International Journal of Psychosocial Rehabilitation, Vol. 27, Issue 03, 2023
ISSN: 1475-7192

Analysis of COVID-19 in World Dataset Using Machine Learning Models

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

Abstract

COVID-19 had a global impact, affecting countries and regions to varying extents. While I can offer general information up to that date, it's essential to bear in mind that the situation is continually changing. For the most current and trustworthy updates, please consult authoritative sources. Machine learning, a branch of artificial intelligence, harnesses statistical methods to empower computers to acquire knowledge and render decisions without explicit programming. It operates on the principle that computers can gain insights from data, identify patterns, and exercise judgment with minimal human intervention. This paper considers covid -19 in world-related dataset like country, continent, total_confirmed, total_deaths, total_recovered, active_cases, serious_or_critical, total_cases_per_1m_population, total_deaths_per_1m_population, total_tests, total_tests_per_1m_population, population. The machine learning approaches which is used to analysis and predict the dataset using linear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: Machine learning, covid -19 in world, decision tree, correlation coefficient, and test statistics.

1. Introduction and Literature Review

COVID-19, it's essential to recognize that the situation is ever-changing. To obtain the most current and accurate information, it's imperative to seek guidance from trustworthy sources, such as the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), and your local health authorities.

Data mining classification is the practice of assigning data into predefined categories or classes by analyzing attributes and patterns. It's a widely-used method in data analysis and machine learning, where algorithms determine the category of new data points based on their characteristics. This process finds applications in tasks like identifying spam emails, performing sentiment analysis, making medical diagnoses, and various other scenarios that require structuring and labeling data into meaningful groups. Supervised machine learning encompasses a wide range of applications, including but not limited to spam email detection, image classification, language translation, and medical diagnosis. This approach is recognized for its efficacy in tasks where the desired outcome is known and serves as the foundation for training the algorithm.

A machine learning technique, known as EAMA, has been proposed for long-term forecasting of COVID-19 related parameters in India and globally. This hybrid model, EAMA, is well-suited for making predictions based on historical and current data. The study leveraged datasets from the Ministry of Health & Family Welfare of India and Worldometers to outline long-term predictions for India and the world. The results showed close alignment between predicted and real-time data. The study also included statewise predictions for India and countrywise predictions worldwide [1].

Humans contract coronaviruses in three ways: mild respiratory disease, zoonotic infections like MERS-CoV, and severe cases like SARS-CoV. Machine learning techniques are employed to classify these three COVID-19 stages by extracting features from data. The TF/IDF method is used for statistical analysis in text data mining of COVID-19 patient records for classification and prediction. This study demonstrates the feasibility of using blood tests and machine learning as an alternative to rRT-PCR for diagnosing COVID-19 patient categories [2].

Supervised machine learning models for COVID-19 infection using algorithms like logistic regression, decision trees, support vector machines, naive Bayes, and artificial neural networks. It used labeled datasets of positive and negative COVID-19 cases in Mexico. The models were trained and tested, and the results indicated high accuracy, sensitivity, and specificity for various models, with decision trees showing the highest accuracy [3].

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

DOI: 10.61841/V27I3/400307

Large-scale data of COVID-19 patients can be harnessed with advanced machine learning algorithms to understand the viral spread, improve diagnosis, develop therapies, and identify susceptible individuals. Advanced machine learning techniques have been applied in various areas such as taxonomic classification of COVID-19 genomes, CRISPR-based COVID-19 detection assays, survival prediction for severe COVID-19 patients, and potential drug discovery against COVID-19 [4].

Data for COVID-19 cases in the USA, Germany, and globally between 20/01/2020 and 18/09/2020 were obtained from the World Health Organization. Machine learning models, including linear regression, multi-layer perceptron, random forest, and support vector machines, were used for time series predictions. SVM achieved the best results, indicating the global pandemic peak at the end of January 2021 with an estimated 80 million cumulative infections [5].

Data mining is a valuable tool for uncovering hidden insights in large databases. This paper focuses on a weather dataset, where various classification algorithms like J48, Random Tree, Decision Stump, Logistic Model Tree, Hoeffding Tree, Reduce Error Pruning, and Random Forest were used to determine whether conditions are conducive for playing golf. The Random Tree algorithm demonstrated the highest accuracy at 85.714% [6].

Google Trends data was used to estimate the number of positive COVID-19 cases in Iran. Linear regression and LSTM models were employed, and their performance was evaluated using 10-fold cross-validation with RMSE as the performance metric [7].

Machine learning classification algorithms were employed to analyze the progress of COVID-19 vaccination worldwide. Decision Tree, K-nearest neighbors, Random Tree, and Naive Bayes algorithms were compared based on accuracy and performance, with Decision Tree emerging as the most accurate and time-efficient choice [8].

Various machine learning algorithms, including Random Forest, Support Vector Machine, and K-Nearest Neighbor, were used to predict the recovery rate of COVID-19 patients in South Asian countries based on their dietary patterns [9].

A study on groundwater level, rainfall, population, food grains, and enterprises data used stochastic modeling and data mining to predict groundwater levels efficiently. Data assimilation analysis was introduced for effective groundwater level prediction [10][11].

A chronic disease dataset was used for classification, employing five different decision tree algorithms. The M5P decision tree approach was found to be the best-performing algorithm for building models compared to other decision tree approaches [12].

2. Backgrounds and Methodologies

A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that decision. The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

2.1 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition.

The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$\mathbf{y = m_x + b} \qquad \ldots (1)$$

Where:

- y is the dependent variable (the one you want to predict or explain).
- x is the independent variable (the one you're using to make predictions or explanations).
- m is the slope of the line, representing how much
- y changes for a unit change in x.

b is the y-intercept, indicating the value of y when x is 0.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. **Input Layer:** This layer consists of neurons receiving input data. Each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network.
ii. **Hidden Layers:** These layers come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output.
iii. **Output Layer:** This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

2.3 SMO

SMO stands for "Sequential Minimal Optimization," an algorithm used for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly well-suited for solving the quadratic programming optimization problem that arises during the training of SVMs.

Step 1. **Initialization:** Start with all the data points as potential support vectors and initialize the weights and bias of the SVM.
Step 2. **Selection of Two Lagrange Multipliers:** In each iteration, the SMO algorithm selects two Lagrange multipliers (associated with the support vectors) to optimize.
Step 3. **Optimize the Pair of Lagrange Multipliers:** Fix all the Lagrange multipliers except the selected two, and then optimize the pair chosen to satisfy certain constraints while maximizing a specific objective function.
Step 4. **Update the Model:** After optimizing the selected pair of Lagrange multipliers, update the SVM model's weights and bias based on the new values of the Lagrange multipliers.
Step 5. **Convergence Checking:** Check for convergence criteria to determine whether the algorithm should terminate.
Step 6. **Repeat:** If convergence hasn't been reached, repeat steps 2 to 5 until it is.

2.4 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition. The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions. The following steps describe what Random Forest works like.

❖ Bootstrap Aggregating (Bagging)
❖ Decision Tree Construction
❖ Voting for Classification, Averaging for Regression

The key advantages of Random Forest are:

❖ Reduced overfitting
❖ Robustness
❖ Feature Importance

Steps involved in Random Forest

Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.5 Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node. The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests. The critical characteristics of Random Trees are as follows:

- Random Feature Subset
- Random Threshold Selection
- No Pruning
- Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:
Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.6 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using a combination of incremental pruning and error-reduction techniques. The key steps involved in building a REP Tree are as follows:

- Recursive Binary Splitting
- Pruning
- Repeated Pruning and Error Reduction

Steps involved in REP Tree

REP Tree (Repeated Incremental Pruning to Produce an Error Reduction Tree) is a machine learning algorithm for classification and regression tasks. It is an extension of decision trees that incorporates pruning to reduce overfitting and improve the model's generalization performance. Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.7 Accuracy Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [14]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [15] and [16].

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \qquad \dots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \qquad ... (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2} \qquad ... (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$RAE = \frac{\Sigma|y_i - \hat{y}_i|}{\Sigma|y_i - \bar{y}|} \qquad \dots (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$\mathrm{RRSE} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}} \qquad \dots (6)$$

Equation 3 to 7 are used to find the model accuracy, which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

3. Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The covid-19 in world dataset include 12 parameters which have different categories of data like country, continent, total_confirmed, total_deaths, total_recovered, active_cases, serious_or_critical, total_cases_per_1m_population, total_deaths_per_1m_population, total_tests, total_tests_per_1m_population, population [17]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. covid -19 in world sample dataset

Country	continent	total_confirmed	total_deaths	total_recovered	active_cases	serious_or_critical	total_cases_per_1m_population	total_deaths_per_1m_population	total_tests	Total_tests_per_1m_population	Population
Afghanistan	Asia	179267	7690	162202	9375	1124	4420	190	951337	23455	40560636
Albania	Europe	275574	3497	271826	251	2	95954	1218	1817530	632857	2871945
Algeria	Africa	265816	6875	178371	80570	6	5865	152	230861	5093	45325517
Andorra	Europe	42156	153	41021	982	14	543983	1974	249838	3223924	77495
Anguilla	North America	2984	9	2916	59	4	195646	590	51382	3368870	15252

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	1.0000	0.9220	1.0000
Multilayer Perceptron	0.9641	0.9410	0.9580
SMOreg	0.9997	0.9105	0.9996
Random Forest	0.8551	0.8293	0.8569

Random Tree	0.7612	0.7835	0.6572
REP Tree	0.7878	0.4588	0.7943

Table 3: Machine Learning Models with Mean Absolute Error

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	0.0012	0.9220	0.0013
Multilayer Perceptron	560756.4229	19498.7619	557213.3999
SMOreg	107678.2424	22136.7724	129116.4153
Random Forest	1206995.7373	25098.9239	1189642.2356
Random Tree	1690399.9239	26308.4810	1755371.8068
REP Tree	2161586.5900	53567.4588	2082212.7229

Table 4: Machine Learning Models with Root Mean Squared Error

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	0.0025	49410.5173	0.0028
Multilayer Perceptron	3097540.5869	42363.9789	3203551.9731
SMOreg	247878.9613	53796.3872	286370.5736
Random Forest	5629232.3025	74158.0746	5463320.9041
Random Tree	6293348.6495	77860.4337	7044809.4084
REP Tree	6293493.1536	127296.1684	6067430.9073

Table 5: Machine Learning Models with Relative Absolute Error (%)

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	0.0000	34.9363	0.0000
Multilayer Perceptron	11.9216	32.7907	12.3031
SMOreg	2.2892	37.2270	2.8508
Random Forest	25.6605	42.2084	26.2669
Random Tree	35.9375	44.2425	38.7580

REP Tree	45.9548	90.0835	45.9745

Table 6: Machine Learning Models with Root Relative Squared Error (%)

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	0.0000	39.0643	0.0000
Multilayer Perceptron	31.9084	33.4932	34.1356
SMOreg	2.5534	42.5317	3.0514
Random Forest	57.9878	58.6298	58.2146
Random Tree	64.8290	61.5569	75.0663
REP Tree	64.8305	100.6411	64.6518

Table 7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	total_confirmed	total_deaths	total_recovered
Linear Regression	0.1800	0.0100	0.0100
Multilayer Perceptron	0.3300	0.1300	0.1100
SMOreg	0.2200	0.0100	0.0200
Random Forest	0.2400	0.0900	0.0500
Random Tree	0.0100	0.0100	0.0100
REP Tree	0.0200	0.0200	0.0100

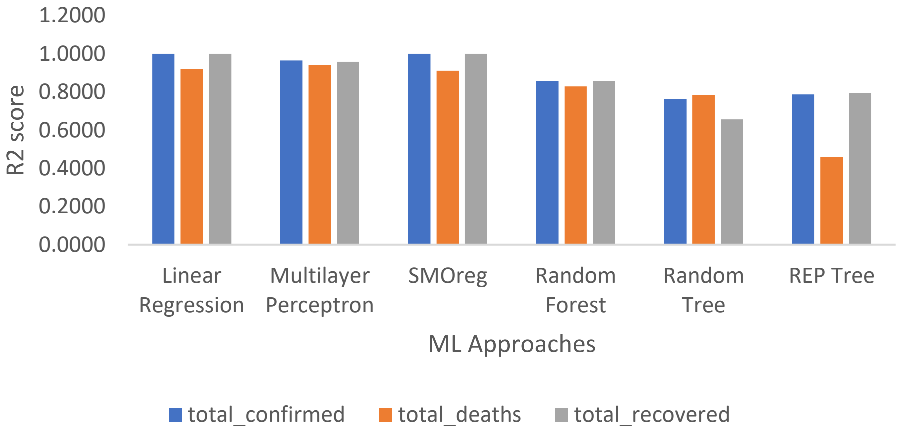

Fig. 1. R2 Score for Machine Learning Approaches

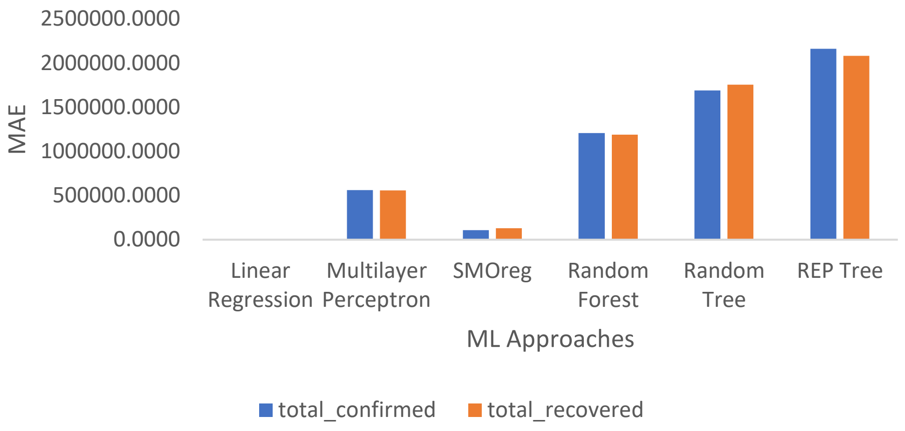

Fig. 2(a). Machine Learning Models with MAE

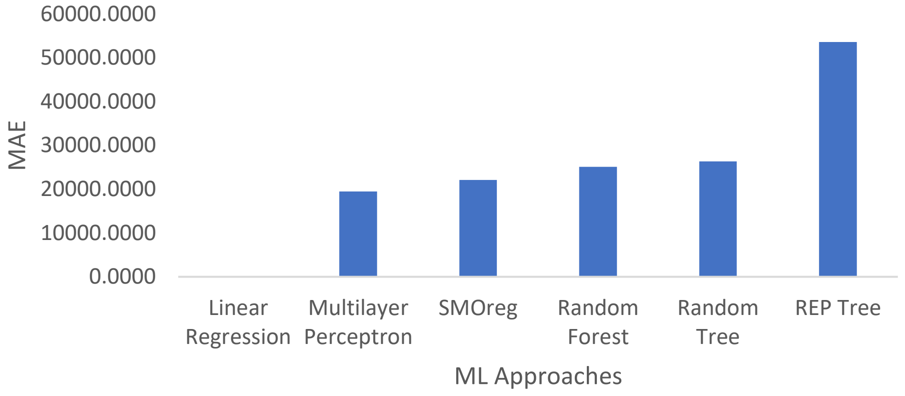

Fig. 2(b). Machine Learning Models with MAE

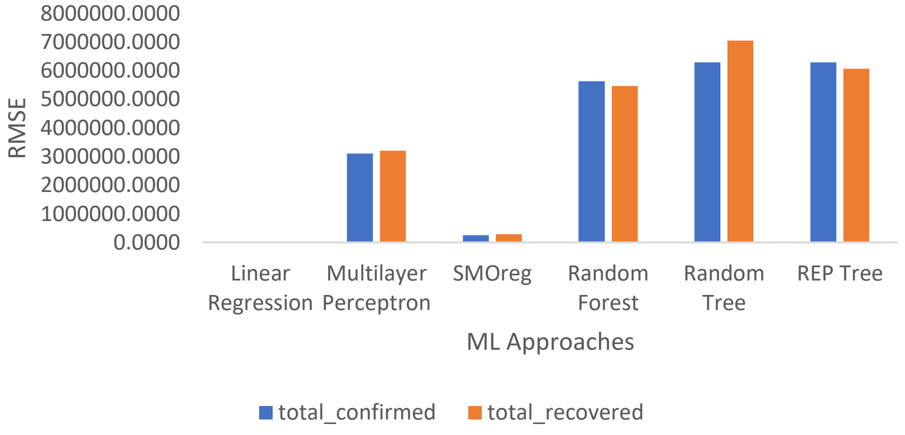

Fig. 3(a). Machine Learning Models with RMSE

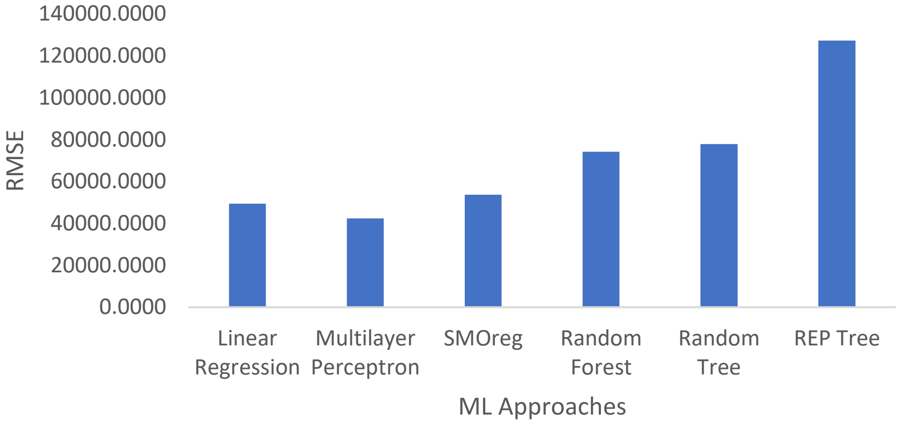

Fig. 3(b). Machine Learning Models with RMSE

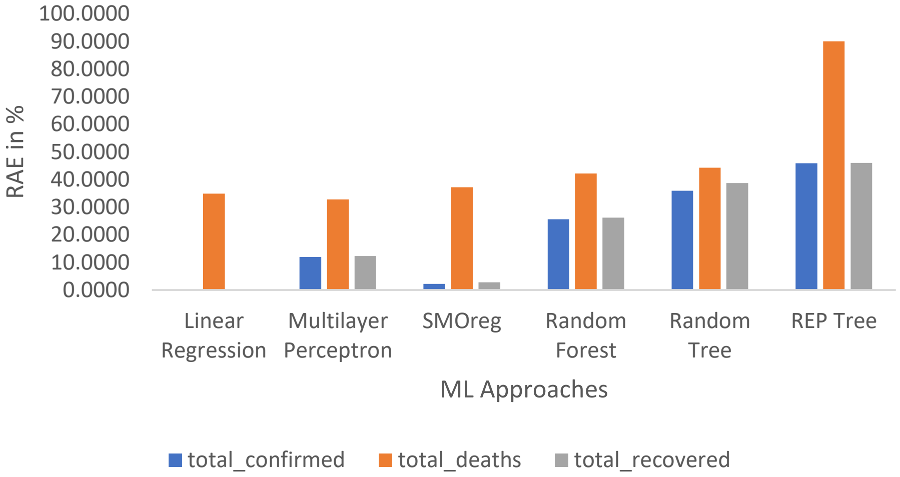

Fig. 4. Machine Learning Models with RAE (%)

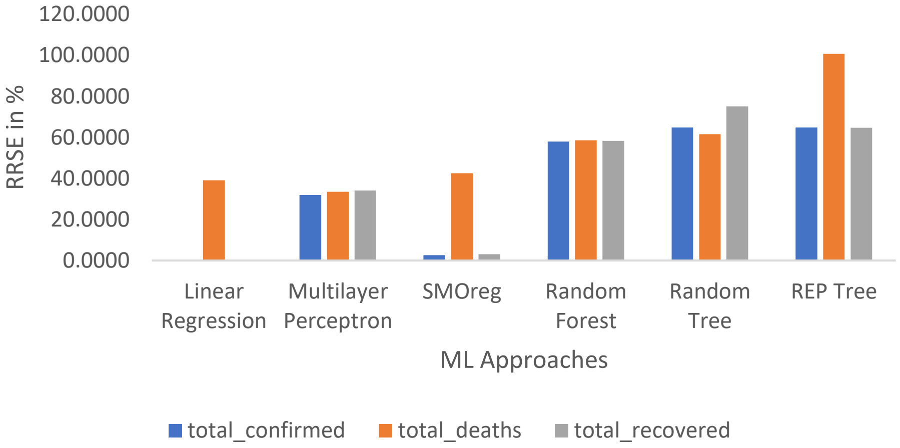

Fig. 5. Machine Learning Models with RRSE (%)

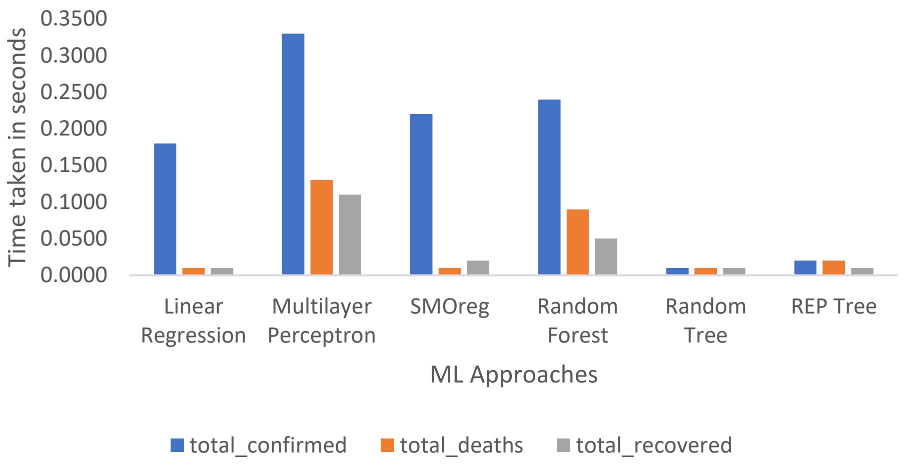

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

4. Results and Discussion

Within Table 1, we delineate 12 parameters encompassing various data categories including country, continent, total_confirmed, total_deaths, total_recovered, active_cases, serious_or_critical, total_cases_per_1m_population, total_deaths_per_1m_population, total_tests, total_tests_per_1m_population, and population. These parameters are subjected to a comprehensive analysis using six machine learning approaches: linear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree. The aim is to unearth concealed patterns and determine the most influential parameter for future predictions. Detailed results and numerical representations can be found in Tables 1 to 7 and Figures 1 to 6.

These analyses are rooted in Equation 2, Table 2, and Figure 1, which are employed to compute the R2 score, facilitating the comparison of the 12 parameters. The numerical data suggests significant variations among these parameters. To address this, a more in-depth analysis involving the combination of three parameters with six distinct ML approaches is conducted. Among these three parameters, namely total_confirmed, total_deaths, and total_recovered, only total_confirmed and total_deaths exhibit a robust positive correlation. Further details are provided in Table 2 and Figure 1.

This research employs six machine-learning algorithms to assess model performance. Mean Absolute Error (MAE) is used to quantify model errors via Equation 3, enabling the determination of the most suitable variable for future predictions. Linear Regression approaches exhibit the lowest error performance, approximately 0.0012, when utilizing only two parameters: total_confirmed and total_deaths. Visual representations are accessible in Table 3 and Figure 2.

Additionally, the Root Mean Square Error (RMSE) is employed to measure the disparity between predicted and actual values, as outlined in Equation 4. In this context, Linear Regression approaches again demonstrate minimal error performance, around 0.0025, when employing the two parameters, total_confirmed and total_recovered. Visual representations can be found in Table 4 and Figure 3.

To assess accuracy, the Relative Absolute Error (RAE), calculated using Equation 5, enables a comparison of predicted and actual values in percentage terms. Linear Regression approaches exhibit minimal error performance, nearly 0, when using only the two parameters, total_confirmed and total_recovered. Visual representations are available in Table 5 and Figure 4. Similar error assessments are presented through RRSE, employing Equation 6, with corresponding numerical illustrations presented in Table 6 and Figure 5. Time efficiency is a crucial factor in machine learning approaches, and Table 7 and Figure 6 showcase the minimal errors associated with the six ML approaches when constructing the model.

5. Conclusion and Future Research

This research unequivocally establishes that the parameters total_confirmed and total_recovered are well-suited for future predictions. Furthermore, we propose potential enhancements and future directions, including the exploration of additional data sources, investigation of superior algorithms and hyperparameters, and fine-tuning of the model to enhance its performance.

6. Reference

1. Mohan, S., Abugabah, A., Kumar Singh, S., Kashif Bashir, A. and Sanzogni, L., 2022. An approach to forecast impact of Covid-19 using supervised machine learning model. Software: Practice and Experience, 52(4), pp.824-840.

2. Ramanathan, S. and Ramasundaram, M., 2021. Accurate computation: COVID-19 rRT-PCR positive test dataset using stages classification through textual big data mining with machine learning. The Journal of supercomputing, 77(7), pp.7074-7088.
3. Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C. and Mohammed, I.A., 2021. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN computer science, 2(1), pp.1-13.
4. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B. and Cheng, X., 2020. Artificial intelligence and machine learning to fight COVID-19. Physiological genomics, 52(4), pp.200-202.
5. Ballı, S., 2021. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals, 142, p.110512.
6. 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.
7. Ayyoub Zadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M. and Kalhori, S.R.N., 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR public health and surveillance, 6(2), p.e18828.
8. 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.
9. 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.
10. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (Vol. 2177, No. 1). AIP Publishing.
11. Rajesh, P. and Karthikeyan, M., 2019. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), pp.18-23.
12. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 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. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).
14. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
15. S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, “Root mean square error (RMSE): A comprehensive review,” International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.
16. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
17. https://www.kaggle.com/code/jeonghyunjhkim/covid-19-spread-and-vaccination-progress-eda/input?select=worldometer_coronavirus_summary_data.csv