DETERMINING THE EFFICIENCY OF PREDICTING COVID-19 INFECTED CASES USING DATA MINING AND MACHINE LEARNING APPROACHES

N. Sankar[1*], S. Manikandan[2]

Article History: Received: 16.01.2023 **Revised:** 01.03.2023 **Accepted:** 14.04.2023

Abstract

Since the start of 2020, the novel coronavirus has spread extensively, prompting numerous efforts to develop vaccines for patient recovery. It is evident now that a swift solution is required worldwide to control the spread of COVID-19. Amidst the COVID-19 pandemic, vaccines have been developed and granted emergency authorization to mitigate the occurrence of COVID-19. Extensive evidence supports the high efficacy of these vaccines in preventing severe illness, hospitalization, and fatality resulting from COVID-19. They serve as a vital resource in curbing the virus's transmission and ultimately ending the pandemic. Non-clinical approaches, such as data mining and machine learning techniques, hold promise in alleviating the strain on healthcare systems and providing an optimal diagnosis of the pandemic. The dataset contains 36 states/UTS, total cases, death cases, and total dose. In this paper, Machine Learning (ML) algorithms were employed to analyze and predict the COVID-19-related dataset for finding the suitable parameters for prediction using M5P and Random Forest. The performance and their accuracy of these algorithms was evaluated using precision, recall, accuracy, and f-measure metrics. Numerical illustrations were also provided to prove the proposed research.

Keywords: COVID-19, Data Mining, Machine Learning, and Accuracy.

[1*]Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India
[2]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram - 608 102, Tamil Nadu, India
Email: [1*]nsankarraj@gmail.com, us.mani.s.mca@gmail.com

[*] Corresponding Author: N. Sankar[1*]
[1*]Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India
Email: [1*]nsankarraj@gmail.com

DOI: 10.31838/ecb/2023.12.1.572

1. Introduction

COVID-19, also called the coronavirus disease 2019, is called illness caused by the SARS-CoV-2 virus. It was initially identified in China, for the year of December 2019 and has since spread globally, resulting in a pandemic. The primary transmission mode for COVID-19 is respiratory droplets expelled when an infected individual coughs, sneezes, talks, or breathes. Additionally, it can be contracted by touching contaminated surfaces and then touching the face. The symptoms of COVID range between mild to severe and can include different symtoms namely fever, cough, shortness of breath, fatigue, muscle or body aches, sore throat, loss of taste, smell, headache, and in some cases, diarrhea. It's important to note that some individuals infected with the virus may remain asymptomatic, meaning they display no symptoms and can still transmit the disease to others. COVID-19 can potentially cause severe respiratory complications and poses a higher risk to older and individuals for affect health conditions like heart disease, diabetes, or weakened immune systems. The virus has led to a significant number of hospitalizations and deaths worldwide. Data mining involves exploring extensive datasets to uncover patterns, trends, and valuable insights. This process involves analyzing and manipulating raw data to extract meaningful information across various domains, including business intelligence, scientific research, and decision-making. Data mining employs diverse techniques and algorithms, such as statistical methodologies, machine learning, and artificial intelligence. These methodologies enable the identification of hidden patterns and relationships within data that may not be immediately apparent, facilitating more informed decision-making processes. One common application of data mining is the identification of consumer behavior patterns, such as purchase histories or browsing habits, which can enhance marketing and sales strategies. It also plays a crucial role for detecting fraudulent activities, discerning trends in healthcare data, and even predicting natural disasters. In summary, data mining is a potent instrument for extracting meaningful insights from large datasets, consequently becoming an indispensable component in numerous industries and fields.

Machine learning, a subset of artificial intelligence, encompasses the creation of algorithms and models that empower computers to enhance their performance through experiential learning. Unlike traditional computer programming, where developers explicitly code tasks, machine learning enables computers to acquire knowledge and refine their abilities by leveraging data and experience. This acquired knowledge then enables the computer to make predictions or decisions autonomously. This research considers the analysis and predictions for the covid dataset using two machine learning approaches, M5P, and Random Forest. The ML approaches return different accuracy parameters: correlation coefficient, Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). The test statistics accuracy returns a strong positive correlation with a minimum error rate. The prediction is reasonable.

Review of Literature

The author discusses different studies related to COVID-19 between 20/01/2020 and 18/09/2020 for the USA, Germany, and the global obtained from World Health Organization. Datasets consist of weekly confirmed and cumulative confirmed cases for 35 weeks. Then the data distribution was examined using the most up-to-date Covid-19 weekly case data, and its parameters were obtained according to the statistical distributions. Furthermore, a time series prediction model using machine learning was proposed to obtain the disease curve and forecast the epidemic tendency. Linear regression, multilayer perceptron, random forest, and support vector machines (SVM) machine learning methods were used. The performances of the methods were compared according to the RMSE, APE, and MAPE metrics, and it was seen that SVM achieved the best trend. According to estimates, the global pandemic will peak at the end of January 2021, and approximately 80 million people will be cumulatively infected [1]. Focus on the COVID-19 World Vaccination Progress using Machine learning classification Algorithms—the findings of the paper show which algorithm is better for a given dataset. Weka is used to run tests on real-world data, and four output classification algorithms (Decision Tree, K-nearest neighbors, Random Tree, and Naive Bayes) are used to analyze and draw conclusions. The comparison is based on accuracy and performance period, and it was discovered that the Decision Tree outperforms other algorithms in terms of time and accuracy [2]. Various researchers try to determine the probability of COVID-19 recovery in South Asian Countries based on healthy diet patterns using data mining and various machine learning algorithms. We have used Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), several machine learning algorithms to predict the recovery rate of Covid-19 affecting patients [3]. The researchers have applied several existing data mining methods (classifiers) available in the Waikato Environment for Knowledge Analysis (WEKA) machine learning library. WEKA was used to understand better how the epidemic spread within Zambia. The classifiers used are the J48 decision tree, Multilayer Perceptron, and Naïve Bayes. The predictions of these techniques are compared against simpler classifiers and those reported in related works [4]. Machine learning algorithms include support vector machines, adaptive boosting, random forest, and k-nearest neighbors. These algorithms are then merged to form ensemble learning which leads to the classification. The results show ensemble learning has the highest actual positive rate of 30%. The obtained results show that regular blood tests do not help much in giving the proper indications for detecting COVID-19 [5]. They are extracting risk factors from clinical data of early COVID-19-infected patients and utilizing four types of traditional machine learning approaches, including logistic regression (LR), support vector machine(SVM), decision tree(DT), random forest(RF), and a deep learning-based method for diagnosis of early COVID-19. The results show that the LR predictive model presents a higher specificity rate of 0.95, AUC of 0.971, and an improved sensitivity rate of 0.82, which makes it optimal for the screening of early COVID-19 infection. We also perform the verification for the generality of the best model (LR predictive model) among the Zhejiang population and analyze the contribution of the factors to the predictive models [6]. Prediction and their performance of death based on the clinical factors (including COVID-19 severity) by data mining methods. Methods: The dataset consists of 1603 SARS-COV-2 patients and 13 variables obtained from an open-source web address. The current dataset contains age, gender, chronic disease (hypertension, diabetes, renal, cardiovascular, etc.), some enzymes (ACE, angiotensin II receptor blockers), and COVID-19 severity, which are used to predict death status using deep learning and machine learning approaches (random forest, k-nearest neighbor, extreme gradient boosting [XGBoost]). A grid search algorithm tunes hyperparameters of the models, and predictions are assessed through performance metrics. Knowledge discovery steps in databases are applied to obtain the relevant information [7].

The authors explain various data mining and machine learning algorithms with accuracy for various decision tree approaches using the WEKA tool to stumble on essential parameters of the tree structure. Seven classification algorithms such as J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP), and Random Forest (RF) are used to measure the accuracy. The data mining tool WEKA (Waikato Environment for Knowledge Analysis) has been used for finding experimental results of weather data sets. Out of seven classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [8]. The main objective of this paper is to analyze the SDGs by various independent metrics in the states of Tamil Nadu, Kerala, and Karnataka in India and by taking into consideration three different state SDGs index using data mining and statistical approaches for retrieving various hidden information. Numerical illustrations are also used to prove the proposed results [9].

Background Knowledge

M5P: M5P [10] is a reconstruction of Quinlan's M5 algorithm [11] for inducing trees of regression models. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. First, a decision-tree induction algorithm is used to build a tree. Still, instead of maximizing the information gain at each inner node, a splitting criterion is used that minimizes the intra-subset variation in the class values down each branch. The splitting procedure in M5P stops if the class values of all instances that reach a node vary slightly or only a few instances remain. Second, the tree is pruned back from each leaf. When pruning, an inner node is turned into a leaf with a regression plane. Third, to avoid sharp discontinuities between the subtrees, a smoothing procedure is applied that combines the leaf model prediction with each node along the path back to the root, smoothing it at each node by combining it with the value predicted by the linear model for that node.

Random Forest (RF)

RF is a standard machine learning decision tree algorithm that belongs to supervised learning methods. In these approaches, working principles are based on classification and regression. RF is generally called ensemble learning, which combines different classifiers to solve various problems with enhanced performance of the model. The Random Forests classifier, compared to others, is the best classifier capable of precisely classifying a massive amount of data. RF decision tree approaches mainly focus on learning procedures for classification and regression methods; it will create many decision trees and the level of the tree at training time for outputs of the class with classes output from single trees [12] and [13].

Step 1:	Select randomly for data k facts of the training stage.
Step 2:	The decision tree creation and associated with the carefully selected given subsets.
Step 3:	Select the total number of N decision trees you desire to build the tree and its level.
Step 4:	The steps follow from 1 to 3 until satisfied with the corresponding condition.

Step 5:	Input new data points to find estimates for every decision tree and allocate the new data points to the corresponding group functions that win.

Correlation Coefficient (CC)

The CC, or coefficient of determination, denoted R2 or r2 score, is used to moderation in the dependent variable means predicted from the independent variables. The r (CC) returns nearly 1.0, which means a strong positive correlation. If the value of r returns nearly -1 means a robust negative correlation, and a return of 0 means no correlation between all the variables [14].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] - [n\sum y^2 - (\sum y)^2]}} \qquad \dots (1)$$

Relative Absolute Error (RAE)

The RAE is used to compute the accuracy for a relative comparison of every predictive model performance—the main reason for calculating the RAE between actual and forecasted value. RAE is very useful for writing the interpretation of the prediction, which means if the RAE <1 means, the model behavior is better. If RAE=0, the model behavior or accuracy is perfect [15]

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|} \qquad \dots (2)$$

Where n is the number of elements in the observations, y(i) is the realized value, y͡(i) is the prediction, and y¯ means the mean values of corresponding variables.

Root Relative Squared Error (RRSE)

RRSE is one of the accuracy metrics for predictive models called regression. It's an accuracy parameter that is used to compute the first result and behavior of the model is performing. It is also an inheritance from RSE. The RRSE parameter for finding the process of square root for the sum of squared errors for the corresponding predictive model with the sum of squared errors.

$$\mathrm{RRSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \dots (3)$$

Where n is called the number of elements in the observations, y(i) is called the realized value and y͡(i) is called the prediction, and ȳ means the mean values of corresponding variables.

Numerical Illustrations

Table 1: Covid'19 Dataset

State Name/UTS	Total Case	Death	Total Dose
Andaman & Nicobar	10751	129	991263
Andhra Pradesh	2339197	14733	110956778
Arunachal Pradesh	66891	296	1924582
Assam	746104	8035	50335764
Bihar	851478	12303	157282765
Chandigarh	99482	1183	2290660
Chhattisgarh	1177891	14147	49167582
Dadra & Nagar	11591	4	1579855
Delhi	2009896	26528	37405098
Goa	259933	4014	2874432
Gujarat	1282202	11055	128102003
Haryana	1057662	10714	45543318
Himachal Pradesh	314259	4217	15325456
Jammu &Kashmir	479708	4785	24782117
Jharkhand	442627	5332	43867143

Karnataka	4077389	40324	122147278
Kerala	6836683	71662	57507197
Ladakh	29446	231	567133
Lakshadweep	11415	52	145280
Maharashtra	8144111	148441	177955713
Manipur	139924	2149	3269104
Meghalaya	96791	1625	2625667
Mizoram	238969	726	1793298
Madhya Pradesh	1055102	10777	133937251
Nagaland	35988	782	1739798
Odisha	1336957	9205	81545463
Puducherry	176059	1976	2274059
Punjab	784977	19291	47048380
Rajasthan	1316083	9661	115717969
Sikkim	44354	500	1360447
Tamil Nadu	3596790	38050	127529632
Telangana	842500	4111	77554632
Tripura	108034	940	5918997
Uttar Pradesh	2128888	23650	392005721
Uttrakhand	449630	7755	20143333
West Bengal	2119019	21533	156098134

Table 2: Machine Learning Approaches and their Correlation Coefficient

Attributes	**M5P**	**Random Forest**
Total Case	1.0000	0.9806
Death	0.8577	0.8038
Total Dose	0.9985	0.8222

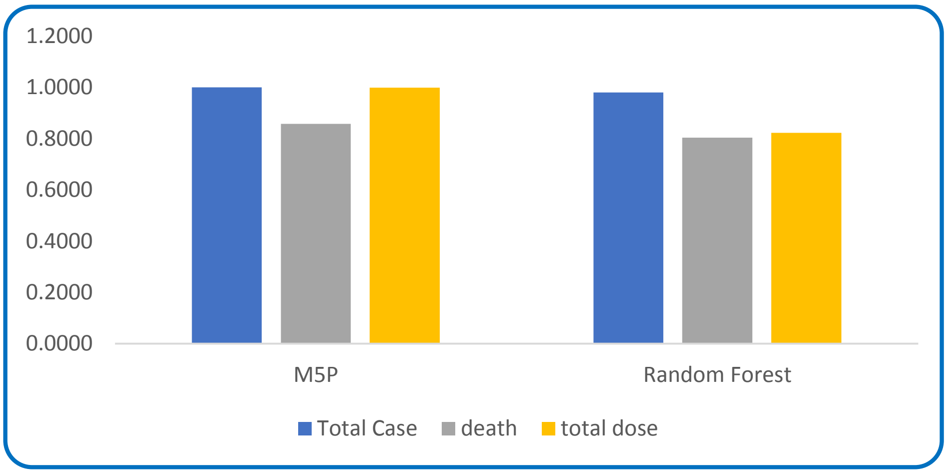

Fig. 1. Correlation Coefficient using M5P and RF

Table 3: Machine Learning Approaches and their RAE

Attributes	M5P	Random Forest
Total Case	0.4007	19.4956
Death	43.0426	37.0647
Total Dose	4.7026	27.9812

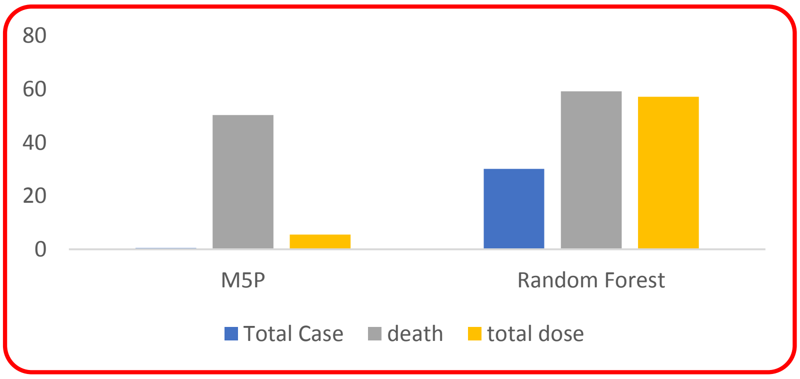

Fig. 2. Machine Learning Approaches and their RAE

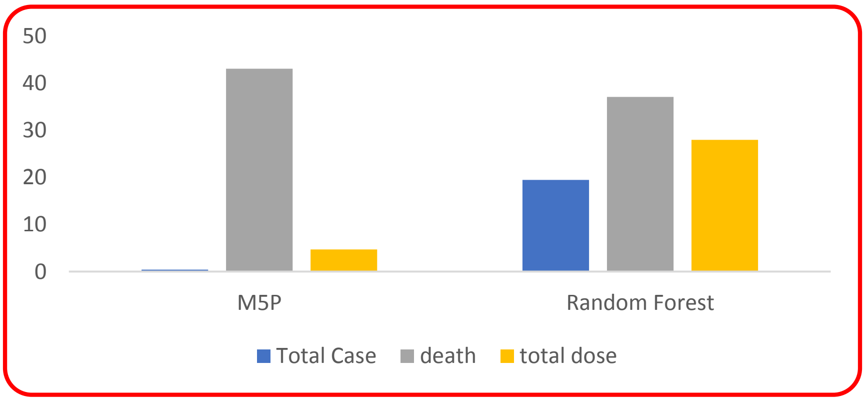

Fig. 3: Machine Approaches and their RRSE

Table 4: Machine Approaches and their RRSE

Attributes	M5P	Random Forest
Total Case	0.4500	30.1442
Death	50.3380	59.2760
Total Dose	5.5458	57.2101

2. Results and Discussion

Table 1 explains four parameters, namely the name of the state/UTS, total cases, death, and total dose and their simple definition. Table 2 indicates different machine learning approaches, namely M5P and Random Forest using correlation coefficient or R2 score. In this case, numerical illustrations explain total cases and their combinations return strong positive correlations (M5P = 1 and RF = 0.9806) for using both ML algorithms, and the remaining two different parameters also return strong positive correlations. Based on the R2 score, the three parameters' future prediction results also have a strong positive correlation. The results and discussions are shown in Table 2 and Figure 1. The R2 score and Relative Absolute Error (RAE) measure accuracy using equation 1 and 2 to compare combinations between predicted and actual values. In this case, the parameter total cases return a minimum error (0.4007), and the total dose returns a minimum error (4.7026) for using M5P. Numerical illustrations are shown in Table 3 and Figure 2. Root Relative Squared Error (RRSE) provides a relative measure of the models with performance, and the result is expressed as a percentage using equation 3. RRSE is used to find the ML algorithms and their accuracy. In this case, the entire case and total dose return a minimum squared error. Based on the RRSE results, the combination of two parameters, namely total case and total dose, is the best independent and dependent variable for future predictions. The similarity study is shown in Table 4 and Figure 3.

3. Conclusion and Future Research

It is essential to consider the limitations of this study. The sample size of each group was relatively small, which could impact the generalizability of the results. The covid analysis and prediction of various parameters influenced, but the total case and total dose is the best parameter for predicting the future based on various illustrations. In the future, add more parameters in the primary table and include different machine learning approaches to increase better performance and increase accuracy or test statistics.

4. Reference

1. Balli, S., 2021. Data analysis of the Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals, 142, p.110512.
2. Abdul Kareem, N.M., Abdulazeez, A.M., Zeebaree, D.Q. and Hasan, D.A., 2021. COVID-19 world vaccination progress using machine learning classification algorithms. Qubahan Academic Journal, 1(2), pp.100-105.
3. Hossen, M.S. and Karmoker, D., 2020, December. Predicting the Probability of Covid-19 Recovered in South Asian Countries Based on Healthy Diet Pattern Using a Machine Learning Approach. In 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1-6). IEEE.
4. Kalezhi, J., Chibuluma, M., Chembe, C., Chama, V., Lungo, F. and Kunda, D., 2022. I am modeling Covid-19 infections in Zambia using data mining techniques. Results in Engineering, 13, p.100363.
5. Almansoor, M. and Hewahi, N.M., 2020, October. Exploring the Relation between Blood Tests and Covid-19 Using Machine Learning. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-6). IEEE.
6. Sun, N.N., Yang, Y., Tang, L.L., Dai, Y.N., Gao, H.N., Pan, H.Y. and Ju, B., 2020. A prediction model based on machine learning for diagnosing early COVID-19 patients. MedRxiv.
7. Kivrak, M., Guldogan, E. and 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, p.105951.
8. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the WEKA tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
9. Rajesh, P. and Santhosh Kumar, B., 2020. Comparative studies on Sustainable Development Goals (SDG) in India using Data Mining approach, Scientific Transactions in Environment and Technovation, 14(2), pp. 91-93.
10. Wang, Y., Witten, I. H.: Induction of model trees for predicting continuous classes. In: Poster papers of the 9th European Conference on Machine Learning, 1997.
11. Ross J. Quinlan: Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348, 1992.
12. Santhosh Kumar, B., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp. 1472–1477.
13. Breiman, L. (2001). Random forests. Machine learning, 45(1), pp. 5-32.
14. https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/
15. https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70