Original Article

Analyzing the Death Ratio of Covid Patients using Multiple Logistic Regression in Comparison with Linear Regression for Improving Accuracy

B. Bharath Kumar Raju[1], N. Deepa [2*]

[1]Research Scholar,Department of Computer Science and Engineering,Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences,Saveetha University, Chennai, Tamilnadu, India. Pincode:602105.
[2]Project Guide, Department of Computer Science and Engineering,Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences,Saveetha University, Chennai, Tamilnadu, India. Pincode:602105.

Abstract

Aim: The aim of the study is to analyze the death ratio of covid patients using Novel Multiple Logistic Regression and linear regression which comes under supervised learning.
Materials and Method: Accuracy is analyzed for a covid dataset of size 239 places. Analyzingthe death ratio of covid patients is performed by a Novel Multiple Logistic Regression of sample size (N=35) and Linear Regression of sample size (N=35), obtained using the G-power value of 80%. These are supervised learning algorithms.
Result: Novel Multiple Logistic Regression accuracy is 96% which is comparatively higher than LR with an accuracy of 86%. The significance value is determined as p=0.030 (p<0.05) for accuracy.
Conclusion: Novel Multiple Logistic Regression performs better in finding accuracy when compared to Linear Regression.
Keywords: Big Data Analytics, Supervised learning, Death ratio, Linear Regression, Novel Multiple Logistic Regression, Machine Learning.

DOI: 10.47750/pnr.2022.13.S04.032

INTRODUCTION

The pandemic of COVID-19 has created havoc on human civilization. Since its appearance in the city of Wuhan (Hebei district) in China, it has been a relentless march of new cases and deaths. Analyzing the ratio between deaths and individuals of covid patients (Kulkarni and Lorenz 2020). In the statistics field, logistic regression is developed and this model is used for studying input and output relations of the numeric variables, and later it is adopted by Ml. It comes under both statistical and machine learning algorithms. It is used to know the number of deaths through covid disease (Wujtewicz et al. 2020). It is also used in Providing information about how covid is spreading in the region (Lombardo et al. 2021). In big data it is used in checking and tracing patients suffering from covid disease and in Machine learning it is used to predict patients diagnosed with COVID-19 disease (Lichtner et al. 2021) and in IOT it is used to monitor affected persons with that particular disease (Shan'an and Qin 2021). The applications of Multiple Logistic Regression are they are used in medical fields and social sciences to predict mortality. The applications of Linear Regression are they are used in medical research to understand the drug dosage and blood pressure in patients.

In the area of Big Data Analytics, several research papers are available in IEEE and Science Direct. From IEEE Xplore digital library 41 journals are identified 23,863 articles from ScienceDirect,18,700 articles from GoogleScholar,14,603 articles from Springer.The most cited article is as follows (Kalish 2015) has a citation of 2242 times, his existing work is mainly based on Analyzing pedagogical components accurately. The major death analysis is done based on C-reactive protein (CRP), CRP concentration, etc. (Cohen et al. 2017) have more citations 2856 times, the research work about analyzing death ratio using Multiple logistic regression. These are the machine learning techniques. The mortality risk of hospitalized patients is analyzed (Kwekha-Rashid, Abduljabbar, and Al Hayani 2021) and has a citation 60 times. The problem and its simple solution are analyzed (Gray et al. 2019) and have a citation 80 times. The mortality assessment and cure are analyzed (Madea 2015) and have a citation of 140.

Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021). The existing system has a drawback in that it only took a few states to find the death ratio but there are more states. The accuracy is less in the existing system. In the proposed system accuracy is more than the existing system and has taken data from more than 200 places of people suffering from the covid disease. Hence the proposed method aims at comparing algorithms to know which algorithm was giving more accuracy than the Linear Regression. The aim of the proposed work is to provide an accurate death ratio due to covid disease in different states and to improve accuracy.

MATERIALS AND METHODS

The research work is carried out in the Image processing laboratory in Saveetha School of Engineering, Saveetha Institution of Medical and Technical Sciences where the laboratory consists of high configuration equipment to gain good results. The number of groups identified for the study was two with a sample size of 35 per group (Smithson and Shou 2019). G-power is used as computation with 80% and has an alpha value of 0.05 and a beta value of 0.95 with a 95% confidence interval.

In sample preparation group 1, Novel Multiple Logistic Regression is used to train the Multiple Logistic Regression. Used statistical analysis at the back end and it falls under supervised learning algorithms. It provides better accuracy for small and simple datasets. It has a resistance to overfitting. This model will help us to find accuracy. This model will help us by providing a death ratio due to covid in multiple states. Multiple Logistic Regression is extended from logistic regression and supports multi-class classification problems. Novel Multiple Logistic Regression is used to predict multinomial probability. Multiple Logistic Regression is used to measure independent variables and to estimate the probability of particular dependent variables. Table 1 represents Multiple Logistic Regression and it is a machine learning technique.

In sample preparation group 2, Linear Regression is used to predict the value of one number based on the value of another number and it is a supervised learning algorithm. It is also used to find accuracy. In this project linear regression is used as a comparison algorithm for my proposed system. Linear regression is a linear model. Linear Regression belongs to both statistical learning and machine learning. Linear Regression is developed in the field of statistics and was lent by ML. It is used for understanding input and output numerical variables relations. In linear regression, observations are independent of each other. In linear regression, the model is estimated. In linear regression, there is a directionality of data. Table 2 represents Linear Regression and it is a machine learning technique.

Jupyter notebook is used for the implementation of this framework. All codes are done in the same notebook itself. Hardware configurations of the system which I worked on consist of 8GB RAM and ROM of 1TB HDD+256 SSD with a processor of 11th gen intel(R) Core i5-1135G7 @2. 40 GHZ. There are two groups: group 1 consists of Multiple Logistic Regression and group 2 consists of Linear Regression.

The dataset named covid19 is downloaded from the Kaggle website. The covid dataset is collected as a record from different locations including several places in India. In these records affected and suffering patients' data is included. Also recovered patients' details are evolving. Added to it recently died in those pandemic situations is also added as one of the columns in the dataset.

Statistical Analysis

Statistical software used is IBM SPSS with version 26.0 to find the standard deviation, mean, standard error mean, mean difference, sig, and F value. Independent T-Test analysis is carried out in this research. Independent variables are Unnamed and State/UTs in the dataset and dependent variables are Active Ratio, Death Ratio, and Discharge Ratio in the dataset.It is used to predict future death rates (Syed et al. 2021) from the inputs which we have taken from the analysis.

RESULTS

Table 1 represents the pseudocode of Novel Multiple Logistic Regression and it is a machine learning technique. At first, the libraries are initialized and the dataset is read into the model. The model splits the dataset into training and testing sets and these sets are assigned to a value and use some functions and calculate the required accuracy.

Table 2 represents the pseudocode of Linear Regression. First, the libraries are initialized and the dataset is read into the model. The model splits the dataset into training and testing sets and uses some functions and calculates the required accuracy. And end the program after getting the accuracy.

Table 3 represents values of mean, standard deviation, and standard error mean are classified based on Multiple Logistic Regression for training accuracy. The accuracy of Multiple Logistic Regression is better when compared to Linear Regression.

Table 4 represents the raw data for accuracy of both Multiple Logistic Regression and Linear Regression.

Table 5 represents the standard error difference and significance of the data where the significance value of accuracy for both Multiple Logistic Regression and Linear Regression is 0.030.

Figure 1 represents the accuracy comparison with 96% significance and 86% accuracy for both Multiple Logistic Regression and Linear regression. Mean accuracy when compared for both Multiple Logistic Regression and Logistic Regression the accuracy for Multiple Logistic regression is higher.

DISCUSSION

The statistical mean accuracy obtained by the proposed system is 96.10 and for the existing system mean accuracy is 86.24 and the significant value for both the proposed and existing algorithms is less than 0.05 where $p = 0.030$ for the selected dataset. Multiple Logistic Regression is with better accuracy when compared to Linear Regression (Table 4).

The research work in this paper has discussed how the covid19 virus has spread all over different countries (Chu 2021). Research work predicted the mortality rate in India using statistical neural network models (Dhamodharavadhani, Rathipriya, and Chatterjee 2020). The research work in this paper has calculated the mortality risks of patients who joined hospitals suffering from covid19 (Sandhu et al. 2021). The research work (Guzmán-Torres et al. 2021) talked about the internal effects caused due to covid in different individuals. The research work in this paper (Devkota 2021) said about direct and indirect ways of how the disease has spread in different developing countries. The research work in this paper (Devkota 2021; Salazar et al. 2020)has discussed how various plasmas are used to decrease covid19 disease. When it comes to analyzing death ratio, the accuracy of Multiple Logistic Regression was superior to that of other Supervised algorithms. The accuracy of the Multiple Logistic Regression classification algorithms depends on the size of the training and testing data set. In our study, accuracy appears to be better than Linear Regression. However, the average error appears to be higher in our proposed work which should be minimized.

The factors that affect the death ratio are lack of medical facilities and lack of cleanliness and hygiene. These are some factors that will affect the death ratio. The limitations of finding the death ratio are that the data is based on large numbers which may be unstable. So that leads to a very drastic change in the death ratio or for any medical purposes including the fetching proper medical records is difficult. In the future, the death ratio can be used to verify how drastically the population changed in that particular year and how many are suffering and cured with that particular disease. In this way, the death ratio can be used.

CONCLUSION

Novel Multiple Logistic Regression and Linear Regression are both machine learning techniques that use averaging to improve accuracy. The work shows the death ratio accuracy of people suffering in the pandemic through covid

disease using Multiple Logistic Regression, Linear Regression, Lasso Regression, Logistic Regression, and Bayesian Linear Regression. It is found that Multiple Logistic Regression gained more accurate results than Linear Regression, Logistic Regression, Lasso Regression, and Bayesian Linear Regression. Hence, it is concluded that Multiple Logistic Regression provides more accuracy when compared with other algorithms.

DECLARATIONS

Conflict of interest

No conflict of interest in this manuscript.

Author's Contributions

Author BKN was involved in the methodology, text analysis, and writing the manuscript. Author ND was involved in review and editing, supervision, and validation.

Acknowledgments

The authors would like to express their gratitude to Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Chu, Jeffrey. 2021. "A Statistical Analysis of the Novel Coronavirus (COVID-19) in Italy and Spain." *PloS One* 16 (3): e0249037.
2. Cohen, Aaron J., Michael Brauer, Richard Burnett, H. Ross Anderson, Joseph Frostad, Kara Estep, KalpanaBalakrishnan, et al. 2017. "Estimates and 25-Year Trends of the Global Burden of Disease Attributable to Ambient Air Pollution: An Analysis of Data from the Global Burden of Diseases Study 2015." *The Lancet* 389 (10082): 1907–18.
3. Cuzick, Jack, Gregory P. Swanson, Gabrielle Fisher, Arthur R. Brothman, Daniel M. Berney, Julia E. Reid, David Mesher, et al. 2011. "Prognostic Value of an RNA Expression Signature Derived from Cell Cycle Proliferation Genes in Patients with Prostate Cancer: A Retrospective Study." *The Lancet Oncology* 12 (3): 245–55.
4. Devkota, Jyoti U. 2021. "Multivariate Analysis of COVID-19 for Countries with Limited and Scarce Data: Examples from Nepal." *Journal of Environmental and Public Health* 2021 (January): 8813505.
5. Dhamodharavadhani, S., R. Rathipriya, and Jyotir Moy Chatterjee. 2020. "COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models." *Frontiers in Public Health* 8 (August): 441.
6. Guzmán-Torres, José A., Elia M. Alonso-Guzmán, Francisco J. Domínguez-Mota, and Gerardo Tinoco-Guerrero. 2021. "Estimation of the Main Conditions in (SARS-CoV-2) Covid-19 Patients That Increase the Risk of Death Using Machine Learning, the Case of Mexico." *Results in Physics* 27 (August): 104483.
7. Kalish, Mia. 2015. "Analyzing Pedagogical Components Using Multiple Regression." *On the Horizon*. https://doi.org/10.1108/oth-03-2014-0008.
8. Kulkarni, Saagar S., and Kathryn E. Lorenz. 2020. "Social Implications of COVID-19 Deaths: Analyzing Race, Ethnicity, Socio-Economic Conditions, Gender, and Age for the US." *Advanced Journal of Social Science*. https://doi.org/10.21467/ajss.7.1.163-180.
9. Lichtner, Gregor, Felix Balzer, Stefan Haufe, NiklasGiesa, FridtjofSchiefenhövel, MalteSchmieding, Carlo Jurth, et al. 2021. "Predicting Lethal Courses in Critically Ill COVID-19 Patients Using a Machine Learning Model Trained on Patients with Non-COVID-19 Viral Pneumonia." *Scientific Reports* 11 (1): 13205.
10. Lombardo, Flavia L., IlariaBacigalupo, Emanuela Salvi, EleonoraLacorte, Paola Piscopo, Flavia Mayer, Antonio Ancidoni, et al. 2021. "The Italian National Survey on Coronavirus Disease 2019 Epidemic Spread in Nursing Homes." *International Journal of Geriatric Psychiatry* 36 (6): 873–82.
11. Salazar, Eric, Paul A. Christensen, Edward A. Graviss, Duc T. Nguyen, Brian Castillo, Jian Chen, Bevin V. Lopez, et al. 2020. "Treatment of Coronavirus Disease 2019 Patients with Convalescent Plasma Reveals a Signal of Significantly Decreased Mortality." *The American Journal of Pathology* 190 (11): 2290–2303.
12. Sandhu, Avnish, Steven J. Korzeniewski, Jordan Polistico, Harshita Pinnamaneni, Sushmitha Nanja Reddy, Ahmed Oudeif, Jessica Meyers,

et al. 2021. "Elevated COVID19 Mortality Risk in Detroit Area Hospitals among Patients from Census Tracts with Extreme Socioeconomic Vulnerability." *EClinicalMedicine*. https://doi.org/10.1016/j.eclinm.2021.100814.
13. Shan'an, Yu, and Yunfei Qin. 2021. "Energy-Efficient IoT Based Improved Health Monitoring System for Sports Persons." *Journal of Intelligent & Fuzzy Systems*. https://doi.org/10.3233/jifs-219015.
14. Smithson, Michael, and YiyunShou. 2019. *Generalized Linear Models for Bounded and Limited Quantitative Variables*. SAGE Publications.
15. Sunyer, J., J. Schwartz, A. Tobías, D. Macfarlane, J. Garcia, and J. M. Antó. 2000. "Patients with Chronic Obstructive Pulmonary Disease Are at Increased Risk of Death Associated with Urban Particle Air Pollution: A Case-Crossover Analysis." *American Journal of Epidemiology* 151 (1): 50–56.
16. Wujtewicz, Magdalena, Anna Dylczyk-Sommer, AleksanderAszkiełowicz, SzymonZdanowski, Sebastian Piwowarczyk, and RadoslawOwczuk. 2020. "COVID-19 - What Should Anesthesiologists and Intensivists Know about It?" *Anaesthesiology Intensive Therapy* 52 (1): 34–41.
17. Syed, Mahanazuddin, Shorabuddin Syed, Kevin Sexton, Melody L. Greer, Meredith Zozus, Sudeepa Bhattacharyya, Farhanuddin Syed, and Fred Prior. 2021. "Deep Learning Methods to Predict Mortality in COVID-19 Patients: A Rapid Scoping Review." *Studies in Health Technology and Informatics* 281 (May): 799–803.
18. Madea, Burkhard. 2015. *Estimation of the Time Since Death*. CRC Press.
19. Gray, William A., Michael R. Jaff, Sahil A. Parikh, Gary M. Ansel, Marianne Brodmann, Prakash Krishnan, Mahmood K. Razavi, et al. 2019. *Mortality Assessment of Paclitaxel-Coated Balloons: Patient-Level Meta-Analysis of the ILLUMENATE Clinical Program at 3 Years*.
20. Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. 2021. "Coronavirus Disease (COVID-19) Cases Analysis Using Machine-Learning Applications. " *Applied Nanoscience*, May, 1–13.

TABLES AND FIGURES

Table 1. Pseudocode for Novel Multiple Logistic Regression algorithm. At first the libraries are intialized and the dataset is readed into the model. The model splits the dataset into training and testing sets and these sets are assigned to a value and use some functions and calculate the required accuracy.

Input: Covid Dataset
Output: Accuracy
Step 1: Import and read the dataset.
Step 2: Select some features from the dataset.
Step 3: Generate the parameter.
Step 4: Analyze the dataset by changing the dependent and independent variables.
Step 5: Predict the output in numerical variables.
Step 6: Predict the output by using functions.

Table 2. Pseudocode of Linear Regression algorithm at first the libraries are intialized and the dataset is readed into the model. The model splits the dataset into training and testing sets and uses some functions and calculates the required accuracy. And end the program after getting the accuracy.

Input: CovidDataset
Output: Accuracy
1)Start
2)Read Number of Data (n)
3)For i=1 to n: Read X_i and Y_i Next i
4) Initialize: sumX = 0 sumX2=0 sumY=0 sumXY=0
5) Calculate Required sum For i=1 to n: sumX = sumX + X_i sumX2 = sumX2 + X_i*X_i sumY=sumY +Yi sumXY=sumXY+ X_i*X_i Next i
6) Calculate Required Constant a and b of y=a+bx: b = (n*sumXY-sumX+sumY)/(n*sumX2-sumX*sumX) a = (sumY-b*sumX)/n
7) Display value of a and b
8) Stop

Table 3. Statistical calculation for independent samples tested between Multiple Logistic Regression and Linear Regression. The mean accuracy of Novel Multiple Logistic Regression is 96.10 and Linear Regression is 86.24. Standard Deviation for Multiple Logistic Regression is 1.834 and Linear Regression is 2.743.

	Algorithm	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	MLR	35	96.10	1.834	.310
	LR	35	86.24	2.743	.464

Table 4. Rawdata table accuracy for both Multiple Logistic Regression and Linear Regression using SPSS statistics.

Group_Id	MLR	LR
1	92	85
2	98	87
3	94	86
4	95	88
5	93	89
6	90	80
7	98	82
8	94	81
9	97	83
10	96	84
11	97	85
12	95	85
13	98	88
14	96	87
15	97	80
16	96	89
17	97	82
18	98	81
19	95	83
20	97	84
21	95	85
22	95	85
23	96	88
24	95	86
25	97	89

26	94	85
27	95	80
28	94	86
29	98	87
30	98	83
31	96	88
32	95	86
33	94	89
34	95	85
35	96	86

Table 5. Statistical independent samples T-test between Novel Multiple Logistic Regression and Linear Regression,confidence interval as 95%. The significance value is determined as 0.030 ($p<0.05$) for accuracy.

		F	Sig.	T	df	Sig((2-trailed)	Mean diff	Std. Error diff	Lower	Upper
Accuracy	Equal variances assumed	4.890	.030	18.900	68	<.001	10.543	.558	9.430	11.656
	Equal variances not assumed			18.900	59.33	<. 001	10.543	.558	9.427	11.659

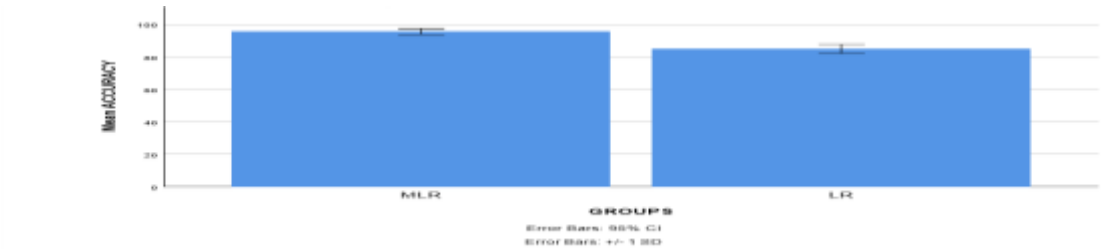

Fig 1. Bar Graph for Comparison of Algorithm and Accuracy. Mean accuracy of MLR is better than LR and standard deviation of MLR is slightly better than LR. X axis: MLR vs LR Y axis: Mean accuracy of detection ± 1 SD.