity*	*Specificity*	*AUC*
RLB	82.60	77.85	78.98	85.02	84.8
RF	79.42	76.37	70.28	85.50	84.0
SVM	79.71	73.18	75.37	82.46	84.8

Binary logistic regression (BLR) performs the best: (Accuracy=82.6%, Precision=77.85%, Sensitivity=78.98%, Specificity=85.02% and AUC=84.8%). The rate of misclassification is respectively: 17. 40% for BLR, 20.58% for RF and 20.28% for SVM.

DISCUSSION

The aim of this study is to determine the discriminatory characteristics between two groups of Covid-19 patients after hospitalisation: survivors and deceased. It is a retrospective, analytic and cross-sectional study covering the period from March 2, 2020, to December 31, 2022. The average age of the patients observed was 56. This is in line with the literature, where the mean age of Covid-19 patients varies between 50 and 70 years; see, for example, Abdelhameed et al. (2020), who found a mean age of 53.1 years for Covid-19 patients hospitalised in Nigeria.

The three prediction models applied to the (randomly selected) training sample allowed us to identify the most significant factors associated with the Covid-19 patient's death:

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

- For the binary logistic regression model, the most significant variables are (in order): Age, Dyspnea, Throat, Cough, Rhinorrhea, and Comorbidity.
- For random forests, the most significant variables are (in order): Age, Dyspnea, Throat, Sex, Rhinorrhea, and Cough.
- For support vector machines, the most relevant variables are (in order): Age, Dyspnea, Throat, Comorbidity, and Rhinorrhea.

We can deduce from this that the most significant characteristics associated with Covid-19 death are: age (Age>=60), breathing difficulties (Dyspnea_YES), sore throat (Gorge_YES) and runny nose (Rhinorrhea_YES) in the first place, followed by the presence of cough (Cough_YES) and the existence of a pathological history (Comorbidity_YES). Finally, the variable Sex contributes to a lesser extent to the prediction of the Outcome variable. In fact, it was observed that men, who represented 63.2% of the studied sample, accounted for 75.4% of deaths and that the Chi-square test revealed that the sex variable was significantly related to patient outcome (p-value=0.0392); this means that men are more likely to die of Covid-19 than women.

In short, the different characteristics which are identified above and linked to Covid-19 death represent discriminatory factors, making it possible to distinguish, during their hospitalisation, patients who will succumb from patients who will survive the disease. This gives a new method of prognosis on the outcome of a Covid-19 patient under treatment and enables prioritising patients with a bad prognosis (death).

CONCLUSION

The aim of this study was to determine discriminatory characteristics between deceased patients and survivors in order to establish a reliable prognosis method on the outcome of a new Covid-19 patient undergoing treatment. This method consists of combining three predictive models: binary logistic regression, random forests and support vector machines, which identified, respectively, the most significant factors associated with Covid-19 death. Age older than 60 appeared to be a highly significant factor in bad prognosis (death) for Covid-19 patients. Similarly, breathing difficulties and runny nose were also found to be significant factors in bad prognosis for Covid-19 patients.

Identifying these discriminatory characteristics can help the medical staff to guide the therapeutic treatment of patients better, thereby reducing Covid-19-related mortality and rationalising the often limited resources.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO

REFERENCES

[1]. Abdulhameed A.O, Mannir A, Usman M, Auwalu I, Lawan A, Ahmad A. S, Hassan S. A, Safiya S.S, Hussaini G, Dikko and Muftahu Z.R, « A Classification Approach for Predicting COVID-19 Patient's Survival Outcome with Machine Learning Techniques ». medRxiv preprint doi: https://doi.org/10.1101/2020.08.02.20129767.
[2]. Asfahan S, Deokar K, Dutt N, Niwas R, Jain P, Agarwal M. Extrapolation of mortality in COVID-19: Exploring the role of age, sex, co-morbidities and health-care related occupation. Monaldi Arch Chest Dis 2020; 90:313–17.
[3]. Bonanad C, García-Blas S, Tarazona-Santabalbina F, et al. The Effect of Age on Mortality in Patients with COVID-19: A Meta-Analysis with 611,583 Subjects. J Am Med Dir Assoc 2020; 21:915–18.
[4]. Breiman, L. "Random Forests," Machine learning, vol. 45, p. 5–32., 200
[5]. Shi Q, Zhang X, Jiang F, et al. Clinical Characteristics and Risk Factors for Mortality of COVID-19 Patients with Diabetes in Wuhan, China: A Two-Center, Retrospective Study. Diabetes Care 2020; 43:1382–91.
[6]. Ngomas J.F, Ifoudji Makao A., Nze Obiang P.C, Nyangui D.E.M, Manga F., Bitegue L., Kombila U.D, Igala M., Ayo Bivigou E., Essola-Rerambiah L., Sima Zué A.. « Aspects Epidémiologiques et Facteurs de Mauvais Pronostic des Patients Atteints de COVID-19 Admis en Réanimation au Centre Hospitalier Universitaire de Libreville au Cours des Trois Premières Vagues de la Pandémie ». Health Sci. Dis: Vol 23 (6) June 2022 pp1-7.
[7]. Guyon, I. and Weston, J. and Barnhill, S. and Vapnik, V. Gene selection for cancer classification using support vector machines. Machine learning, 46,2022, pages 389-422.
[8]. Molka O. et al. (2021). Facteurs prédictifs de mortalité liée à la Covid-19 : Revue de la littérature. Rapport de l'Observatoire National des Maladies Nouvelles et Emergentes, Ministère de la Santé Tunisienne.
[9]. Uddin S., Khan A., Hossain M. E. and Ali M.M, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no.281, 2019.
[10]. Tian W, Jiang W, Yao J, et al. Predictors of mortality in hospitalised COVID-19 patients: A systematic review and meta-analysis. J. Med. Virol. 2020. doi:10.1002/jmv.26050.
[11]. Yu C, Lei Q, Li W, et al. Clinical Characteristics, Associated Factors, and Predicting COVID-19 Mortality Risk: A Retrospective Study in Wuhan, China. Am J Prev Med 2020; 59:168–75.
[12]. Zhou F., Yu T., Du R., G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li,. X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen and B. Cao, "Clinical course and risk factors for mortality of adult in patients with COVID-19 in Wuhan, China: a retrospective cohort study," Lancet, vol. 395, p. 1054–62, 2020.

Article DOI: 10.52589/AJMSS-VGAF69PO
DOI URL: https://doi.org/10.52589/AJMSS-VGAF69PO