A Study on Variable Selections and Prediction for Covid-19 and Delta Dataset Using Machine Learning Approaches

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

[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.

Abstract

Ongoing initiatives to combat COVID-19 prioritize widespread vaccination, ongoing research, and implementing public health measures to mitigate disease transmission and its impact. Staying informed about the most current guidance provided by local health authorities and the World Health Organization (WHO) is crucial for safeguarding your well-being and that of your community. Data mining entails exploring patterns, trends, relationships, and valuable insights within extensive datasets by employing various techniques and algorithms. This process aims to extract useful information from structured and unstructured data sources. This paper considers COVID-19-related dataset like state name, state code, district, confirmed, Active, deceased, recovered, delta confirmed, delta deceased, delta recovered. The machine learning approaches are used to analyze 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, decision tree, correlation coefficient, and test statistics.

1. Introduction and Literature Review

Research on COVID-19 has played a crucial role in enhancing our comprehension of the virus, shaping public health strategies, and propelling the creation of vaccines and treatments to address the pandemic. This field remains dynamic and progressive, marked by ongoing studies and new revelations.

Data mining involves sifting through extensive datasets to uncover patterns and relationships, which can be harnessed for data analysis to address business challenges. Employing data mining techniques and tools empowers organizations to forecast future trends and enhance the quality of their decision-making in the business realm. Machine Learning serves as a global platform for research into computational methods for learning. The journal publishes articles that present significant findings related to a diverse set of learning techniques applied to various learning challenges.

A multi model machine learning technique known as EAMA is introduced for the long-term prediction of COVID-19-related parameters in both India and globally. EAMA, a hybrid model, is adept at making predictions based on historical and current data. The study utilizes two datasets sourced from the Ministry of Health & Family Welfare of India and World Meters. These datasets are employed to outline long-term predictions for both India and the world, with the predicted data closely resembling real-time values. Additionally, state-wise predictions for India and country-wise predictions globally are included in the Appendix [1].

Coronaviruses infect humans through three distinct pathways: mild respiratory disease, zoonotic infections like MERS-CoV, and highly fatal cases, exemplified by SARS-CoV. This study employs machine learning techniques to classify these three stages of COVID-19 using feature extraction from data retrieval processes. Text data mining, utilizing TF/IDF, is applied for statistically evaluating COVID-19 patient records to classify and predict coronavirus cases. The study demonstrates the potential of using blood tests and machine learning as an alternative to rRT-PCR for categorizing COVID-19-positive patients [2].

Author explain supervised machine learning models for COVID-19 infection, employing learning algorithms such as logistic regression, decision trees, support vector machines, naive Bayes, and artificial neural networks. These models are trained using epidemiology labeled datasets for positive and negative COVID-19 cases in Mexico. Correlation coefficient analysis is conducted to assess relationships between dependent and independent features within the dataset. The study finds that the decision tree model achieves the highest accuracy at 94.99%, the Support Vector Machine Model attains the highest sensitivity at 93.34%, and the Naïve Bayes Model records the highest specificity at 94.30% [3].

Large-scale data on COVID-19 patients can be harnessed and analyzed using advanced machine learning algorithms to gain deeper insights into the virus's spread patterns, improve diagnostic accuracy and speed, develop innovative therapeutic strategies, and identify individuals at higher risk based on personalized genetic and physiological characteristics. Notably, machine learning techniques have been swiftly deployed in tasks such as taxonomic classification of COVID-19 genomes, CRISPR-based COVID-19 detection, survival prediction for severe COVID-19 cases, and the identification of potential drug candidates against COVID-19 since the outbreak of the pandemic [4].

Utilizes COVID-19 data for the USA, Germany, and the global scenario between 20/01/2020 and 18/09/2020, sourced from the World Health Organization. The datasets encompass weekly confirmed cases and cumulative confirmed cases for 35 weeks. The data's distribution is analyzed, and parameters are determined based on statistical distributions. The study proposes a time series prediction model using machine learning, employing linear regression, multi-layer perceptrons, random forest, and support vector machines. Among these, the support vector machine achieves the most accurate predictions. The study estimates that the global pandemic will peak at the end of January 2021, with approximately 80 million cumulative infections [5].

Predicting the incidence of COVID-19 in Iran. Google Trends data is employed, and linear regression and long short-term memory (LSTM) models are used to estimate the number of positive COVID-19 cases. All models are evaluated through 10-fold cross-validation, with root mean square error (RMSE) serving as the performance metric [6].

Investigates COVID-19 vaccination progress worldwide using machine learning classification algorithms. Real-world data is analyzed using Weka, and the study employs Decision Tree, K-nearest neighbors, Random Tree, and Naive Bayes algorithms to draw conclusions based on accuracy and performance period. The Decision Tree algorithm is found to outperform others in terms of both time and accuracy [7]. Predict COVID-19 recovery rates in South Asian countries based on healthy diet patterns, utilizing machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) [8].

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 [9].

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:

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

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 [10]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [11] and [12].

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.

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

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

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

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

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

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.

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

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.

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

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-source Kaggle data repository. The COVID-19 dataset includes ten parameters with different data categories like state name, state code, district, confirmed, Active, deceased, recovered, delta confirmed, delta dead, and delta recovered [13]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. Covid-19 sample dataset

state name	state code	District	confirmed	Active	deceased	recovered	Delta confirmed	Delta deceased	Delta recovered
Andhra Pradesh	AP	Foreign Evacuees	111	111	0	0	49	0	0
Andhra Pradesh	AP	Anantapur	136	40	4	92	0	0	0
Andhra Pradesh	AP	Chittoor	208	110	1	97	0	0	0
Andhra Pradesh	AP	East Godavari	59	16	0	43	0	0	0
Andhra Pradesh	AP	Guntur	426	90	8	328	0	0	0

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	1.0000	0.7491	-0.0023	-0.0050
Multilayer Perceptron	0.9513	0.9429	0.0013	-0.0025
SMOreg	1.0000	0.9012	-0.0294	-0.0006
Random Forest	0.8695	0.9481	0.0380	0.0769
Random Tree	0.9263	0.9313	0.0072	0.1953
REP Tree	0.6707	0.9356	0.0570	0.0101

Table 3: Machine Learning Models with Mean Absolute Error

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	0.0027	74.3333	3.6555	0.7981
Multilayer Perceptron	39.9048	33.5953	2.5969	0.5847
SMOreg	0.8658	37.4284	1.6379	0.4357
Random Forest	64.3023	28.6722	2.0775	0.7261
Random Tree	70.8706	32.0450	2.4527	0.7493
REP Tree	163.2373	40.1516	2.3877	0.7406

Table 4: Machine Learning Models with Root Mean Squared Error

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	0.7406	463.2245	24.2028	3.9750
Multilayer Perceptron	507.3871	159.5173	12.7985	4.1584
SMOreg	6.6373	203.4987	12.3972	3.5373
Random Forest	749.1648	158.2081	12.9505	3.7107
Random Tree	682.0731	170.5346	17.4973	4.5567
REP Tree	1087.3562	164.9180	12.5772	3.5478

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

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	0.0010	64.2080	153.6426	107.7675
Multilayer Perceptron	14.5380	29.0191	109.1490	78.9536
SMOreg	0.3154	32.3300	68.8421	58.8375
Random Forest	23.4264	24.7666	87.3164	98.0451
Random Tree	25.8193	27.6800	103.0888	101.1818
REP Tree	59.4700	34.6824	100.3570	100.0047

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

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	0.0025	99.1556	195.9226	112.8038
Multilayer Perceptron	35.7205	34.1455	103.6041	118.0093
SMOreg	0.4673	43.5600	100.3561	100.3834
Random Forest	52.7419	33.8653	104.8350	105.3044
Random Tree	48.0186	36.5038	141.6411	129.3124
REP Tree	76.5509	35.3016	101.8129	100.6805

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

ML Approaches	Confirmed	Recovered	Delta confirmed	Delta recovered
Linear Regression	0.2200	0.0300	0.0100	0.0000
Multilayer Perceptron	0.6800	0.3800	0.3900	0.3600
SMOreg	0.2300	0.1000	0.0700	0.0300
Random Forest	0.6300	0.1300	0.2000	0.1200
Random Tree	0.0000	0.0000	0.0000	0.0000
REP Tree	0.0300	0.0100	0.0100	0.0200

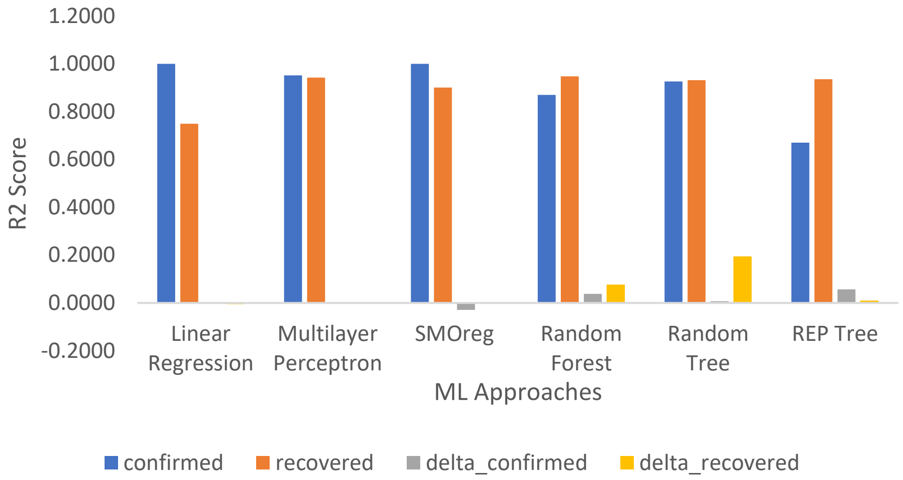

Figure 1. R2 Score for Machine Learning Approaches

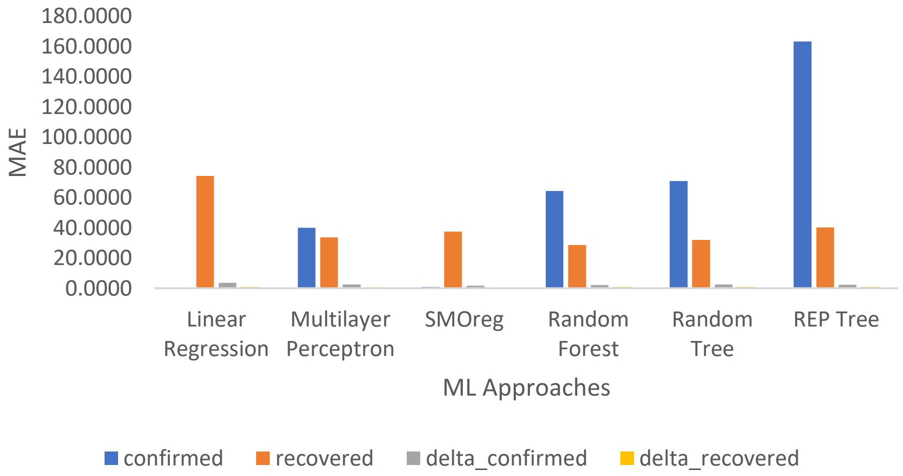

Figure 2. Machine Learning Models with MAE

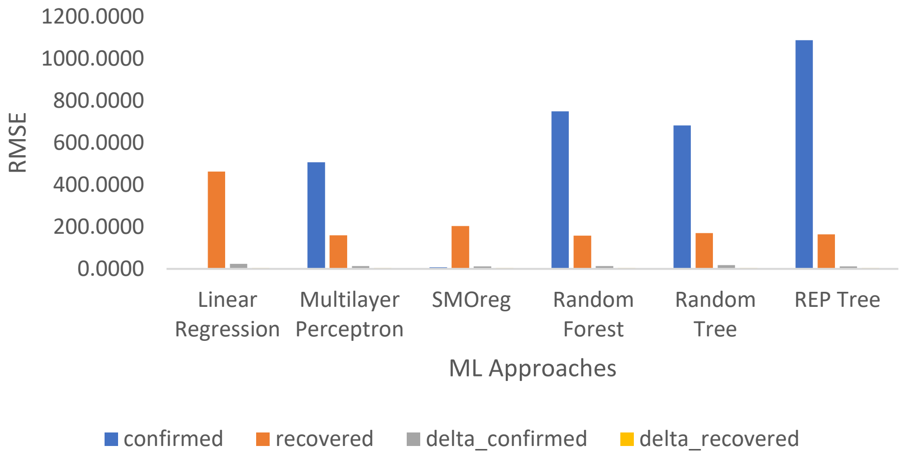

Figure 3. Machine Learning Models with RMSE

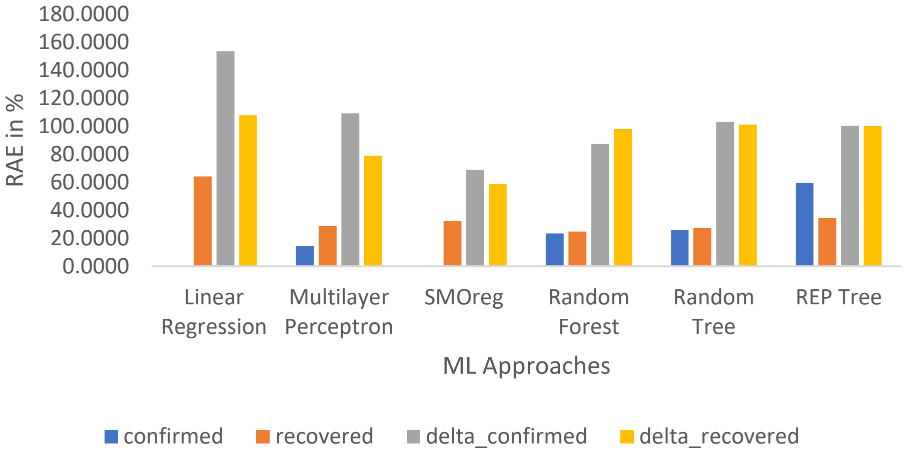

Figure 4. Machine Learning Models with RAE (%)

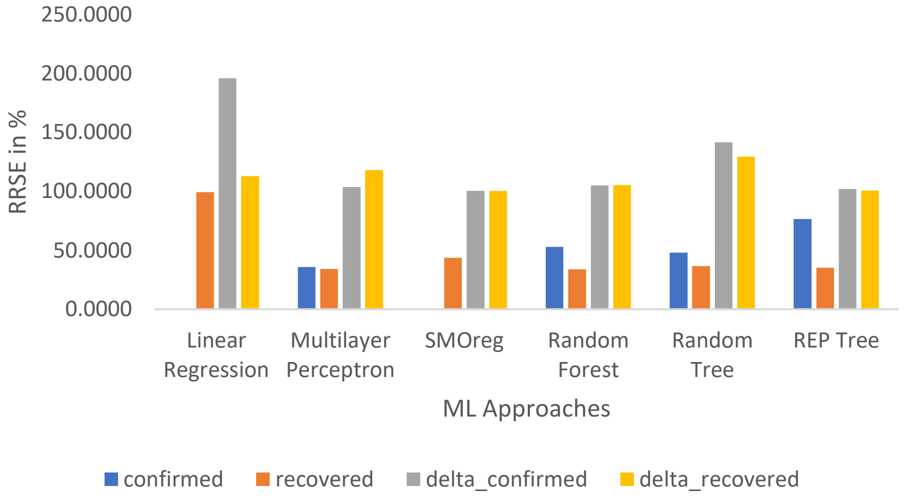

Figure 5. Machine Learning Models with RRSE (%)

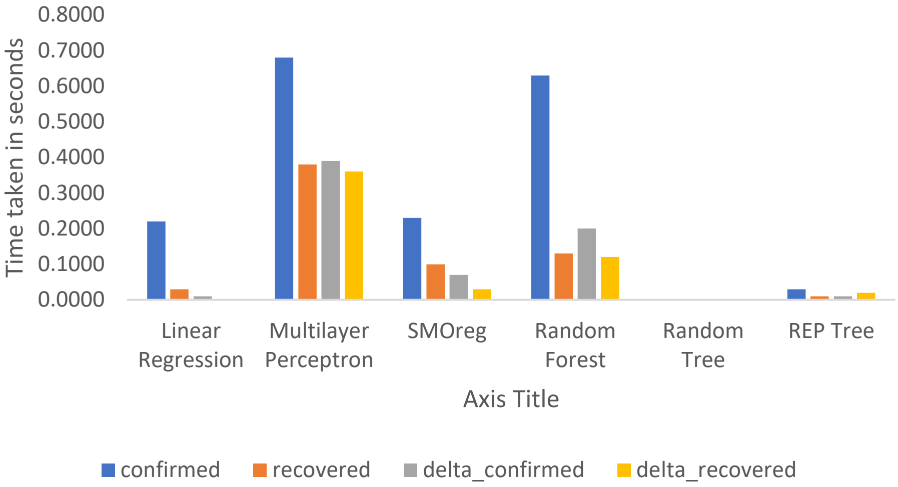

Figure 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

4. Results and Discussion

Table 1 explains ten parameters with diverse data categories, including state name, state code, district, confirmed, active, deceased, recovered, delta confirmed, delta deceased, and delta recovered. The dataset analyzes and unveils hidden patterns through six machine learning approaches: linear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree. The goal is to determine the most influential parameter for future predictions. The results and numerical illustrations are showcased across Tables 1 to 7 and Figures 1 to 6. These analyses are based on Equation 2, Table 2, and Figure 1, which facilitate the calculation of R2 scores by comparing the ten parameters. The numerical illustrations suggest notable differences among these parameters. In this context, a comprehensive analysis is conducted, combining four parameters using six distinct machine learning approaches. Out of these four parameters (confirmed, recovered, delta confirmed, delta recovered), only two (confirmed and recovered) exhibit a strong positive correlation, while the remaining parameters, such as delta confirmed and delta recovered, show negative correlations. The corresponding results and discussions are depicted in Table 2 and Figure 1.

Six machine-learning algorithms are employed in this study, with Mean Absolute Error (MAE) being used to assess model errors, as per Equation 3. The objective is to determine the most suitable variable for future predictions among the six machine learning approaches. The linear regression and SMOreg methods demonstrate minimal error performance, nearly approaching zero when using the confirmed parameter. These findings are visualized in Table 3 and Figure 2.

The Root Mean Square Error (RMSE), as defined in Equation 4, quantifies the disparity between predicted and actual values. Similar to the MAE analysis, linear regression and SMOreg approaches yield minimal error performance, nearly reaching zero for the confirmed parameter. These results are depicted in Table 4 and Figure 3.

Relative Absolute Error (RAE) is employed to gauge accuracy, comparing the disparity between predicted and actual values in percentage, as per Equation 5. The linear regression and SMOreg approaches consistently deliver minimal error performance, approaching zero when employing the confirmed parameter. These insights are presented in Table 5 and Figure 4. Correspondingly, similar error analysis is reflected through Relative Root Mean Square Error (RRSE) using Equation 6, with corresponding numerical representations provided in Table 6 and Figure 5. Notably, time taken is a crucial aspect in machine-learning methodologies, with Table 7 and Figure 6 demonstrating that the six machine learning approaches result in minimal errors in building the model.

5. Conclusion and Future Research

This research unequivocally asserts that the parameters "confirmed" and "recovered" are highly suitable for predicting future outcomes. Additionally, it proposes potential enhancements and future steps, including exploring additional data sources, investigating more robust algorithms and hyperparameters, and fine-tuning the model to improve its performance.

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. 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.
7. 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.
8. 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.
9. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).
10. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
11. 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.
12. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
13. https://www.kaggle.com/code/thejaskiran/covid19-karnataka/input?select=district_level_latest.csv