International Journal of Computer Science and Software Engineering (IJCSSE), Volume 9, Issue 5, May 2020
ISSN (Online): 2409-4285 www.IJCSSE.org Page: 30-34

Prediction of Mortality Rate of COVID-19 Patients using Machine Learning Techniques in Nigeria

Agbelusi, Olutola[1] and Olayemi Olufunke C.[2]

[1] Computer Science Department, Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria
[2] Computer Science Department Joseph Ayo Babalola University, Ikeji, Osun state, Nigeria

[1]tola52001@yahoo.com, [2]ocolayemi@jabu.edu.ng

ABSTRACT

Corona virus disease pandemic have highly destructive effects around the world and this virus has affected both developed and developing nations. In this paper, predictive model for the mortality rate of patients infected with corona virus in Nigeria using data mining techniques is developed. Oral interview was conducted with virologist at health institution (The Federal medical centre, Owo, Ondo state, Nigeria) to ask for some basic factors that causes mortality in infected corona virus patients. Online survey was done based on these ten basic factors and three hundred and two responses were collected and preprocessed. A ten fold cross validation
technique was used to partition the datasets into training and testing data in which predictive models were developed using data mining algorithms (Multilayer Perception, Naïve Bayes', Decision Tree and Decision Rule) . Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the models and the result shows that the four models developed have the capability to forecast mortality rate of corona virus adequately. Conclusively, multilayer perception has the highest level of performance with 85% accuracy. Multilayer Perception model is effective, reliable and is recommended to forecast the rate of mortality of patients infected with corona virus. Moreover, this prediction is important because the death of any patients is emotional and physically challenging to the morning families.

Keywords: *Corona Virus, Mortality, Pandemic, Data mining, Algorithms.*

1. INTRODUCTION

Corona Virus otherwise known as COVID is a chronic worldwide epidemic that causes sicknesses ranging from common cold to more severe diseases such as severe Acute Respiratory Syndrome. This pandemic originated from Wuhan in China and spread to nations of the world in which Nigeria is not exempted. Sadly, mortality among children and adults were recorded of which no one in Nigeria can forget in a rush. The total number of confirmed cases globally as at the end of May, 2020 has risen to 6,203,615 with a rising figure of confirmed cases in Nigeria of 9,855 which is devastating and needs to be urgently controlled hence, this study (WHO, 2020). This research is carried out during a lockdown process in Nigeria and we can only get responses by conducting an online survey. Four data mining techniques ((Multilayer Perception, Naïve Bayes', Decision Tree and Decision Rule) were used with the aim of adopting the one with the highest level of accuracy and Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the models.

The reviewed works of authors who had carried out researches on the prediction of mortality rate of other diseases has really assisted in carrying out this research.

Zhang *et al* (2020) presented a research work with the aim of showing more light on the risk factors that are linked with COVID infection, unimprovement and its death. The study work included all the 663 COVID patients that were admitted at the study clinic. Multivariate logistic regression analysis was used for the analysis. The result shows that the older patients having some underlying illness and some other risk factors were the patients that had severe COVID and with the highest mortality rates.

Zhou *et al* (2020) worked on a research with the aim of investigating the risk factor that are responsible with mortality and morbidity of COVID confirmed patients that were admitted and treated in two different Community Clinics. Univariable and Multivariable logistic regression methods were used to analyze the dataset extracted from the electronic medical records of the COVID patients. The result of this research work shows that older age, higher SOFA and elevated d-dimer of patients while on hospital bed were the risk factors for mortality with COVID patients.

Singh *et al* (2020) presented a paper work with the aim of using machine learning and Deep learning models to understand the daily increase in the spread of corona virus. The result shows a polynomial regression spread of the virus.

Svetoslav B. (2020) presented a paper work with the aim of using logistic growth model in evaluating the Epidemic of Corona virus in New York City. The result shows that the proposed model was not capable of

forecasting both long- and short-term Epidemic predictions.

Zhao *et al* (2020) worked on the Epidemic of Corona virus using exponential growth model. The signs exhibited by the travelers in and out Wuhan were compared with the existing Corona virus symptoms and adopted the development of the proposed model. The result shows that the model has the capability to analyze the spread of the virus. The limitation of the work is that, insufficient numbers of data were used.

Hamzaha *et al* (2020) developed predictive modelling for the purpose of predicting COVID-19, mortality and morbidity and their likely recoveries. Actual data enquiry and live pictures were used on their website. The inquired data was used for Susceptible-Exposed-Infections-Recovered (SEIR) Predictive modelling. The inquired information was classified into two parts (negative and positive opinions) so as to understand the impact of the information. The result of the study shows that COVID is still an uncertain infectious disease meaning that SEIR modelling can only be correct after the end of the epidemic.

Olayemi *et al* (2017) presented a paper titled '' the development of a predictive model for paediatric patients with lower respiratory tract infection using data mining approach''. The paper was motivated by the use of Naïve Bayes' classifier for predicting the risk of lower respiratory tract infections. Respiratory Rate, Temperature, cough and Cyanosis were used as the independent variables for the prediction of the infection. The result shows that the model used was suitable in carrying out the predictive task with minimum 92% accuracy. There was no feature selection method used to actually know the most relevant features for predicting the infection and also, only one machine learning algorithm was used for the prediction with just four variables.

Olayemi *et al* (2018) worked on a research that aims at using machine learning algorithms to diagnose the occurrence of Respiratory Infections patients alongside with some identified risk factors. Highly correlated features were attained from patient's records along side with the whole features that were present in the records of patients. Two machine learning techniques (Naïve Bayes' and K-nearest neighbor) were used alongside with java programming language for the implementation of the work. The performance of the models was evaluated based on accuracy, sensitivity, specificity and precision. The result shows that Naïve Bayes with information- based feature selection method performs better than the other method used in the research work.

2. RESEARCH METHODOLOGY

Extensive literature review on Corona virus mortality prediction and other related publications was carried. This section describes the dataset, data mining algorithm and the performances analysis. Ten basic factors were identified and online survey was done based on these factors in which three hundred and two responses were collected. These data was preprocessed and 10-fold cross validation technique was used to partition the datasets into training and testing data. Predictive models were developed using data mining algorithms (Multilayer Perception, Naïve Bayes', Decision Tree and Decision Rule) and Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the models.

2.1 A detailed description of each of the variables in the dataset is as follows:

a) Age: is a description of the min age (60 years and above proposed by Nigeria centre for disease control (NCDC), 2020) of the patient receiving treatment.

b) Underlying Illness: is a description of the presence of other infection like tuberculosis, cancer, asthma etc. The presence of these diseases reduces the chances of survival of corona virus infected patient and of a nominal value classified as yes or no – it is one of the input variables used in predicting the mortality of infected patient.

c) Respiratory Illness: is a description of the presence of other infection like tuberculosis, cancer, asthma etc. The presence of these diseases reduces the chances of survival of corona virus infected patient and of a nominal value classified as yes or no – it is one of the input variables used in predicting the mortality of infected patient.

d) No/Poor medical facilities: This is the description of no or poor availability of medical facilities to tackle this virus ravaging the world at large.

e) Lack of accepted vaccine: The description of no accepted vaccine to cure infected patient during the period of carrying out this research.

f) Superstitious belief of corona virus: Many infected patient did not belief that corona virus is existing and that the government of Nigeria is deceiving the entire nation.

g) Poverty: Description of financial status of infected patients and poor family background.

h) Weight: is a description of the weight of patient at the age of receiving medication; the weight is a numeric value measured in Kilograms (Kg). The

age and weight of each patient altogether are required in determining the nutritional status of each infected patient.

i) Treatment with wrong local herbs: This is the description of infected patients treating the virus with wrong local herbs.

j) Ignorance: Lack of knowledge of this virus despite the government effort to create awareness.

3. IMPLEMENTATION

3.1 Naïve Bayes' prediction model

Bayesian classification provides a useful perspecti-ve for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.
The Bayes' Theorem:

$$P(\text{Data}|\text{Class}) = \frac{P(\text{Class}|\text{Data}) * P(\text{Data})}{P(\text{Class})}$$

Where
P(Class): Prior probability of class
P(Data): Prior probability of training data
P(Class|Data) : Probability of Class given the data
P(Data|Class) : Probability of Data given the class.

After using the Naïve Bayes' Classifier to train the data and test the model developed using 10-fold cross validation, the level of accuracy is 77%. Figure 1 shows that 55 data were correctly classified as Yes and 11 were correctly classified as No. Also, 12 were misclassified as Yes and 22 as No.

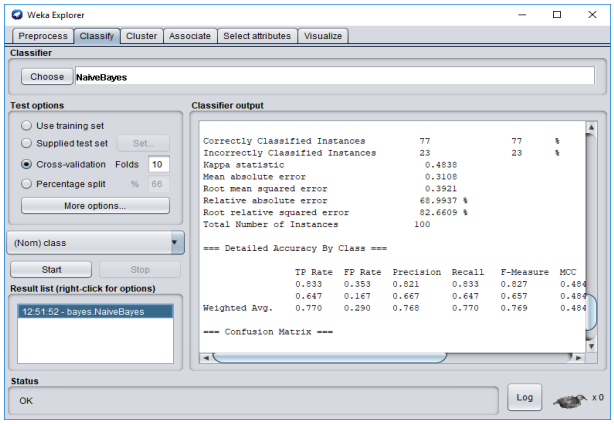

Fig. 1. Naïve Bayes' Model

3.2 Multilayer Perceptron Model

Multilayer perceptron is divided into three layers, the input, hidden and output layer.
The mathematical model for the artificial neural network in figure 3.4 is as follows:

THE INPUT LAYER– at this part of the multi-layer perception (MLP) the input values are entered into the MLP system and the weights, w_i of each input, x_i are applied after which the summation, U_k is sent to the hidden layer for the activation function, to take effect.

THE HIDDEN LAYERS– at this part of the MLP the summation of the input variables are all sent to the activation function which is fired through all the hidden layers (for the purpose of this research 20 layers were used) and finally producing the required output variable, y_k which determines whether the survival is YES/NO.

$$y_k = \varphi(v_k) = \varphi(u_k + b_k), \quad y_k = \pi r^2$$
$$\varphi(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + b_k$$

Where the activation function is:

$$\varphi(v) = \frac{1}{1 + e^{-av}}$$

THE OUTPUT LAYER– the value of the output (Survival) is determined with the error rate as low as possible. Also, the back-propagation algorithm is applied which tries to reduce the error rate, of the model via gradient descent. At iteration n (the nth row in the training set), the error for neurons in the output layer is calculated as:

$$e_j(n) = d_j(n) - y_j(n)$$
$$v_j = \sum_{i=0}^{m} w_{ji}(n) y_i(n)$$
$$y_j = \varphi_j(v_j(n))$$

After using the Multilayer Percepton Classifier to train the data and test the model developed using 10-fold cross validation, the level of accuracy is 85%. Figure 2 shows that 59 data were correctly classified as Yes and 7 were as No. Also, 8 were misclassified as Yes and 26 as No.

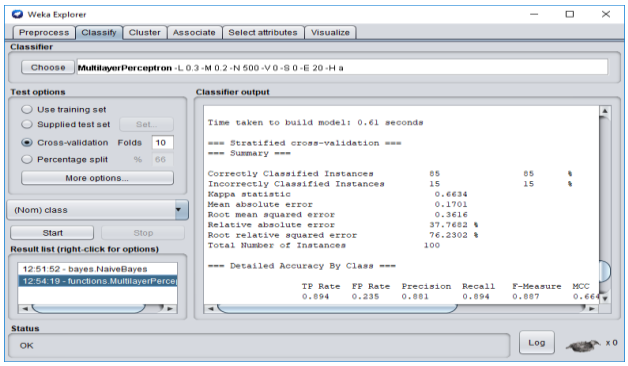

Fig. 2. Multilayer Perception

3.3 Decision Rule Model

After using the Decision Rule Classifier to train the data and test the model developed using 10-fold cross validation, the level of accuracy is 77%. Figure 1 show that 54 data were correctly classified as Yes and 12 as No. Also, 11 were misclassified as Yes and 23 as No

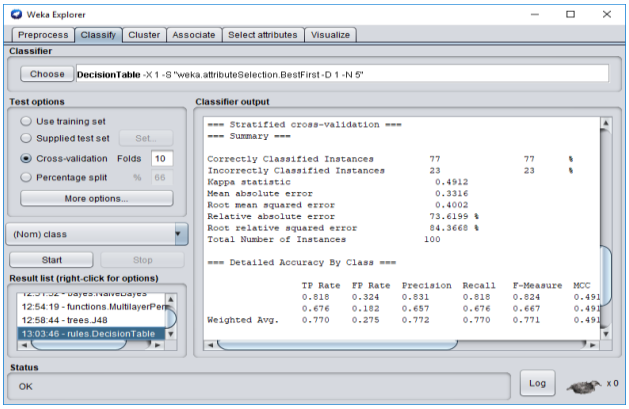

Fig. 3. Decision Rule Model

3.4 J48 Decision Tree Model

Decision tree is a learning model designed to predict the value of a target variable based on a number of input variables. After using the J48 Decision Tree classifier to train the data and test the model developed using 10-fold cross validation, the level of accuracy is 77%. Figure 4 shows that 53 data were correctly classified as Yes and 13 misclassified as No. Also, 10 were wrongly classified as Yes and 24as No

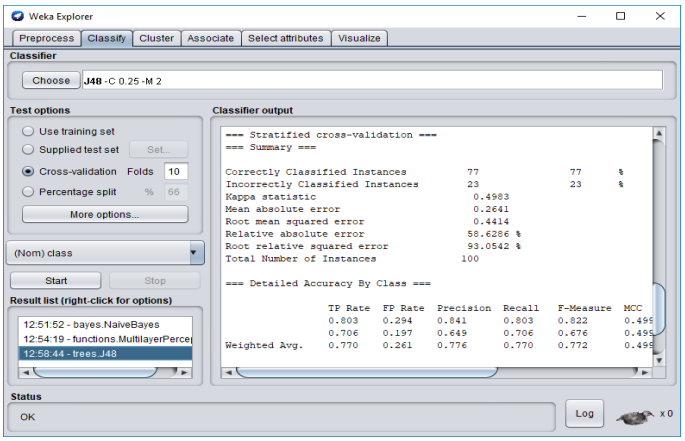

Fig. 4. J48 model

Table 1: Comparison of all four models on the datasets provided

Detailed Accuracy by Class	Naïve Bayes' Model	Decision Trees Model	Multilayer perception	Decision Rule
Accuracy	77%	77%	85%	77%
Correct Classification	77	77	85	77
Incorrect Classification	23	23	15	23
TP rate	0.770	0.770	0.094	0.770
FP rate	0.290	0.261	0.235	0.275
Precision	0.768	0.776	0.001	0.772
RMSE	0.3921	0.4414	0.3616	0.4002
MAE	0.3108	0.2441	0.1701	0.3316
RAE	46.9937%	59.6266%	37.7602%	73.6199%

Table 1 compared all three models developed and the report shows that Naïve Bayes', Decision trees and Decision rue have the same accuracy of 77% while Multilayer perceptron has an accuracy of 85. The root mean error of Naïve Bayes' is 46.9937%, a Decision tree is 59.6266%, Multilayer perceptron is 37.7602% and Decision rule is 73.6199%. This result shows that Multilayer perceptron has the lowest root absolute error. Multilayer perceptron is the most preferred based on the accuracy and error rate.

4. COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

Figure 5 below shows the comparative analysis of the four different kinds of data mining techniques used in this study. Multilayer Perceptron has the highest level of accuracy of 85% while Naïve Bayes', Decision Tree and Decision Rule have the same level of accuracy which is 77%

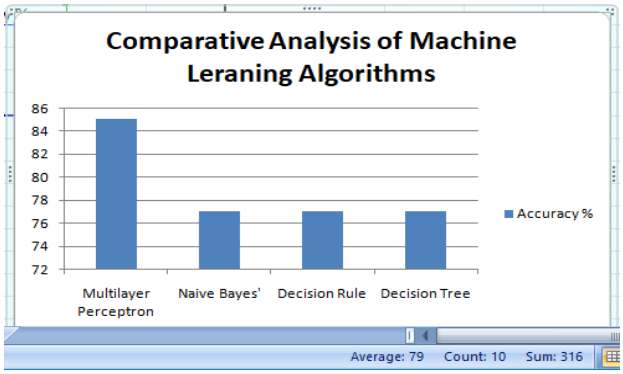

Fig. 5. Comparative Analysis of Machine Learning Algorithms

5. CONCLUSION

Conclusively, Multilayer Perception is effective and reliable and is recommended to forecast the rate of mortality of patients infected with corona virus. Moreover, this prediction is important because the death of any patients is emotional and physically challenging to the morning families.

REFERENCES

[1] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen and Bin Cao(2020) ''Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study'' Lancet 2020; 395: 1054–62 Published Online March 9, 2020 https://doi.org/10.1016/ S0140-6736(20)30566-3

[2] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. Corona Tracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Submitted]. Bull World Health Organ. E-pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695

[3] O.C. Olayemi,, O.O. Olasehinde and B.A Ojokoh (2017) "Development of a Predictive Model for Paediatric Patients with Lower Respiratory Tract Infection Using Bayesian Rule Approach", JABU Journal of Science and Technology, Vol. 3, N0.1, pp 92-101.

[4] Olufunke.C. Olayemi, Olumide .S. Adewale, Olayemi. O. Olasehinde, Bolanle A. Ojokoh, Adebayo. O. Adetunmbi, (2018) "Application of Machine Learning to the Diagnosis of Lower Respiratory Tract Infection in Paediatric Patients" Paper presented at the 2nd International Conference on Information and Communication Technology and its Applications (ICTA). Federal University of Technology, Minna

[5] Narinder Singh Punn,, Sanjay Kumar Sonbhadra and Sonali Agarwal (2020) ''COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms'' https://doi.org/10.1101/2020.04.08.20057679.

[6] Svetoslav Bliznashki (2020) ''A Bayesian Logistic Growth Model for the Spread of COVID-19 in New York'' https://doi.org/10.1101/2020.04.05.20054577

[7] Qingyuan Zhao, Yang Chen and Dylan S Small (2020)'' Analysis of the epidemic growth of the early 2019-nCoV outbreak using internationally confirmed cases'' https://doi.org/10.1101/2020.02.06.20020941

[8] Zhang. J, Wang y, X. , Jia y, X. , J. Li 1 , Hu.K , Chen.G , Wei.J , Gong.Z , Zhou .C ,Yu.H ,Yu.M , Lei.H Cheng .F ,Zhang .B , Xu. Y ,Wang. G, Dong . W (2020) ''Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China'' Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2020.04.012.