International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

Prediction of death rate among COVID-19 patients in the age group of 10 to 19 yrs. using machine learning

Debanjan Paul[1], Mayurima Sarkar[2], Dr. Avijit Kumar Chaudhuri[3], Sulekha Das[4], Moumita Chakraborty[5]
[1]*UG-Computer Science and Engineering, Techno Engineering College Banipur,*
[2]*UG-Computer Science and Engineering, Techno Engineering College Banipur,*
[3]*Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur,*
[4]*Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur,*
[5] *Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur,*
Corresponding Author Orcid ID: 0009-0007-0855-578X

ABSTRACT
COVID-19, a contagious disease which resulted in a worldwide pandemic just like "The Great Plague of Marseille" in 1720, "The First Cholera Pandemic" in 1820, and "The Spanish Flu" in 1920, claimed many lives. In December 2019 this disease was identified in the Chinese city Wuhan. After that researchers have found many relations in the ethnicity, occurrence, symptoms, severity, age and death. The study of these factors and their co-relation has helped in identifying who are at a greater risk of the disease and its repercussions.
In this paper, authors have predicted the death rate of COVID-19 patients (Sex, Current Status, ICU, Medical Condition, Hospitalization) in the age group 10-19 years. Data has been analysed through Logistic Regression Analysis (LR). The proposed model is tested on the "COVID-19_Case_Surveillance_Public_Use_Data" from the UCI Machine Learning Repository.
Keywords- Logistic Regression, Confusion Matrix, Methodology, 10Fold Cross Validation

1. Introduction

COVID-19(Coronavirus disease 2019), caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) was officially declared a pandemic on March 11, 2020. It created havoc across the world. However, it was noticed that the severity of the disease varies. Some people had no symptoms. Some had mild symptoms which were similar to upper respiratory disease, like a common cold. Whereas in certain people, symptoms were severe enough to require hospitalization. People with mild cases recovered at home typically within two weeks. However, among those hospitalized, some had to be taken to the ICU, or needed to be put on ventilators. While anyone can contract the disease, it most commonly affects middle-aged and older adults. As of 13 February 2023, there have been 6,836,825 deaths reported to WHO. It has been noticed that among those who died, most of them were in critical condition, were of advanced age, or suffered from some chronic condition (like diabetes mellitus, COPD, heart failure, or chronic kidney disease). Not only that, even gender and ethnicity seemed to affect death. According to epidemiological findings reported across different parts of the world, morbidity and mortality rate in males were higher than in females.
In this paper, a machine learning method is applied to investigate information regarding death during covid-19 dataset, to assess the prescient intensity of these systems. To get this aim, a logistic regression (LR) is first developed to predict covid positive in the first stage as the feature selection algorithm can affect the performance of the LR model. This enables the model to achieve better accuracy in the prediction and prognosis stages. Besides, the value of the parameter k in the LR is determined experimentally using an iterative approach. The performance of the proposed algorithm is assessed when it applies to a covid-19 database.

2. Literature Review

The ML models based on the algorithms like Multilinear Regression, Logistic Regression, Random Forest Classifier, and XG Boost Classifier are built using the two datasets. The R Squared (coefficient

DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

of determination) regression score and accuracy are computed with a train and test dataset ratio of 70:30. The feature importance for Covid-19-India Dataset is shown in the figure. The shows the Coefficient of Determination (CoD), also called R-Squared, and Accuracy for the models built on COVID-19-India Dataset. This shows the CoD and Accuracy for the models built on COVID-19-Data Dataset. The results show that the Random Forest Classifier and the Random Forest Regressor outperformed the other ML Models.

Sudhir Bhandari & Ajit Singh Shaktawat [1] made an analysis on prediction of mortality risk in COVID-19 patients from Routine Hematological parameters using logistic regression. They aimed to develop a predictor model of mortality risk from symptoms, random blood glucose and complete blood count. They observed that the statistically significant regressors are - differential neutrophil count (%) & random blood glucose (RBS in mg/dL). Their accuracy reached 70%.

Dr Satish Kumar et al. (2019) [10] have conducted a research work on predicting campus placement probability using binary logistic regression. They aimed to study the nature of campus placements and build a model to predict the probability of a random student that he or she will be placed or not. Also, they wanted to identify the factors that are influencing the placement chances of a student.

A new multiple regression model for predictions of urban water use S. Alireza Eslamian, S. Samuel Li ∗, Fariborz -Haghighat Department of Building [11], Civil and Environmental Engineering, Concordia University Montreal, QC H3G 1M8, Canada. This paper is about shortages of fresh water in many regions across the globe. They have collected the data and used MLR to predict the daily use of water.

Table 1: Dataset Overview

In this paper, the data we have been working on has been taken from the ***Kaggle Website*** [2]. The dataset is based on the deaths in COVID-19.

ATTRIBUTES	DEFINITION	MEAN	STANDARD DEVIATION
Current status	Patients are covid positive or not	0.941	0.94
Sex	male or female	0.528	0.49
Hospitalized Condition	hospitalized or not	0.733	0.45
ICU	whether shifted to ICU	0.542	0.50
Medical Condition	Patient's medical condition is good or bad during the hospitalized condition	0.991	0.78
Death	It is a dependent field on the above independent conditions that how many people died for those conditions.	0.677	0.46

3. Methodology

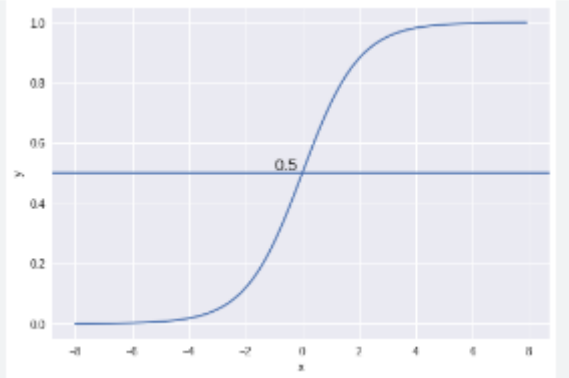
Machine Learning [4] is the usage and development of computer systems that can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. One such technique is called Multiple Logistic Regression (MLR). In MLR, the data is analyzed and some independent variables are identified against a single dependent variable.

As mentioned earlier, a statistical technique namely **Logistic Regression** (LR) has been used for classification of death rate among the age group of 10 to 19 years [3]. After thorough research of data set, authors have found that the death rate depends on different independent features, such as sex(x00), ICU(x04), hospitalization(x03), Medical condition(x05) & current status(x01). We now construct a stepwise logistic relation between them.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

3.1 Logistic Regression

Logistic regression [9] is mainly for classification. The biggest difference between it and linear regression is that its data points are not arranged in line rows. It may be a bunch here, there is a pile, each pile represents a category, and each type of data point has the same category label. For logistic regression, we want to find the boundary line of the classification, which is represented by the regression formula. The training classifier uses the optimization algorithm to find the best regression coefficient in the regression. Logistic regression-based classification is given an arbitrary set of inputs, and then the output is obtained by a function, which is the classification of the input data. For example, to simplify processing, when classifying, the function output 0 or 1 in the two classifications represents two classes. According to the actual needs and the above analysis, the above function argument range is from positive infinity to negative infinity. The dependent variable range is 0 or 1. There are many functions that satisfy the above conditions. The most intuitive one is the 0-1 step function. However, the step function is not steerable at the step point, which is not conductive to mathematical processing.



A few steps need to be followed in LR before moving forward toward our final model [7].

STEP 1: Checking Assumptions

The data set consists of 8 features. Out of this 8 features death rate (dependent feature of the data set) of 10 to 19 ages is fully dependent on 5 other features those are sex(x00), ICU(x04), hospitalization(x03), Medical condition(x05) & current status(x01). Next, we try to develop a logistic relation between the dependent & independent variables. We then split the data into two parts defined as training data and testing data and work on them in the following proportion: 80-20%, 66-34% & 50-50%.

Cross-validation or rotation estimation or out-of-sample testing is a method in which various parts of the data are tested and trained individuals and the accuracy of the model in practice is calculated. Here we divided the dataset into 10 parts, each time we select a part out of the 10 as the testing data and the remaining as training parts.

STEP 2: - We weren't satisfied with the accuracy we got by dividing the data sequentially. So, to obtain a higher accuracy, we have used 10-fold cross validation model.

STEP 3 – After applying the Multiple Linear Regression the value of dependent variables has been obtained. After that logistic regression has been applied for the value of dependent variables. Then predicted value of death of 10 to 19 ages has been obtained.

3.3 Mathematical Expression of Logistic Regression

b0=avg00-(b1*avg000+b2*avg001+b3*avg003+b4*avg004+b5*avg005)
y=b0+b1*x00[i]+b2*x01[i]+b3*x03[i]+b4*x04[i]+b5*x05[i]

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

$$b_i = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \quad \textbf{where I=1, 2....n}$$

$$p = \frac{e^{y}}{1 + e^{-y}}$$

Where,

b0 = coefficient of all the average independent variables of the cofactors.

b1 = coefficient cofactors of the first controlled variable (Sex Field).

b2 = coefficient cofactors of the second controlled variable (Current Status Field).

b3=coefficient cofactors of the third controlled variable (Hospitalized Condition Field).

b4 = coefficient cofactors of the fourth controlled variable (ICU Condition Field)

b5= coefficient cofactors of the fifth controlled variable (Medical Condition Field)

y00= variable representing death

x00= variable representing sex

x04= variable representing ICU

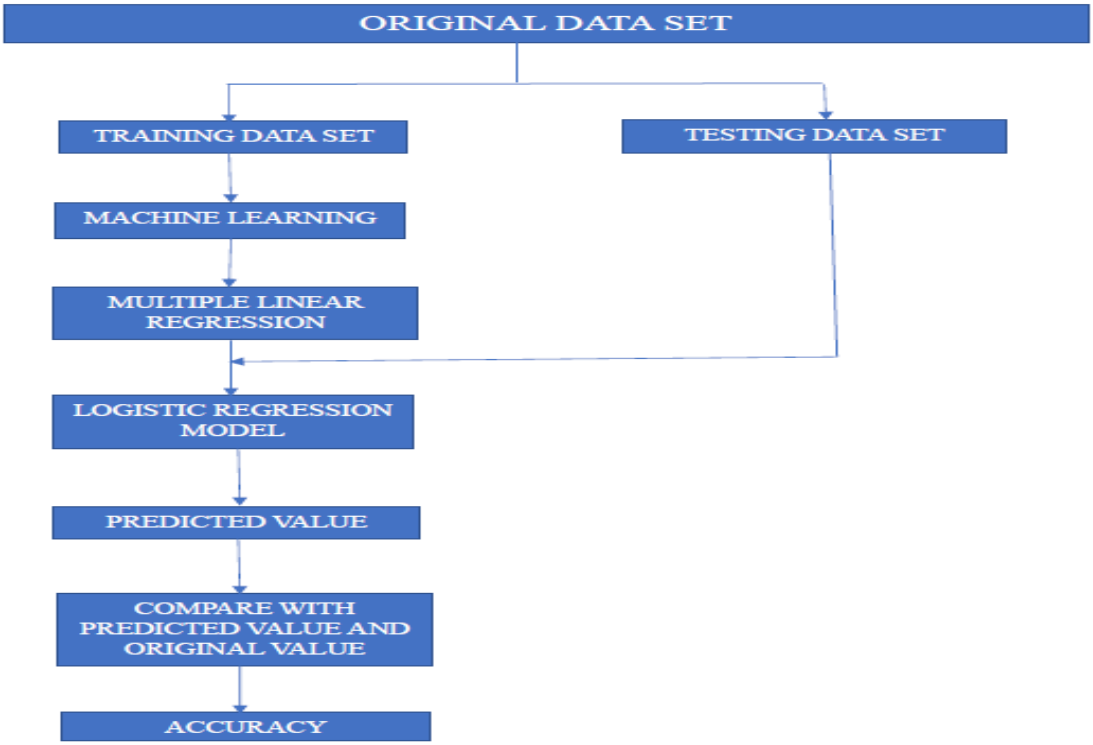
x03= variable representing the hospitalized condition

x05= variable representing the medical condition

x01= variable representing the current status

$\bar{y}$= Average of death variable

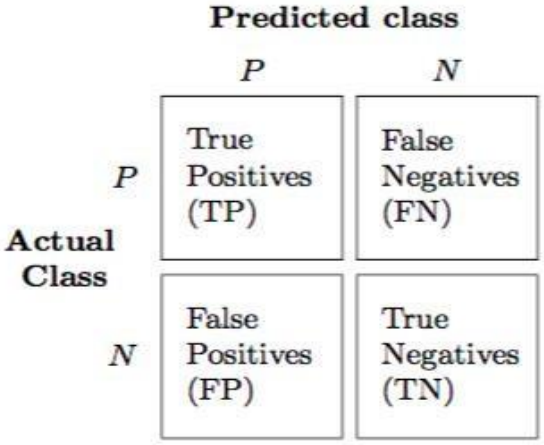
$\bar{x}$= Average of all independent fields particularly



Confusion Matrix: - A Confusion Matrix also known as an error matrix is a matrix that is used to determine the overall performance of an algorithm or a clarification model for a given set of test data.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

In the field of statistical analysis, a confusion matrix shows a set of test data for which the values are true or not.



TP=TRUE POSITIVE
TN=TRUE NEGATIVE
FP=FALSE POSITIVE
FN=FALSE NEGATIVE

Accuracy must be calculated to determine how precisely or how close the measured value reflects the originals.

Next, **Specificity** must be calculated. It refers to the test accuracy at identifying the probability of a negative test, provided the condition is absent.

After that, **Sensitivity** refers to the test accuracy in identifying the probability of a positive test, provided the condition is present.

The **precision** study refers to how precisely or accurately, the model is measured. We develop precision investigations to check whether we are getting the correct results or not.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad \text{Specificity}=\frac{TN}{TN+FP} \quad \text{Sensitivity}=\frac{TP}{TP+FN} \quad \text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad \text{F1\_Score}=\frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

After analyzing this model, we get the results that are given below.

Table 2: For 80-20% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	759 03 277 277
Accuracy	78.72
Sensitivity	99.60
Specificity	50.0
Precision	73.26
Recall	99.60
F1_Score	84.42

Table 3: For 66-34% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	1437 03 327 327
Accuracy	84.24
Sensitivity	99.79
Specificity	50.0
Precision	81.46
Recall	99.79
F1_Score	89.70

Table 4: For 50-50% train-test split

ATTRIBUTES	VALUE RANGE (%)
Confusion Matrix	1462 262 327 548
Accuracy	77.337
Sensitivity	84.80

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

Specificity	62.62
Precision	81.72
Recall	84.80
F1_Score	83.23

Table 5: For 10-fold cross-validation

TEST CASES	ACCURACY	SENSITIVITY	SPECIFICITY	STANDARD DEVIATION	Recall	PRECISION	F1_SCORE
01	60.93	95.07	50.0	0.456	95.07	37.84	54.13
02	72.22	100.0	50.0	0.466	100.0	61.53	76.19
03	91.53	99.78	50.0	0.477	99.78	90.94	95.16
04	82.14	100.0	50.0	0.473	100.0	78.26	87.81
05	73.84	89.74	50.0	0.472	89.74	72.91	80.45
06	57.49	98.56	50.0	0.449	98.56	26.44	41.70
07	73.96	100.0	50.0	0.468	100.0	64.80	78.64
08	93.86	100.0	50.0	0.479	100.0	93.46	96.62
09	89.19	100.0	50.0	0.477	100.0	87.88	93.55
10	70.43	99.01	50.0	0.464	99.01	58.60	73.63

CONCLUSION

While there are many machine-learning methods available in the literature whose performances depend on different aspects including the dataset they are applied to, in this paper, a machine-learning method called logistic regression, is applied to classify the risks of covid positive patients. Moreover, the best value for the number of neighbors in the logistic regression method was determined using an algorithm coded in Python. These experiments and results were analyzed carefully by a covid specialist. Moreover, the specialist analyzed the outcomes using some patients' clinical data and compared their real condition with the class that the model devoted to that patient. Future works may involve the use of other machine-learning classification algorithms or employing other population-based feature selection meta-heuristics and comparing their performances to the one obtained by the proposed approach.

References

1. Bhandari, S., Shaktawat, A., Tak, A., Patel, B., Shukla, J., Singhal, S., ... & Dube, A. (2020). Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina Journal of Medicine and Biomedical Sciences, 12(02), 123-129.
2. The Covid-19 Dataset is taken from Kaggle Website (https://www.kaggle.com/datasets/imdevskp/corona-virus-report).
3. Koushik Paul, Saheb Karan, Siddhartha Kuri, Sulekha Das, Avijit Kumar Chaudhuri "Placement Prediction Using Multiple Logistic Regression Method". International Journal of Advanced Research in Computer and Communication Engineering. ISSN (O) 2278-1021 ISSN (P) 2319-5940. Volume 11, Issue 3, March – 2022.
4. Saha, S., Mondal, J., Arnam Ghosh, M., Das, S., & Chaudhuri, A. K. Prediction on the Combine Effect of Population, Education, and Unemployment on Criminal Activity Using Machine Learning.
5. D. Satish Kumar, Zailan Bin Siri, D.S. Rao, and S. Anusha, "Predicting Student's Campus Placement Probability using Binary Logistic Regression". International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019
6. S. Taruna, and Mrinal Pandey, "An Empirical Analysis of Classification Techniques for Predicting Academic Performance". 2014 IEEE International Advance Computing Conference (IACC).
7. Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy', "A Comparative Analysis of Techniques

DOI:10.46647/ijetms.2023.v07i02.036 ISSN: 2581-4621

for Predicting Academic Performance", 37th ASEE/IEEE Frontiers in Education Conference, IEEE, 2007.
8. Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, and Keshav Kumar, "PPS - Placement Prediction System using Logistic Regression". 2014 IEEE International Conference on MOOC, Innovation, and Technology in Education (MITE).
9. Logistic Regression, Wikipedia, https://en.wikipedia.org/wiki/Logistic_regression#cite_note-1
10. Kumar, D. S., Siri, Z., Rao, D. S., & Anusha, S. (2019). Predicting student's campus placement probability using binary logistic regression. International Journal of Innovative Technology and Exploring Engineering, 8(9), 2633-2635.
11. Eslamian, S. A., Li, S. S., & Haghighat, F. (2016). A new multiple regression model for predictions of urban water use. *Sustainable Cities and Society*, *27*, 419-429.