Comparative Analysis of Machine Learning Algorithms in Predicting Risk of Death in COVID-19 Patients

Aasika ES[1a], Jeeva D[1b], Sangavi G[1c], DR. Lakshmanan S[2*]

[1]PG Student, Division of Mathematics, School of Advanced Sciences, VIT Chennai

Email: [a]ashika.2023@vitstudent.ac.in

[b]jeeva.2023@vitstudent.ac.in

[c]sangavi.g2023@vitstudent.ac.in

[2]Associate Professor, Division of Mathematics, School of Advanced Sciences, VIT Chennai

Email: [2]lakshmanan.s@vit.ac.in

ABSTRACT:

This paper is a comparative analysis of machine learning (ML) algorithms for predicting the risk of death in COVID-19 patients. The proposed study analyzes the performance and accuracy of the ML algorithms using relevant medical data. The study includes data importing, preprocessing, and visualizing the data of COVID-19 patients. This data set contains 10,485,75 data points and 21 attributes. Subsequently, numeric data scaling is implemented here to balance the imbalanced data set. After balancing, the data set will contain 10,251,52 data points and 19 attributes. Consider one particular algorithm among the other ML algorithms that offers the best accuracy in predicting mortality risk. This study is compared with four different algorithms, such as logistic regression, random forest, decision tree, and support vector machine. From the outcome, we can conclude that the decision tree has more accuracy than other algorithms.

Keywords:
Machine-learning Algorithms; Covid-19 patient; Risk of death prediction; Decision tree accuracy

1. INTRODUCTION:

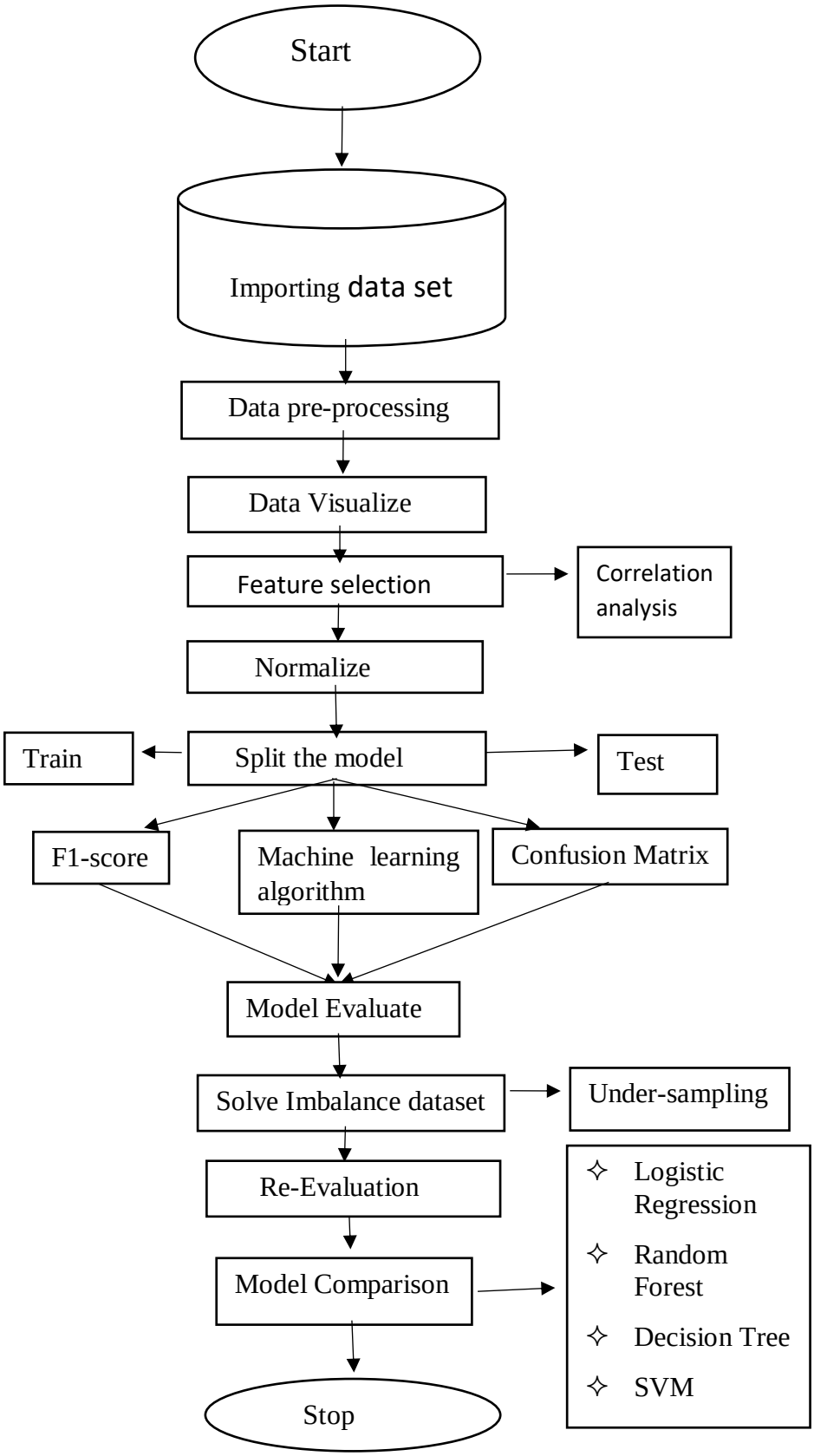
Corona virus (COVID-19) disease is considered to be a contagious transmissible condition typically caused by the novel Corona virus. The individuals with pre-existing health conditions are more vulnerable to get affected by Covid-19.Then there arises the crucial need for modern predictive measures to approach the high risk of death among infected individuals.

Machine learning(ML) algorithms have the potential to bring out high-quality predictions from the raw data sets of affected patients. Regarding this condition, our study aims to establish a comparative analysis of Machine learning algorithms including Logistic regression(LR), Decision tree(DT), Random forest(RF), and Support Vector Machines(SVM) to predict the high-risk of mortality in Covid-19 patients. The presented system in this paper comprises data importing, data preprocessing, balancing missing values in the data set, data visualization, and feature selection from the data of patients which includes their medical history, physiological conditions, and demographic details. So after applying those four algorithms of the Machine learning model, we will be able to identify the best predictive algorithm model with more accuracy rate which helps in finding out the patients who are all at a high risk of death. By this study, we get a solution for better prediction of mortality in Covid-19 patients.
The authors in [1] highlighted the application of machine learning (ML) methods to predict mortality in hospitalized COVID-19 patients was emphasized. The random forest (RF) method fared notably better than the others with an accuracy of 95. 03%. In [2], the authors compared eighteen machine learning algorithms for COVID-19 patients in order to predict ICU admission and mortality. The authors created models using a sizable data set,

validated them both internally and externally, and showed that ensemble-based models perform better than other models in predicting mortality. In [3] addressed about the logistic regression algorithm's improved calibration and discrimination metrics, COVID-19 Community Mortality Risk Prediction (CoCoMoRP), an online prediction tool available to the public, was developed using it.The author in[4] was to create a model that would forecast the prediction of individuals with severe COVID-19 infection.The prediction model used is logistic regression model because of its interpretation. An AUROC of 0. 881 with a 50% threshold likelihood of death was obtained through external validation. In [5], they discovered five important variables—incubation, vasopressor demand, age, gender, and PaO2/FiO2 ratio—integrated into the DT model for predicting 28-day ICU outcomes by valuating data from 14 hospitals to help identify patients who are in higher risk in order to effectively allocate and make crucial decisions.The author[6] demonstrated that Machine learning models are able to provide the most precise and accurate form of results of predicting the death rate in the 28-day period of in-hospital care for the severely affected patients. In study[7] of a large number of datasets which contains the list of over 2.6 million patients in 146 countries worldwide by making sure of diversity. In this paper, the overall accuracy of prediction is found out to be 89.98% by applying various Machine learning algorithms such as Random forests and Neural forests model becomes more reliable for prediction.This study[8] comprises of both the Machine Learning and Deep Learning key algorithms and conceptual models to detect the outcome of the affected patients of COVID-19 by the application of Machine learning algorithms like Logistic Regression, Decision Tree, Random forest, Extreme gradient boosting and K-nearest neighbour .The implementation in the study[9] of Machine learning algorithms With the data set, they applied a well-developed line of action to select the features in their data set by the fusion of Lassocv, Spearman's correlation of rank to recognize the required research-oriented features with the timely execution of Random forest, Logistic regression and Support Vector Machines to get the desired result. The incorporation in the study [10]of available insights that are collected from virologists and also through an online survey established a well-precise Machine learning model for the case of prediction for the mortality rate of affected patients during the Covid-19 pandemic.The efficacy of this study [11]of various machine-learning algorithms in forecasting the death risk of COVID-19 patients. This study highlights random forest approach performed the best with an accuracy of 95.03%, sensitivity of 90.70%, specificity of 95.10%, precision of 94.23%, and ROC value of 99.02% compared to the other models. The study in[12]compared the performance of various machine learning methods, including random forest (RF), k-NN, Boost, and deep learning. Boost had the best prediction performance, with a 99.7% accuracy. The paper in[13]highlighted that the effectiveness of three alternative ML models for calculating the chance of critical COVID-19 depending on patient condition at admission was examined. The study in [14] explores the model that was built with a number of machine learning methods, including the least absolute shrinkage and selection operator (LASSO), linear support vector machine (SVM), SVM with radial basis function kernel, random forest (RF), and k-nearest neighbors (KNN).The paper in [15]is to highlight the machine learning (ML) approaches to classify the mortality of persons with underlying health issues affected by COVID-19 by using Logistic regression, Random Forest, Support vector machine, Naive Bayes, and Threshold selector. The main objective of this study mentioned here as follows,

- In our paper we take 10 lakh of data points with 21 attributes to enhance the accuracy with comparison of different models.

2. Flow of work:



3. Methodology:

3.1. Data Importing:

Collect and import the data in CSV format. The data set contains patient records like Age, Sex, Classification, Symptoms. Import some library files Such as Numpy, Pandas, matplotlib, and Seaborn. This data set contains “1048575” data points and “21” Attributes. The Attributes are Sex, age, classification, patient type, pneumonia, pregnancy, diabetes, COPD, asthma, inmsupr, hypertension, cardiovascular, renal chronic, other disease, obesity, tobacco, USMER, medical unit, Intubed, ICU, death.

3.2. Data Pre-processing:

Clean and pre-process the data set to deliver the data quality issues, such as missing values, duplicates, and outliers. We find the missing value percentage in our data set. The most missing in data are Pregnant, Intubed, and ICU. Encoded some features to be converted into binary classification. Encoded the death-died features: 2 for “Alive” and 1 for “Death”. For missing values in pregnant encoded as 2. 1 means “yes” and 2 means “no”. Removed the Intubed and ICU features. After pre-processing the data set contains “1025152” data points and “19” attributes. The attributes are USMER, Medical unit, sex, patient type, pneumonia, age, pregnancy, diabetes, COPD, asthma, Inmsupr, hypertension, other disease, cardiovascular, obesity, renal-chronic, tobacco, classification, and death.

3.3. Data Visualization:

Visualize the data set to find out about variable distribution, correlations, and probable trends. By using various kinds of plots, graphs, and charts to understand our data.

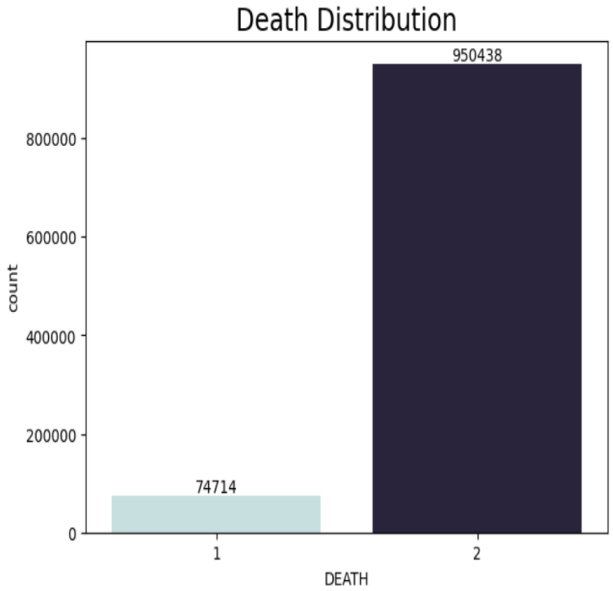


Fig 1. Represent the bar plot of death distribution

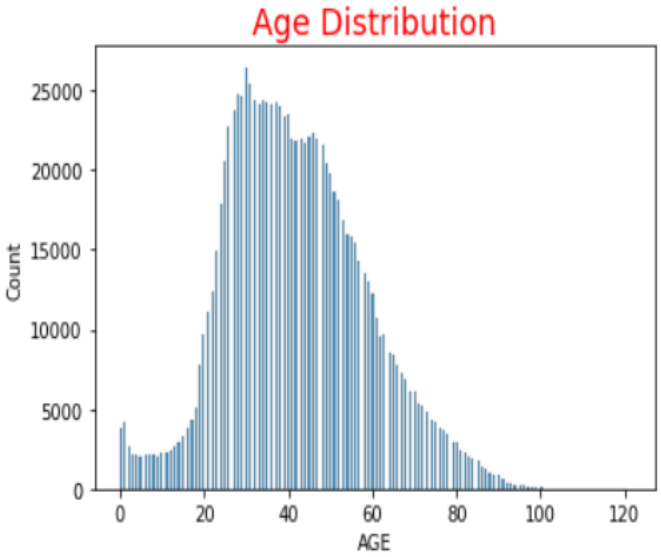


Fig 2. Represent the histogram of Age distribution

Fig 2. Shows Between 20 to 60 age people are chance to die. The chance to death this group age people .They intake lack of healthy foods, lack of regular exercise, consuming smoking and alcohol for this reason they have lack of immune power so easily they have a chance to die.

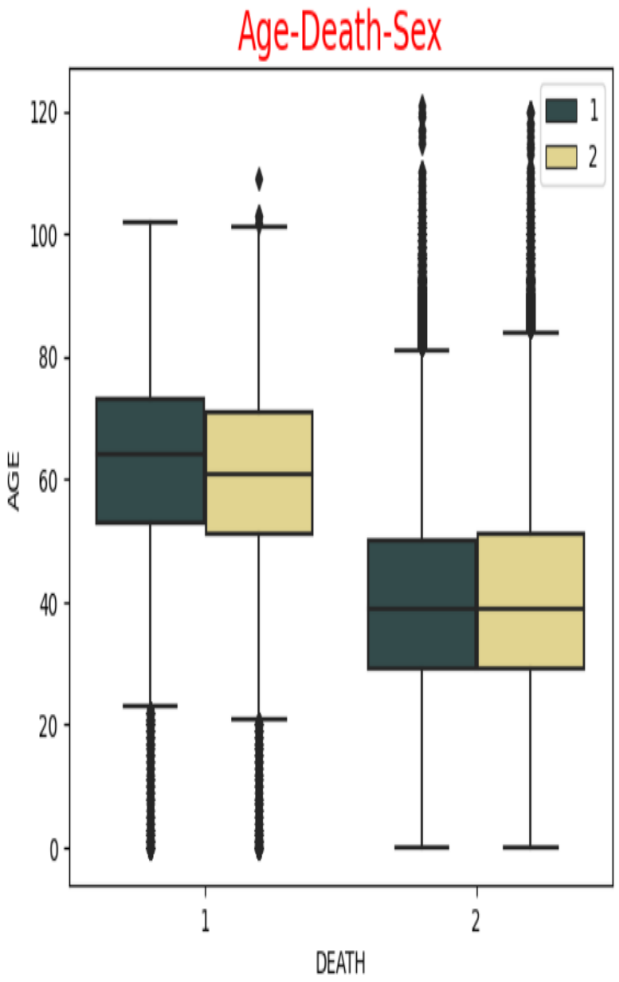


Fig 3. Represent the box plot of age-death-sex

Fig 3. shows that older age people have a chance to die compared to young ones. There is no difference between men and women. Everyone dies eventually, regardless of their age or sex.

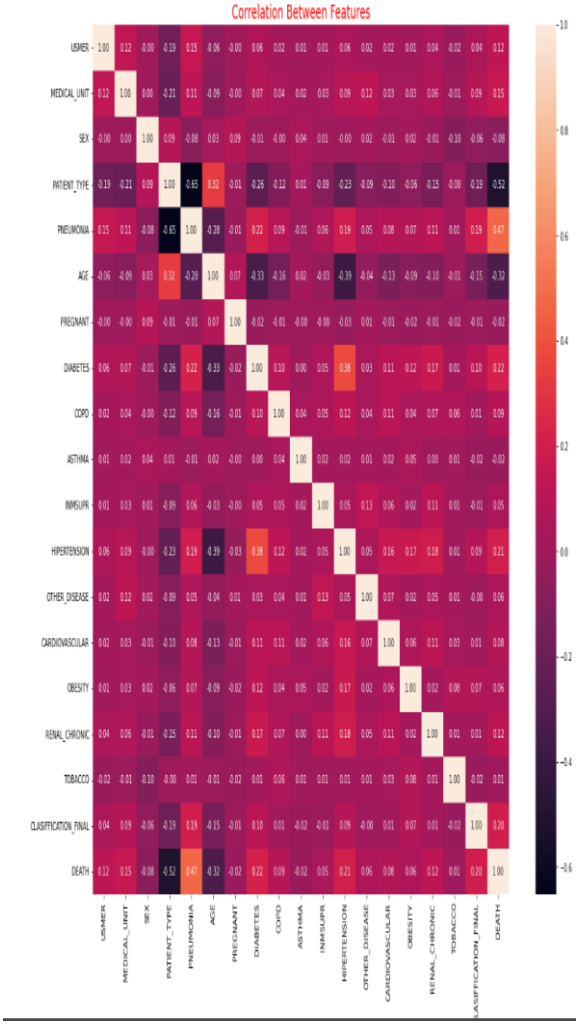


Fig 4. Represent the correlation between the features

Fig 4. Says by using correlation to remove some features that do not have a positive correlation with the death feature. Calculate the Karl Pearson correlation coefficient between each feature and the death feature. Remove any feature with a correlation coefficient less than zero.

3.4. Feature Selection and Scaling:

Choose features depending on their unique properties, such as statistical significance or association with the variable being studied. By using correlation remove some features which do not have a positive correlation.Scaling numeric features plays an important role during model training. Normalize the numeric feature. There are some ways to normalize the data. Here handled with

the standardized process to normalize the data. The numeric feature is “Age”

Model Evaluation Techniques:

Table I. Confusion matrix

Actual class/ Prediction class	**C1**	**¬ C1**
C1	**TP**	**FN**
¬ C1	**FP**	**TN**

Table I. indicates the Confusion matrix for actual and prediction classes. ¬ C1 refers to the model predicted as negative.TP – True positive, FP – False positive, TN – True Negative, and FN-False Negative.

Accuracy:

The rate at which my data points are correctly classified

- Accuracy= (TP +TN) / (TP+FP+FN+TN)

Precision, Recall, and F-measures (F1 or F-score):

Precision is the true positive out of all predicted positives. Recall is the true out of all actual positives. The F1-score is the mean of precision and recall.

- Precision=TP/TP + FP
- Recall=TP/TP+FN
- F1=2*Precision*Recall/Precision*Recall

3.5. Creating Model and Model Evaluation:

Split the data into train and test and choose appropriate machine learning algorithms for classification tasks. The target variables are essential in machine learning. The target variable represents what we are trying to forecast. In this paper, the target variable is "death," which is represented as y, and the feature variable is x. The feature variable is used to predict the target variable. By using F1-score and confusion evaluate the model whether it is balanced or imbalanced.

- F1-score = (0.50658847 0.96717964)

Here the majority class is dominant the minority class

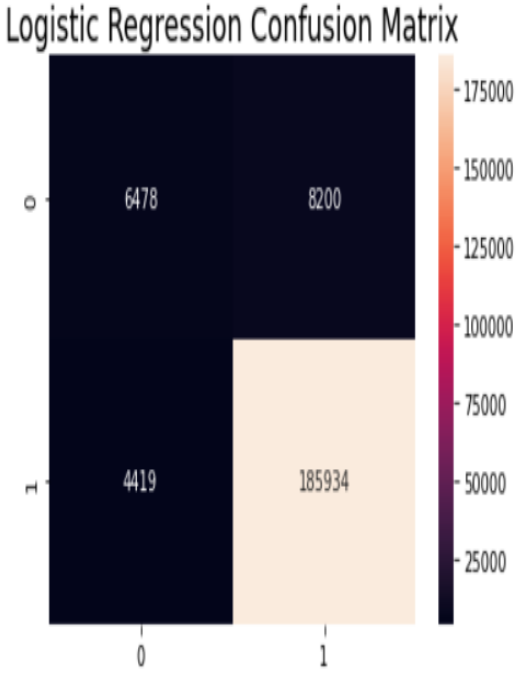


Fig 5. Represent the confusion matrix for logistic regression

Fig 5. Uses whether the data set is balanced equally. In binary classification Confusion matrix helps to evaluate whether the data set is equally Balanced. Here is Fig 5. Represent the data set is not balanced equally.

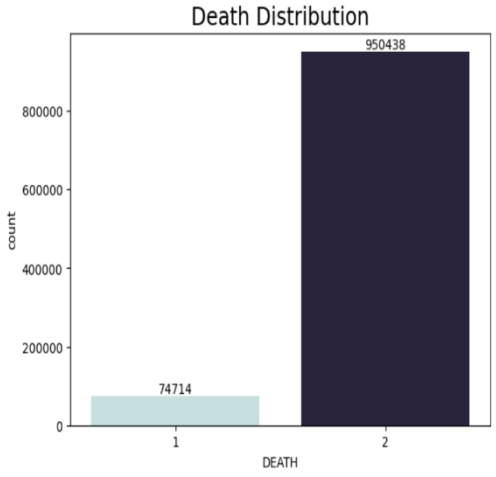


Fig 6. Represent the bar plot of death distribution here the data are imbalanced.

3.6. Solving Imbalance Data set Problem with Under-sampling:

There are two types of sampling under-sampling and over-sampling.

If under-sampling is not possible to balance our data set then proceed with over-sampling.

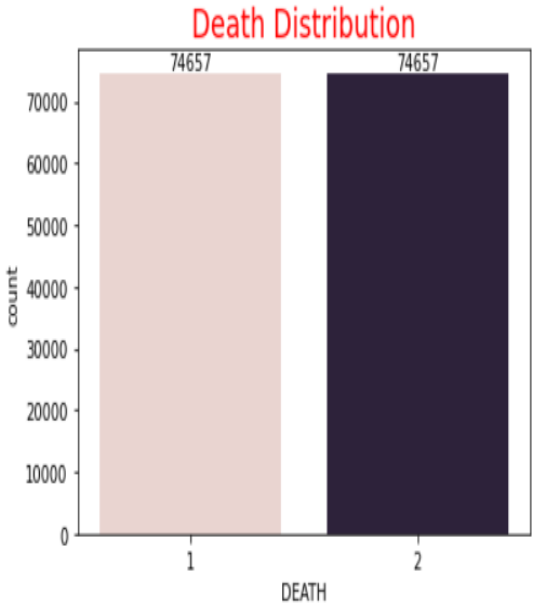


Fig 7. Represent the bar plot that the data are equally balanced

Fig 7. Highlights that the data are balanced after the under-sampling. By doing under-sampling before we check with the help of the confusion matrix and F-measure whether the data are eventually balanced.

3.7. Model Re-Evaluation:

Re-evaluate the model after under-sampling by using the F1-score and Confusion Matrix.

After under-sampling f1-score:
[0.91204116 0.9094428]

After under-sampling Confusion Matrix is

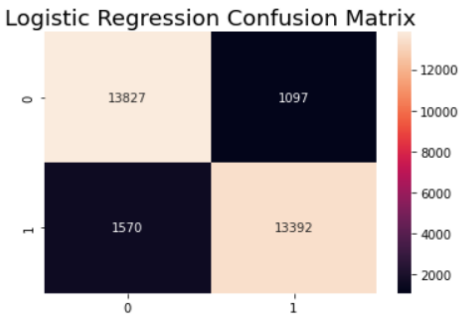


Fig 8. Represent the re-evaluate the data in the confusion matrix

Fig 8. Indicates the data are balanced. Now we are able to proceed with the model comparison. Otherwise, we can't get the proper accuracy rate in model comparison.

3.8. Model Comparison:

Logistic regression, decision classifiers, random forest classifiers, support vector Machine Classifier (SVM). These are our approaches to compare which will give more accuracy. The accuracy in logistic regression is 91.08%, the decision tree is 91.58%, the random forest is 90.56% and SVM is 91.17%

4. Conclusion and Future work:

This paper compares machine learning techniques for predicting the risk of death in COVID-19 patients. It is beneficial to study the data during the data visualization process. The histogram demonstrates that people aged 20 to 60 can be easily affected and die. The plot in Box demonstrates how elderly people die compared to young people (sex, pregnancy, asthma, COPD, cardiovascular, obesity, other disorders, Tobacco, and inmsupr).Correlation aids in the selection of data set features. It eliminates features that do not have a positive link with death features. Scaling normalizes the numeric feature "age," which aids in accuracy when employing a machine-learning algorithm. The under-sampling procedure aids in the balancing of the data set.

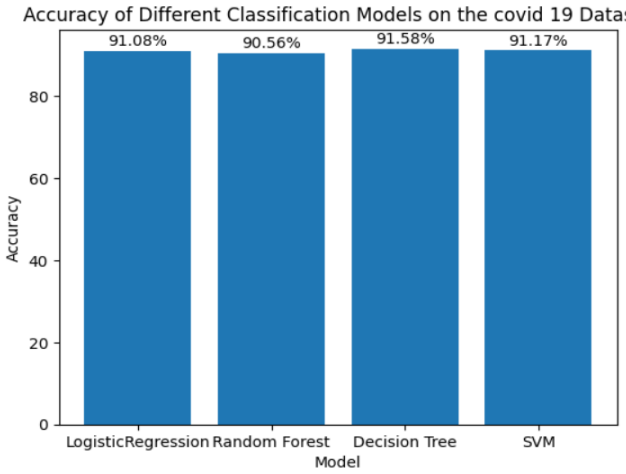


Fig 9. Represent the bar plot of the accuracy of different models

Fig 9. Indicates in the Model comparison, the "decision tree classifier" has more accuracy with 91.58% for our data set.

In the future, studies will combine some clinical data modalities such as imaging data, genetic data, and social determinants of health to increase the accuracy of risk prediction models. Concentrate on creating interpretative machine learning models that can assist in predicting which patients are at high or low risk of mortality.

REFERENCES:

[1] Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. *BMC* medical informatics and decision making, 22(1), 1-12.

[2] Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., ... & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ* digital medicine, 4(1), 87.

[3] Das, A. K., Mishra, S., & Gopalan, S. S. (2020). Predicting community mortality risk due to CoVID-19 using machine learning and development of a prediction tool. *MedRxiv*, 2020-04.

[4] Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., ... & Chen, X. (2020). Early prediction of mortality risk among patients with severe COVID-19, using machine learning. International journal of epidemiology, 49(6), 1918-1929.

[5] Elhazmi, A., Al-Omari, A., Sallam, H., Mufti, H. N., Rabie, A. A., Alshahrani, M., ... & Arabi, Y. M. (2022). Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. Journal of infection and public health, 15(7), 826-834.

[6] Churpek, M. M., Gupta, S., Spicer, A. B., Hayek, S. S., Srivastava, A., Chan, L., ... & Leaf, D. E. (2021). Machine learning prediction of death in critically ill patients with coronavirus disease 2019. Critical Care Explorations, 3(8).

[7] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart health, 20, 100178.

[8] Khan, I. U., Aslam, N., Aljabri, M., Aljameel, S. S., Kamaleldin, M. M. A., Alshamrani, F. M., & Chrouf, S. M. B. (2021). Computational intelligence-based model for mortality rate prediction in COVID-19 patients. International journal of environmental research and public health, 18(12), 6429.

[9] Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., Huang, L., & National Traditional Chinese Medicine Medical Team. (2022). Comparing different machine learning techniques for predicting COVID-19 severity. Infectious diseases of poverty, 11(1), 19.

[10] Agbelusi, O., & Olayemi, O. C. (2020). Prediction of mortality rate of COVID-19 patients using machine learning techniques in nigeria. International journal of computer science and software engineering, 9(5), 30-34.

[11] Anandhanathan, P., & Gopalan, P. (2021). Comparison of machine learning algorithm for COVID-19 death risk prediction.

[12] Kivrak, M., Guldogan, E., & Colak, C. (2021). Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. Computer methods and programs in biomedicine, 201, 105951.

[13] Assaf, D., Gutman, Y. A., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., ... & Tirosh, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine, 15, 1435-1443.

[14] An, C., Lim, H., Kim, D. W., Chang, J. H., Choi, Y. J., & Kim, S. W. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Scientific reports, 10(1), 18716.

[15] Mohammad, M. A., Aljabri, M., Aboulnour, M., Mirza, S., & Alshobaiki, A. (2022). Classifying the mortality of people with underlying health conditions affected by COVID-19 using machine learning techniques. Applied Computational Intelligence and Soft Computing, 2022.