BENIN JOURNAL OF
STATISTICS
**ISSN 2682-5767**
*Vol. 7, pp. 69– 83 (2024)*

# Statistical Analysis of COVID-19 Death Cases in Nigeria Using Machine Learning Approaches and Count Data Regression Models

**F. E. Runyi**[1, *] , **M. T. Nwakuya,** [2] **and M. A. Ijomah** [3]

[1]*Department of Statistics, Federal Polytechnic Ugep, Ugep, Nigeria*
[2,3]*Department of Mathematics and Statistics, University of Port Harcourt, Port Harcourt, Nigeria*

(*Received: 22 November 2023; Accepted: 18 April 2024*)

**Abstract.** In this paper, different methods suitable for analyzing count data based on relevant data characteristics were studied using two of the most popular count regression models (i.e., Poisson regression and negative binomial regression) and some machine learning algorithms, namely decision tree, support vector machine, and neural networks. This study model confirmed cases of COVID-19 death in Nigeria. The data used in the study are the number of cumulative COVID-19 confirmed cases, the number of discharged patients, the number of active cases on admission as predictors, and the number of COVID-19 deaths as the response variable obtained from the Nigeria Centre for Disease Control (NCDC). It is observed that out of the three machine language algorithms considered, the support vector machine learning algorithm provides a better result and outperforms the count regression models in terms of minimum MSE, RMSE and MAE. Then, the support vector machine learning algorithm is recommended for modeling the dynamics of COVID-19 death in Nigeria.

**Keywords:** COVID-19, Count Data, Poisson Model, Negative Binomial Model, Machine Learning Algorithms, Regression

**Published by: Department of Statistics, University of Benin, Nigeria**

## 1. Introduction

Regression modeling is a statistical methodology that shows the relationship between two or more variables. The main goal is to investigate whether data points of a dependent variable can be predicted from another independent variable(s) in a regression model. The linear regression model utilizing ordinary least squares (OLS) estimation is the most widely used traditional regression model in research (Runyi and Maureen, 2022). However, data counted with a

*Corresponding author. Email: runyiemmanuel@gmail.com

positively skewed distribution may not fit well in the OLS linear regression model.

Counts are non-negative integers that represent the number of occurrences of an event within a fixed period. Count data is often modeled using probability distributions suitable for discrete data. One of the most commonly used distributions for count data is the Poisson distribution, which assumes that the events occur independently and at a constant rate. In some cases, count data may exhibit higher variability than what can be explained by the Poisson distribution, giving rise to the problem of overdispersion. Heterogeneity in sample values usually causes overdispersion, outliers, correlated variables, omission of relevant predictors, or zero inflation (Payne *et al.,* 2018). In such cases, alternative models, such as Negative Binomial distribution or Zero-Inflated Models, account for the excess variability. Most of the count data models belong to Generalized Linear Models.

Due to the limitations in some count regression statistical methods and the diversity of data, new techniques for analyzing count regression data using machine-learning algorithms have been developed. In machine learning, data science solves relevant problems and predicts results by considering data quality. Various research fields have combined machine-learning algorithm techniques to consolidate statistical analyses.

For this research, we considered modeling confirmed cases of COVID-19 deaths in Nigeria. The COVID-19 pandemic poses multiple threats to the world's environmental health. It is worth noting that the coronavirus is a disease that affects the old, the young, the rich, and, contrary to popular belief, the poor also, and many lives have been lost due to COVID-19. Developing a predictive model for COVID-19 death will assist in policy formulation. Several research has been carried out to investigate the dynamics of Covid-19 in Nigeria. Roseline *et al.,* (2020) developed a predictive model for confirmed cases in Nigeria using a regression approach. Similarly, Ayinde *et al.,* (2020) conducted a study on modeling Nigerian COVID-19 cases with a comparative analysis of different regression models. Stephen *et al.,* (2021) presented a study on modeling and analyzing the daily incidence of COVID-19 in eighteen countries using count regression models. Oztig and Askin (2020) studied and modeled the relationship between human mobility and COVID-19 using a negative binomial regression. Nwakuya and Nwabueze (2022) applied Poisson regression and negative binomial regression on count data from road fatality during the COVID-19 era in Nigeria. Samuel *et al.,* (2020) modeled COVID-19 cases in Nigeria using selected count data regression models like Poisson regression, Negative Binomial regression, and Generalized Poisson Regression model. Using the negative binomial regression model, Mouhammed *et al.,* (2022) identified explanatory factors for the number of daily COVID-19 death cases in Senegal. Giroh and Nathan (2020) examine COVID-19 cases in the region and its likely effects on food security using the Poisson regression model. Nwakuya and Nkwocha (2023) investigated the robustness of quantile regression of count data over negative binomial regression. Oyindamola and Linda (2015) compared the performance of the Poisson, Negative binomial, and generalized Poisson regression models in predicting Antenatal care visits in Nigeria. In addition, to the best of our knowledge, most of the research carried out so far focused more on the

confirmed cases as not much has been done about COVID-19 death in Nigeria with a focus on the comparative performance of count data regression models and Machine Learning models.

This study presents machine learning (ML) regression algorithm techniques: decision tree, support vector machines (SVM) and neural networks in analyzing COVID-19 count data. Unlike statistical analysis methods, ML algorithms make no distributional assumptions about the response or predictor variables. Some ML techniques accommodate zeros in the response and the predictor variables, making them unique and robust alternatives to statistical methods. We test the estimation techniques on a real-life COVID-19 count data set and compare the performance of the regression count models and machine learning algorithms to assess the fitness and forecast the performance of these different models. Hence, this study aims to compare the performance of Poisson regression, Negative binomial regression, and Machine Learning algorithms in modeling COVID-19 deaths in Nigeria.

## 2. Materials and Methods

### 2.1 *Data Description*

The data for the study were extracted from the official website of the Nigeria Centre for Disease Control (NCDC) on A - State basis. The data extracted were used to determine confirmed cases of COVID-19, number of admissions, number discharged, number of deaths, and active cases of COVID-19. The dependent variable was the number of COVID-19 deaths, while the independent variables were the confirmed cases, number of admissions, number of discharged, and active cases of COVID-19. The data is analyzed using two-count regression models (Poisson and Negative binomial regression) and machine learning algorithms.

Table 1: Confirmed cases by states

| States | Lagos | FCT | Rivers | Kaduna | Plateau | Oyo | Edo | Delta | Ogun |
|---|---|---|---|---|---|---|---|---|---|
| **Number of Confirmed Cases in the Laboratory** | 104,286 | 29,535 | 18,112 | 11,672 | 10,365 | 10,352 | 7,928 | 5,858 | 5,810 |
| **Number of Cases on Admission** | 1,143 | 9 | 3 | 2 | 4 | 0 | 0 | 576 | 11 |
| **Number of Discharged** | 102,372 | 29,277 | 17,960 | 11,581 | 10,286 | 10,150 | 7,606 | 5,170 | 5,717 |
| **Number of Death** | 771 | 249 | 155 | 89 | 75 | 202 | 322 | 112 | 82 |

Table 2: Confirmed cases by states

| States | Kano | Ondo | Akwa Ibom | Kwara | Gombe | Osun | Enugu | Nassarawa | Anambra |
|---|---|---|---|---|---|---|---|---|---|
| **Number of Confirmed Cases in the Laboratory** | 5,429 | 5,173 | 5,010 | 4,691 | 3,313 | 3,311 | 2,952 | 2,846 | 2,825 |
| **Number of Cases on Admission** | 11 | 315 | 6 | 452 | 8 | 29 | 13 | 462 | 46 |
| **Number of Discharged** | 5,291 | 4,749 | 4,960 | 4,175 | 3,239 | 3,190 | 2,910 | 2,345 | 2,760 |
| **Number of Death** | 127 | 109 | 44 | 64 | 66 | 92 | 29 | 39 | 19 |

Table 3: Confirmed cases by states

| States | Bayelsa | Ada-mawa | Niger | Cross River | Sokoto | Jigawa | Yobe | Kebbi | Zam-fara | Kogi |
|---|---|---|---|---|---|---|---|---|---|---|
| **Number of Confirmed Cases in the Laboratory** | 1,373 | 1,312 | 1,183 | 947 | 822 | 669 | 638 | 480 | 375 | 5 |
| **Number of Cases on Admission** | 2 | 134 | 165 | 0 | 0 | 2 | 4 | 10 | 0 | 0 |
| **Number of Discharged** | 1,343 | 1,140 | 998 | 922 | 794 | 649 | 625 | 454 | 366 | 3 |
| **Number of Death** | 28 | 38 | 20 | 25 | 28 | 18 | 9 | 16 | 9 | 2 |

Table 4: Confirmed cases by states

| States | Imo | Ekiti | Kastina | Benue | Abia | Ebonyi | Bauchi | Borno | Taraba |
|---|---|---|---|---|---|---|---|---|---|
| **Number of Confirmed Cases in the Laboratory** | 2,691 | 2,466 | 2,418 | 2,317 | 2,263 | 2,064 | 2,028 | 1,629 | 1,517 |
| **Number of Cases on Admission** | 3 | 0 | 0 | 88 | 0 | 28 | 2 | 5 | 32 |
| **Number of Discharged** | 2,630 | 2,438 | 2,381 | 2,204 | 2,229 | 2,004 | 2,002 | 1,580 | 1,451 |
| **Number of Death** | 58 | 28 | 37 | 25 | 34 | 32 | 24 | 44 | 34 |

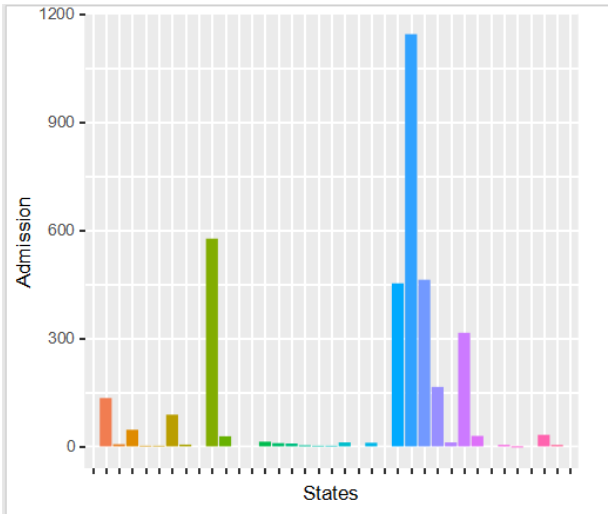


Figure 1: Number of cases on admission

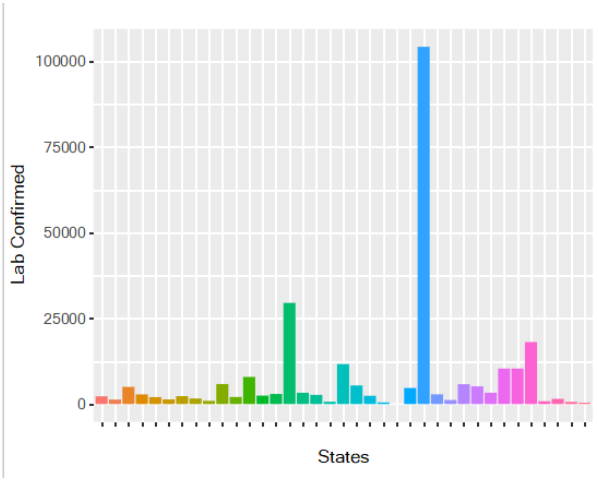


Figure 2: Number of confirmed cases in Laboratory

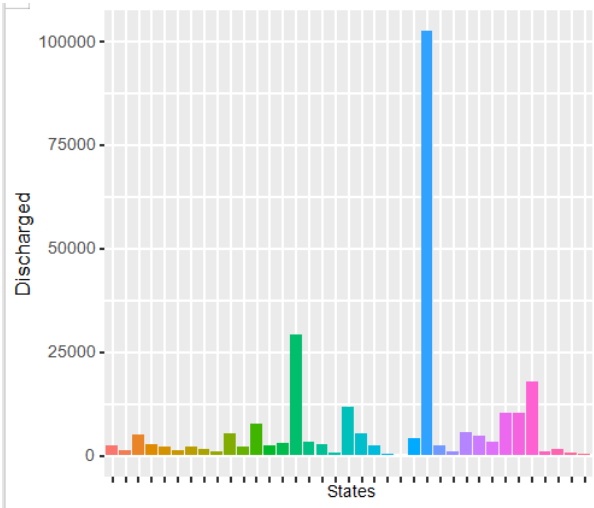


Figure 3: Number of discharged case

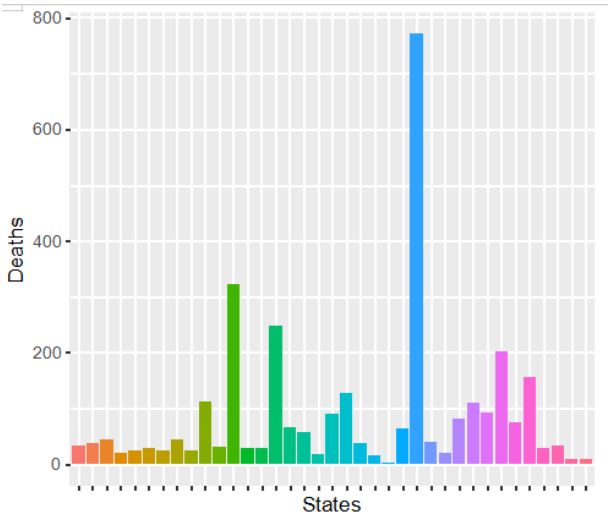


Figure 4: Number of death case


Figures (1-4) above present the bar chart plot of the cumulative COVID-19 confirmed cases recorded in the labs, active cases on admission, number of discharged patients, and number of deaths extracted from 27$^{th}$ February 2020 to 31$^{st}$ March 2023. 266,665 cases were confirmed in the laboratory, 3,559 cases on admission were recorded, 259,951 cases were discharged, and 3,155 COVID-19 deaths were recorded in 36 states and the FCT.

## 2.2 *Poisson Regression*

Poisson regression is a type of nonlinear regression analysis of the Poisson distribution used to predict a response variable that consists of discrete or count data given by one or more independent variables. Poisson regression contains overdispersion if the variance exceeds the mean value (Hardin and Hilbe, 2007; McCurllagh and Nelder, 1983). Overdispersion has the same impact as the assumption that if the offense discrete data occurred over dispersion but still used Poisson regression, the parameter estimates of the regression coefficients remain inconsistent but inefficient (Philip and Sabastian, 2017).
Let Y be a random variable that represents Covid-19 death. The random variable *Y* is said to follow a Poisson distribution with parameter if the probability function is given by;

$$P\left(Y=y\right)=\frac{\lambda^{y}e^{-\lambda}}{y!}, \tag{1}$$

where $y$ represents the number of occurrences of the event $Y$ with mean . One of the valuable properties of the Poisson distribution is that the variance depends on the mean, and the variance is equal to the mean.
Suppose the Covid-19 death rate $\lambda$ is determined by a set of three (3) regressor variables, i.e., $X_1$ = Number of confirmed cases in the lab., $X_2$ = Number of cases on admission, $X_3$ = Number of discharged patients.

The Poisson Regression model will be given as follows:

$$Log\,(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \tag{2}$$

The parameter $\beta$ represents the expected change in the logarithm of the mean per unit change in the predictor $X_i$ which can be estimated using the maximum likelihood estimator (MLE) method.
Taking the exponential of both sides of equation (2), we have,

$$\lambda_i = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \tag{3}$$

Also, the likelihood function is

$$L = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda_i}}{y!} \tag{4}$$

Taking the logarithm of the likelihood function, we obtain the log-likelihood function as:

$$\ell = Log\,(L) = \sum (y_i \beta' x_i - e^{\beta' x_i} - Log(y_i!)). \tag{5}$$

### 2.3 *Negative Binomial Regression*

The negative binomial model is a generalization of the Poisson model by allowing the Poisson parameter to vary randomly following a gamma distribution (Hilbe, 2011). Negative binomial regression is a type of generalized linear model in which the dependent variable Y is a count of the number of times an event occurs. It can be further explained that this occurs when the observed variance is higher than the variance of a theoretical model, and then over-dispersion is said to occur (Samuel *et al.,* 2020). On the other hand, under-dispersion means less variation in the data than predicted (Shaw-Pin, 1993).
Assuming $Y$ is the dependent variable and $X_1$, $X_2$ and $X_3$ are the independent variables and $P(Y = y|x_1,x_2,x_3)$ represents the probability that $Y = y$ when $X_1 = x_1$, $X_2 = x_2$ and $X_3 = x_3$. $X_1$ = Number of confirmed cases in the lab, $X_2$ = Number of cases on admission, $X_3$ = Number of discharged patients.
A convenient parameterization of the negative binomial distribution is given as:

$$P\,(Y) = P\,(Y = y) = \frac{\Gamma\left(y + \frac{1}{\alpha}\right)}{\Gamma\,(y+1)\,\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}, \tag{6}$$

where $\mu > 0$ is the mean of Y and $\alpha > 0$ is the heterogeneity parameter.
The traditional negative binomial regression model is given as follows:

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik.} \tag{7}$$

where the prediction regression coefficient $x_1, x_2, \ldots x_k$ are given and the regression coefficient $\beta_0, \beta_2, \ldots \beta_k$ are to be estimated. Using this notation, the fundamental negative binomial regression model for an observation $i$ is given as:

$$P\left(Y = y_i|\mu_i, \alpha\right) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{8}$$

for $i = 0, 1, 2, \ldots$
The regression coefficient is estimated using the method of maximum likelihood. The likelihood function is:

$$L\left(\alpha, \mu_i\right) = \prod_{i=1}^{n} P(y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}. \tag{9}$$

The log-likelihood function is:

$$\ln L\left(\alpha, \beta\right) = \sum_{i=1}^{n} y_i \ln\alpha + y_i\mu_i - (y_i + \frac{1}{\alpha})\ln\left(1+\mu\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma\left(y_i+1\right) - \ln\Gamma\left(\frac{1}{\alpha}\right). \tag{10}$$

## 2.4 *Machine Learning Regression*

In this study, we considered some machine learning (ML) regression techniques to compare results with count regression models. Machine learning is used to teach machines how to handle data more efficiently. With the abundance of datasets available, the demand for machine learning is rising. Machine learning relies on different algorithms to solve data problems. We present three algorithms used in machine learning for regression analysis. These machine learning algorithms are among the most versatile statistical models, able to automatically classify multidimensional data into two or multiple classes (Pisano *et al.,* 2023)

## 2.5 *Decision Tree*

A decision tree is a graph representing choices and their output results in a tree. The nodes in the graph represent an event or choice, and the graph's edges represent the decision rules or conditions responsible for the output. Each tree consists of nodes and branches. The algorithm selects the best feature to split the data at each node based on a certain criterion (Pisano *et al.,* 2023). Each node represents attributes (A) in a group to be classified, and each branch represents a value (v) the node can take. A decision tree algorithm creates a tree model using values of only one attribute at a time. At first, the algorithm sorts the dataset on the attribute's value. Then, it looks for regions in the dataset containing only one class and marks those regions as leaves. The algorithm chooses another attributes for the remaining regions with more than one class. It continues the branching process with only the number of instances in those regions

until it produces all leaves, or no attribute can produce one or more leaves in the conflicted regions.

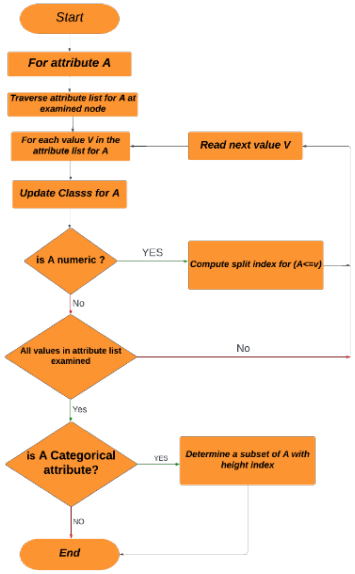


Figure 5: Flowchart of Decision Tree Algorithm

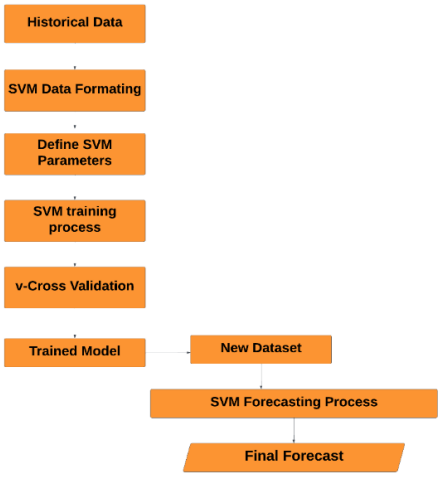


Figure 6: Flowchart of SVM Algorithm

## 2.6 *Support Vector Machine*

A support vector machine (SVM) developed by Boser *et al.,* (1992) is a supervised machine learning technique that analyzes data and isolates patterns that are generally good and applicable to both classification, regression, and forecasting (Claris and Caston, 2023). The classifier helps choose between two or more possible outcomes that depend on continuous or categorical predictor variables. The SVM algorithm assigns the target data into categories based on training and sample classification data. The data is represented as points in space, and categories are mapped in linear and non-linear ways. Data is provided to the algorithm. These inputs are structured correctly to be read. Then, the training process begins. The sample is divided into v parts. One subset is used as a validation part, and the remaining are used to train the model. This process prevents the over-fitting problem and gives the trained model good generalization performance. Once the trained model is created, an unknown data set is provided to the algorithm. SVM produces a forecast output for the unknown sample based on the trained algorithm. Figure 6 describes the flowchart of the algorithm's modeling process.

### 2.7 *Neural Network*

A neural network is an algorithm that endeavors to recognize underlying relationships in a data set through a process miming how the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial. Neural networks can adapt to changing input to generate the best possible result without redesigning the output criteria. A neural network generally consists of three layers, i.e., the input layer, the hidden layer, and the output layer (Azeem *et al.,* 2023). The input layer takes input, The hidden layer processes the input, and the output layer sends the calculated output. The process used for training the network is called a learning algorithm, whose work is to change the function weights of the network to obtain the desired objective.

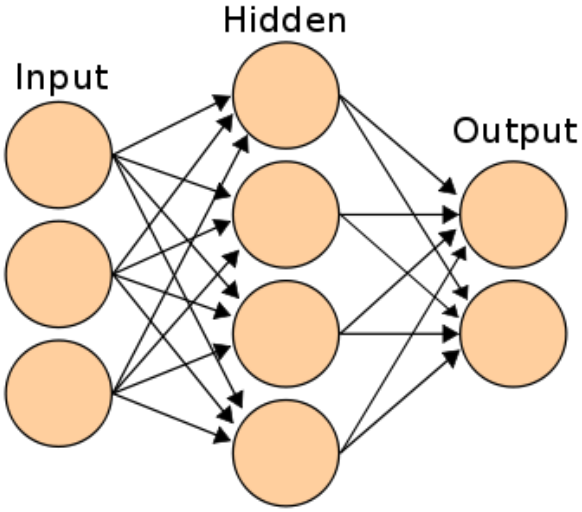


Figure 7: The layers of neural network

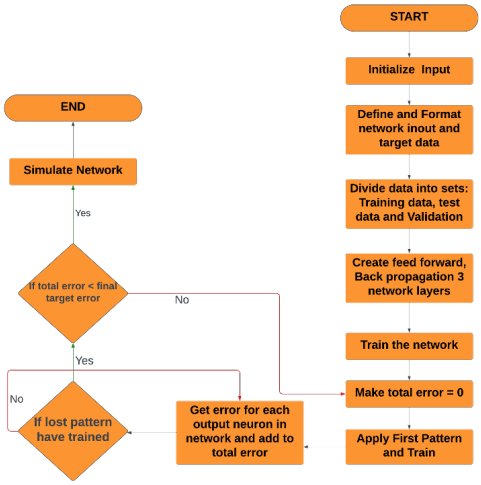


Figure 8: The layers of neural network

## 3. Results and Discussion

Table 5: Descriptive statistics showing the number of confirmed cases in the laboratory, number of patients on admission, number on discharge, and COVID-19 deaths in Nigeria

| Variables | Min | Max. | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| Confirmed cases | 5 | 104286 | 7207.162 | 17349.88 | 5.191 | 29.007 |
| Admission | 0 | 1143 | 96.35135 | 227.924 | 3.324 | 12.425 |
| Discharge | 3 | 102372 | 7025.703 | 17058.84 | 5.177 | 28.872 |
| COVID-19 death | 2 | 771 | 85.27027 | 134.89 | 4.033 | 19.083 |

The result of the descriptive statistics presented in Table 5 reveals that the average COVID-19 death was 85.27, with a maximum deaths of 771 persons. The average confirmed cases, number of admissions, number on discharge, and active cases were obtained to be 7207.16, 96.35, and 7025.70, respectively. The skewness obtained for all variables was above 0 (zero), meaning that these variables are positively skewed, with confirmed cases showing more excess kurtosis

(29.007) than other variables. Two different count data regression models were fitted to the data, and the summary results are presented in Tables 6, 7 and 8.

Table 6: Summary result of the Poisson regression for modelling confirmed cases of Covid-19 death in Nigeria

| **Variables** | **Coefficient** | **Std. Error** | **Z-value** | **P-value** | **95% confidence interval** | |
|---|---|---|---|---|---|---|
| Confirmed cases | 0.00774 | 0.000232 | 33.25 | 0.000∗ ∗ | 0.0072867 | 0.0081996 |
| Admission | -.008464 | 0.000261 | -32.46 | 0.000∗ ∗ | -0.0089747 | -0.007953 |
| Discharge | -.007764 | 0.000235 | -33.09 | 0.000∗ ∗ | -0.0082233 | -0.00730 |
| Constant | 3.66012 | 0.034060 | 107.46 | 0.000∗ ∗ | 3.59336 | 3.726871 |
| Summary Statistic | | | | | | |
| Pseudo $R^2$ | 0.8295 | | | | | |
| Log-likelihood | -373.5584 | | | | | |
| Prob > chi2 | 0.0000∗ ∗ | | | | | |

∗ ∗**significant at 1% (p<.01),** ∗ **significant at 5% (p<.05)**

Table 6 shows the summary results of the Poisson regression with Pseudo $R^2$ of 0.8295, which implies that 82.95% of the variation in COVID-19 death was accounted for by the independent variables (number of confirmed cases in the laboratory, number of cases on admission, and number discharged cases). A p-value of 0.000 (p<.01) was obtained, which shows that the predictor variables accounted for significant variation in the dependent variable (CoVID-19 death). Based on the results of Poisson regression, all three independent variables were found to impact COVID-19 death significantly. For the confirmed cases, the contribution was positive and significant ($\beta$ = .0077431, SE= 0.0002329, Z-cal. = 33.25, p = 0.000, p<.01). At the same time, number on admission showed significant negative contribution to Covid-19 death ($\beta$ = -.0084637, SE= 0.0002607, $Z$-cal. = -32.46, $p$ = 0.000, p<.01). The impact of number of cases on admission and number on discharge cases were significantly negative (p<.01). But despite this relatively high Pseudo $R^2$ of the Poisson regression, one of its significant limitations is that it cannot handle the issue of overdispersion because Poisson regression assumes that the mean and the variance are equal (equi-dispersed).

Table 7: Testing for overdispersion in the Poisson regression model

| **Test statistics** | **Statistic** | **P-value** | **Remark** |
|---|---|---|---|
| Deviance goodness-of-fit | 544.4974 | 0.0000∗ ∗ | Significant |
| Pearson goodness-of-fit | 527.0675 | 0.0000∗ ∗ | Significant |

∗ ∗**significant meaning there is evidence of overdispersion**

Overdispersion test was carried out using Deviance goodness-of-fit (test statistic= 544.4974, $p$ = 0.000, p<.01) and Pearson goodness-of-fit (test statistic = 527.0675, p = 0.000, p<0.01). The p-values are less than 0.05 (p<.05), which implies that there is evidence of overdispersion. This means the result of the Poisson regression is spurious; hence, negative binomial regression was fitted to the data to correct this, and the results are presented in Table 8.

Table 8: Summary result of negative binomial regression for modelling confirmed cases of Covid-19 death in Nigeria

| **Variables** | **Coefficient** | **Std. Error** | **Z-value** | **P-value** | **95% confidence interval** | |
|---|---|---|---|---|---|---|
| Confirmed cases | 0.0119369 | 0.0023482 | 5.08 | 0.000∗ ∗ | 0.0073346 | 0.0165393 |
| Admission | -0.0122407 | 0.0024993 | -4.9 | 0.000∗ ∗ | -0.0171391 | -0.0073422 |
| Discharge | -0.0119822 | 0.0023644 | -5.07 | 0.000∗ ∗ | -0.0166163 | -0.007348 |
| **Constant** | 3.311564 | 0.1525812 | 21.7 | 0.000∗ ∗ | 3.01251 | 3.610617 |
| Summary statistic | | | | | | |
| Pseudo R2 | 0.1458 | | | | | |
| Log-likelihood | -172.27097 | | | | | |
| Prob > chi2 | 0.000∗ ∗ | | | | | |

∗ ∗**significant at 1% (p<.01),** ∗**significant at 5% (p<.05)**

Results in negative binomial are also significant, and the joint contributions of the independent variable were significant in the models (p<.01). The Pseudo $R^2$ of 0.1458 was obtained for the negative binomial regression, indicating that the independent variables accounted for 14.58% of the variation in the confirmed cases of Covid-19 death in Nigeria.

Table 9: Result of the ML algorithms for Covid-19 death

| **S/N** | **ML models** | **$R^2$** | **RMSE** | **MSE** | **MAE** |
|---|---|---|---|---|---|
| 1 | Decision Fine Tree | 0.350 | 218.16 | 47596 | 115.18 |
| 2 | **SVM** | **0.9600** | **53.099** | **2819.5** | **32.717** |
| 3 | Neural Network | 0.830 | 112.61 | 12681 | 56.309 |

***$R^2$- coefficient of determination, MSE- Mean Square Error, RMSE- Root Mean Square Error.***

Three different Machine Learning algorithms were considered namely decision tree, SVM, and neural network. The result reveals that the SVM has the highest $R^2$ (0.9600), least MSE (2819.50), RMSE (53.099), and MAE (32.717). This indicates that the SVM is the best among the competing Machine learning algorithms.

Table 10: Comparison of the best count regression model with the best machine learning algorithm

| **Models** | **$R^2$** | **MSE** | **RMSE** | **MAE** |
|---|---|---|---|---|
| Negative binomial regression | 0.1443 | 44580.0362 | 211.1398 | 64.36507 |
| **SVM** | **0.9600** | **2819.5000** | **53.0909** | **32.7170** |

Furthermore, Table 10 shows the results of the best count regression models, in this case, the negative binomial regression with that of the SVM. The result, as presented in Table 10, shows that SVM gave the highest $R^2$ (**0.960** vs. 0.1493), least MSE (**2819.5** vs. 44580.03), least RMSE (**53.099** vs. 211.14), and least MAE (**32.717** vs. 64.37). This indicates that the SVM machine-learning algorithm is superior to the negative binomial regression in estimating COVID-19 death based on the predictor variables.

## 4. Conclusion

Based on the findings presented in this study, it is evident that the SVM algorithm outperforms both count regression models (Poisson regression and negative binomial regression) in modeling cases of COVID-19 death in Nigeria. With the global threat posed by the COVID-19 pandemic, accurate modeling and prediction of COVID-19 mortality rates is crucial for effective public health planning and intervention strategies. The comparison between count regression models and machine learning algorithms highlights the superiority of SVM in terms of least MSE, RMSE, and MAE. Therefore, SVM is highly recommended for modeling the dynamics of COVID-19 death in Nigeria. This recommendation underscores the importance of leveraging advanced machine learning techniques to enhance our understanding and response to public health crises such as the COVID-19 pandemic.

## References

Ayinde, K., Lukman, A. F., Rauf, I. R., Alabi, O. O., Okon, C. E., and Ayinde, O. E. (2020). Modelling Nigerian COVID-19 cases: A comparative analysis of models and estimators. Chaos, Solicitors & Fractals, 138, 109911. https://doi:10.1016/j.chaos.2020.109911.

Azeem M, Javaid S, Khalil R. A, Fahim H, Althobaiti T, Alsharif N, and Saeed N.(2023). Neural Networks for the Detection of COVID-19 and Other Diseases: Prospects and Challenges. Bioengineering (Basel). 10(7), 850. doi: 10.3390/bioengineering10070850.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92), Pittsburgh, 144-152.

Claris, S. and Caston, S. (2023). Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data. American Journal of Infection Control, 51(10), 1095-1107. https://doi.org/10.1016/j.ajic.2023.03.010.

Francis, R. and Nwakuya, M. (2022). Logistic Estimation in the presence of Collinearity and its application. International Journal of Data Science and Analysis, 6(8), 187-193.

Giroh, Y. D. and Nathan, N. (2020). A Poisson Regression Analysis of COVID-19 Pandemic: Implication on Food Security in North Eastern Nigeria. SSRN Electronic Journal.

Hilbe, J. M. and Hardin, J. W. (2007). Generalized linear models. Chapman and Hall, London.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.

Mouhamed, N., Aba, D., Cheikh, T., Ibrahima, F., Abdourahmane, N., and Idrissa, S. (2022). A negative binomial regression analysis for COVID-19 death cases in Senegal. International Journal of Statistics and Applied Mathematics, 7(6), 06-12. doi:10.22271/maths.2022.v7.i3a.817.

Nwakuya, M. T. and Nwabueze, J. C. (2022). A Negative Binomial Regression on Road Accident Fatalities During COVID-19 Hit Era in Nigeria, International Journal of Statistical Distributions and Applications, 8(3), 40-46. doi: 10.11648/j.ijsd.20220803.11

Nwakuya, M. and Nkwocha, C. (2023). Quantile Regression for Count Data as a Robust Alternative to Negative Binomial Regression. African Journal of Mathematics and Statistics Studies, 6(9), 1-11.

Oyindamola, B. Y. and Linda, O. U. (2015). On the Performance of the Poisson, Negative Binomial and Generalized Poisson Regression Models in the Prediction of Antenatal Care Visits in Nigeria. American Journal of Mathematics and Statistics, 5(3), 128-

136. doi: 10.5923/j.ajms.20150503.04

Oztig, L. I. and Askin, O. E. (2020). Human Mobility and Coronavirus disease 2019 (COVID-19): a negative binomial regression analysis. The Royal Society of Public Health 185, 304-367.

Payne, E. H., Gebregziabher, M., Hardin, J. W., Ramakrishnan, V., and Egede, L. E. (2018). An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. Communications in Statistics-Simulation and Computation, 47, 1722-1738.

Philip, N. and Sabastian, N. (2017). Application of Poisson regression on traffic safety. Degree project, in second level mathematical statistics Stockholm. htttps://www.divaportal. org/smash/get/diva2:816402/FULLTEXT01.pdf.

Pisano, F., Cannas, B., Fanni, A., Pasella, M., Canetto, B., Giglio, S. R., Mocci, S., Chessa, L., Perra, A., and Littera, R. (2023). Decision trees for early prediction of inadequate immune response to Coronavirus infections: a pilot study on COVID-19. Front. Med. 10:1230733. doi:.3389/fmed.2023.1230733

Roseline, O. O., Adewale, F. L., Golam, B. M., Joseph, B. A., and Benedita, B. A. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. Infectious Disease Modelling, 5, 543-548. https://doi.org/10.1016/j.idm.2020.08.003.

Samuel, O. A., Muhammad, A. B., Samuel, O. O., Haruna, U. Y., et al., (2020). Modeling COVID-19 Cases in Nigeria Using Some Selected Count Data Regression Models. International Journal of Healthcare and Medical Sciences, Academic Research Publishing Group, 6(4), 64-73.

Stephen, C., Jeffrey, C., Yuanyuan, Z., and Saralees, N. (2021). Count regression models for COVID-19, Physica A: Statistical Mechanics and its Applications, 5639(23), 120-135. https://doi.org/10.1016/j.physa.2020.125460.

Shaw-Pin, M. (1993). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regression. Center for Transportation Analysis, Energy Division Oak Ridge national Laboratory. 14

**Appendix**

```
Support Vector Machine - Regression
set.seed(2221022)
library(caret)
COVID19D < − read.csv("C:/Users/Runyi E. Francis/Desktop)covid19data.csv",header=T)
COVID19D2- COVID19D [,-c(1)]
intrain< −createDataPartition(y= COVID19D2 death,p=0.8,list=FALSE)
training< − COVID19D2[intrain,]
testing< −COVID19D2 [-intrain,]
anyNA(COVID19D2)
trctrl < − trainControl(method="repeatedcv",number=10,repeats=3)
svmLinearReg < − train(Price.,data=training,method="svmLinear",trControl=trctrl,preProcess=c("c
gth=10)
svmLinearReg
vimp<varImp(svmLinearReg)
vimp test pred< −predict(svm Linear Reg,newdata=testing)
plot(testpred)
plot the predicted values (as line) and original value together x=1:length(testing$death)
plot(x,testing$death,pch=18,col="red")
lines(x, testpred, lwd = "1", col = "blue") plot the predicted values (as points) and original
value together x=1:length(testing$death)
plot(x,testing$death,pch=18,col="red")
points(x, testpred, pch = 18, col = "blue")
Neural Network - Regression
library(plyr)
library(readr)
library(dplyr)
library(caret)
library(neuralnet)
library(NeuralNetTools)
library(nnet)
set.seed(1000)
COVI19D< −read.csv("C:/Users/Runyi E. Francis/Desktop/covid19data.csv",header=T)
COVID19D2< − COVID19D[,-c(1)]
COVID19D2< − COVID19D$death- COVID19D$discharge head(COVID19D2)
intrain< −createDataPartition(y= COVID19D2$death,p=0.8,list=FALSE)
training< −COVID19D2[intrain,]
testing< − COVID19D2[-intrain,]
anyNA(COVID19D2) trctrl¡ trainControl(method="repeatedcv",number=10,repeats=5)
nnetmodel train(death/2065.6, training, method = "nnet", trControl = trctrl, preProcess =
c("scale", "center"))nnetmodel
plot(nnetmodel)
nnetpredictionstest< −predict(nnetmodel, testing)
nnet predictions test*2065.6
plot(nnet predictions test*2065.6)
results < −data.frame(testing$death,prediction=nnet predictions test*2065.6)
head(results)
tail(results)
plot the predicted values (as line) and original value together x=1:length(testing$death)
plot(x,testing$death, pch = 18, col = "red")
lines(x, nnetpredictionstest 2065.6, lwd = "1", col = "blue")
plot the predicted values (as points) and original value together x=1:length(testing$death)
plot(x,testing$death,pch=18,col="red")
points(x,nnet predictionstest 2065.6, pch = 18, col = "blue")
test.mse< −(sum(testing$death-nnet predictions test*2065.6)2)/(nrow(testing))
rmse.nnet=sqrt(test.mse)
```

```
rmse.nnet
Decision Tree
plot
COVID19D < – read.csv("C:/Users/Runyi E. Francis/Desktop/covid19data.csv",header=T)
COVID19D Class < – as.factor(ifelse(COVID19DClass == 0, "nodeath", "death"))
ggplot(COVID19D,aes(x=confimedcase,discharge,onadmission,y=death,color=Class))+geompoint()
geomsmooth()
set.seed(0220)
library(caret)
COVID19D< –read.csv("C:/Users/Runyi E. Francis/Desktop/
covid19data.csv",header=T)
COVID19D2< –COVID19D[,-c(1)]
intrain< –createDataPartition(y=data2$Class,p=0.8,list=FALSE)
training< – COVID19D2[intrain,]
testing< – COVID19D2[-intrain,]
anyNA(COVID19D2) convert to factor training[["Class"]]=factor(training[["Class"]])
trctrl< –trainControl(method="repeatedcv",number=10,repeats=3)
cardtree< –train(Class.,data=training,method="rpart",trControl=trctrl) cardtree
library(rpart.plot)
rpart.plot(cardtree< finalModel)
test pred-predict(cardtree,newdata=testing)
confusionMatrix(table(test pred,testing$Class))
set.seed(0221)
trctrl-trainControl(method="repeatedcv",number=10,repeats=3)
grid< –expand.grid(cp=c(0,0.0001,0.001,0.06,0.07,0.35,0.46,0.65))
cardtree
grid< – train(Class.,data=training,method="rpart",trControl=trctrl,tuneGrid=grid) test
pred
grid< –predict(cardtree grid,newdata=testing) test pred gridconfusionMatrix(table(test
pred grid,testingClass)) rpart.plot(cardtree grid$finalModel)
varImp(cardtree grid)
```