Machine Learning Models to Analysis and Prediction for Covid Vaccine Dataset Using Classification 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,
Email: nsankarraj@gmail.com

[2]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram - 608 102, Tamil Nadu, India,
Email: us.mani.s.mca@gmail.com

Abstract:

The most up-to-date and precise updates about the distribution and accessibility of COVID-19 vaccines at a state or regional level, it is advisable to contact the official health department or agency of the specific country you are inquiring about. These organizations serve as the primary and most reliable resource for current information on COVID-19 vaccination initiatives and distribution at the state or regional level. Data mining underpins essential business functions like fraud detection, risk management, and cybersecurity planning, among other crucial applications. Additionally, it holds significance in various sectors such as healthcare, government, scientific research, mathematics, sports, and more. This paper considers covid vaccine state wise-related dataset like updated on, state, total doses administered, sessions, sites, first dose administered, second dose administered, covaxin (doses administered), covishield (doses administered), total individuals vaccinated. The machine learning approaches which is used to analysis and predict the dataset usingLinear Regression, Multilayer Perceptron, SMO reg,Decision Stump,M5P,random forest and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: Machine learning, covid vaccine, decision tree, R2 Score, and test statistics.

DOINUMBER:10.48047/NQ.2022.20.19.NQ99450 NEUROQUANTOLOGY2022;20(19):4884-4895

1. INTRODUCTION AND LITERATURE REVIEW

Leveraging data mining and machine learning techniques for COVID-19 vaccine distribution and administration at the state level has the potential to provide valuable insights. This research can contribute to the optimization of vaccine allocation, the anticipation of vaccine demand, and the identification of variables impacting vaccination rates.

Data mining is typically characterized as the utilization of computer technology and automation to sift through extensive datasets, uncovering patterns and trends that can be transformed into valuable business insights and predictive information. Going beyond mere data search, data mining harnesses data to assess future probabilities and generate actionable analyses. Machine learning (ML) entails the systematic application of algorithms and statistical models to proficiently execute a defined task without explicit instructions, leaning on patterns and inference instead. Essentially, ML methods provide a quantitative approach for systematically analyzing qualitative data.

Machine learning techniques to identify optimal predictive models based on their performance. The study focused on data from 435 suspected COVID-19 cases recorded at Imam Khomeini Hospital between May 9, 2020, and December 20, 2020. The Chi-square method was utilized to identify key diagnostic features for COVID-19, and eight data mining algorithms, including multilayer perceptron (MLP), J-48, Bayesian Net (Bayes Net), logistic regression, K-star, random forest, Ada-boost, and sequential minimal optimization (SMO), were employed. The most suitable diagnostic model for COVID-19 was determined through a comparative evaluation of these algorithms [1].

A dataset containing information on 1500 eligible patients (1386 survivors and 144 deaths) at Ayatollah Taleghani Hospital in Abadan, Iran. Multiple machine learning algorithms were trained to predict COVID-19 mortality. To assess the models' performance, metrics derived from the confusion matrix were calculated [2].

The global distribution of COVID-19 incidence introduces an artificial intelligence approach based on a deep convolutional neural network (CNN) to detect COVID-19 patients using real-world datasets, specifically chest X-ray images. The study highlights the value of X-ray analysis in COVID-19 diagnosis, given its rapid availability and cost-effectiveness. Empirical findings from 1000 X-ray images confirmed the effectiveness of our system, achieving an F-measure range of 95–99%. Additionally, three forecasting methods—Prophet Algorithm (PA), Autoregressive Integrated Moving Average (ARIMA) model, and Long Short-Term Memory Neural Network (LSTM)—were used to predict COVID-19 confirmations, recoveries, and deaths over the next 7 days, showing promising performance with high accuracy in Australia and Jordan. The study also identifies coastal areas as being significantly more impacted by COVID-19 spread [3].

The prediction performance of death status based on demographic and clinical factors, including COVID-19 severity, using data mining methods. The dataset comprises 1603 SARS-COV-2 patients and 13 variables obtained from an open-source web address. Machine learning approaches, including deep learning and random forest, were employed, with hyperparameters tuned through a grid search algorithm. The study applies the steps of knowledge discovery in databases to obtain relevant information [4].

Cox_COVID_19 and Deep_Cox_COVID_19, designed for survival analysis of COVID-19 patients. Cox_COVID_19 is based on Cox regression, while Deep_Cox_COVID_19 combines an autoencoder deep neural network with Cox regression to enhance prediction accuracy. The clinical dataset comprises 1085 patients, with results showing that applying an autoencoder before Cox regression improves concordance, accuracy, and precision. Deep_Cox_COVID_19 outperforms Cox_COVID_19 in accuracy and identifies age, muscle pain, pneumonia, and throat pain as the most important features affecting mortality. Both systems can provide valuable information for doctors to reduce mortality [5].

Data mining is a valuable tool for analyzing large databases, and in this study, weather data is used to determine the suitability for playing golf. The dataset includes attributes such as Outlook, Temperature, Humidity, Windy, and a Boolean Play Golf class variable. Seven classification algorithms, including J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP), and Random Forest (RF), were used to measure accuracy. The Random Tree algorithm outperformed other algorithms with an accuracy of 85.714% [6].

Clinical data is encoded into fixed-length feature vectors and subjected to efficient feature selection. The study applies this approach to two COVID-19 patient datasets and uses various machine learning algorithms for classification. The research demonstrates that with efficient feature selection, prediction accuracy exceeds 90% in most cases and highlights the importance of specific attributes for studying the disease [7].

A hybrid machine learning approach, combining adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm

(MLP-ICA), to predict COVID-19 data from Hungary. The models predict a substantial drop in the outbreak and total mortality by late May, with validation confirming their accuracy for nine days. The paper showcases the potential of machine learning in future research [8].

COVID-2019 cases from March 1, 2020, to April 11, 2020, and predicts the number of patients in India. Six regression analysis models, including quadratic, third degree, fourth degree, fifth degree, sixth degree, and exponential polynomial, were used to analyze the COVID-2019 dataset. The sixth degree polynomial regression model yielded the lowest root mean square error and is recommended for forecasting the next 6 days of COVID-2019 data in India. The model can assist Indian doctors and the government in planning for the next 7 days and potentially for longer-term forecasting [9].

Stochastic modeling and data mining approaches to assess groundwater levels, rainfall, population, food grains, and enterprises datasets. The study introduces a novel data assimilation analysis for effective groundwater level prediction. The results indicate the robustness of this approach [10][11].

Chronic disease data with attributes including topics, questions, data values, low confidence limit, and high confidence limit. Five classification algorithms are used to assess and compare the accuracy of the decision tree models. The M5P decision tree approach is identified as the best algorithm for building the model 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:

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

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
Step 2. Selection of Two Lagrange Multipliers
Step 3. Optimize the Pair of Lagrange Multipliers
Step 4. Update the Model
Step 5. Convergence Checking
Step 6. Repeat

2.4 Decision Stump

A decision stump is a straightforward machine-learning model for binary classification tasks. It is the most basic form of a decision tree with a single level or depth of one. In a decision stump, only one feature (attribute) of the data is used to decide, and it splits the data into two subsets based on a threshold value for that feature. The decision stump can be understood as choosing one part, choosing a threshold, assigning classes, and predicting.

Steps involved in decision stump

Step 1. Selecting the feature
Step 2. Choosing the threshold
Step 3. Assigning class labels
Step 4. Making predictions
Step 5. Training and evaluation

2.5 M5P

M5P is a machine learning algorithm used for regression tasks. It is an extension of the decision tree-based model called M5, which Ross Quinlan developed. The M5 algorithm combines decision trees and linear regression to create more accurate and flexible regression models. M5P, specifically, stands for M5 Prime. It enhances the original M5 algorithm to improve its predictive performance. M5P uses a tree-based model to divide the data into subsets based on feature values recursively and then fits linear regression models to each of these subsets. The result is a piecewise linear regression model, where different linear regressions are used for other regions of the input feature space.

Steps involved in the M5P

Step 1. Building the initial decision tree (M5 model): Recursive Binary Splitting and Pruning (optional)
Step 2. Linear Regression Model: Leaf Regression Models and Model Parameters
Step 3. Piecewise Linear Regression: Piecewise Prediction
Step 4. Model Evaluation: Training and Testing.

2.6 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.

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)

2.7 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.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):

2.8 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 followsRecursive Binary, Splitting, Pruning and Repeated Pruning and Error Reduction. Steps involved in REP Tree.
Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.9 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.

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

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

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

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

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

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.

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

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.

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

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.

Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The covid vaccine state wisedataset include 10 parameters whichhavedifferent categories of data likeUpdated On, State, Total Doses Administered, Sessions, Sites, First Dose Administered, Second Dose Administered,Covaxin (Doses Administered), CoviShield (Doses Administered), Total Individuals Vaccinated [17]. Adetailed description of the parameters is mentioned in the following Table 1.

Table 1. Covid vaccine state wise sampledataset

Updated On	State	Total Doses Administered	Sessions	Sites	First Dose Administered	Second Dose Administered	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
16-01-2021	India	48276	3455	2957	48276	0	579	47697	48276
17-01-2021	India	58604	8532	4954	58604	0	635	57969	58604
18-01-2021	India	99449	13611	6583	99449	0	1299	98150	99449
19-01-2021	India	195525	17855	7951	195525	0	3017	192508	195525
20-01-2021	India	251280	25472	10504	251280	0	3946	247334	251280

Table 2: Machine Learning Models with R-squared

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	0.9934	1.0000	1.0000
Multilayer Perceptron	0.9964	0.9999	0.9999
SMOreg	0.9958	1.0000	1.0000
Decision Stump	0.9006	0.9269	0.9250
M5P	0.9981	0.9999	1.0000
Random Forest	0.9993	0.9997	0.9997
REP Tree	0.9975	0.9988	0.9986

Table 3: Machine Learning Models with Mean Absolute Error

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	131132.2532	34769.8182	31955.2884
Multilayer Perceptron	137621.9203	109645.7821	138265.3357
SMOreg	98117.5358	26059.3773	33419.9267
Decision Stump	522021.0132	3805845.7606	3507049.4272
M5P	71067.0586	120094.7516	27895.2694
Random Forest	52597.3627	307738.0301	277473.0669
REP Tree	78792.3954	477474.4085	443379.9195

Table 4: Machine Learning Models with Root Mean Squared Error

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	285465.3245	113170.9304	177058.7001
Multilayer Perceptron	214127.1096	219000.6966	314297.9169
SMOreg	239439.2672	130363.7540	177364.5882
Decision Stump	1085746.2635	7545976.8981	6966856.6054
M5P	154288.0046	245722.5221	177337.8218
Random Forest	96800.1943	467856.3594	451365.1236
REP Tree	175565.2558	974070.4246	979201.4825

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

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	15.9279	0.5298	0.5327
Multilayer Perceptron	16.7161	1.6706	2.3048
SMOreg	11.9178	0.3970	0.5571
Decision Stump	63.4069	57.9861	58.4609
M5P	8.6321	1.8298	0.4650
Random Forest	6.3887	4.6887	4.6253
REP Tree	9.5705	7.2748	7.3909

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

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	11.4282	0.5628	0.9652
Multilayer Perceptron	8.5722	1.0890	1.7133
SMOreg	9.5856	0.6483	0.9669
Decision Stump	43.4662	37.5241	37.9783
M5P	6.1767	1.2219	0.9667
Random Forest	3.8752	2.3265	2.4605

REP Tree	7.0285	4.8438	5.3379

Table7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Covaxin (Doses Administered)	CoviShield (Doses Administered)	Total Individuals Vaccinated
Linear Regression	0.3400	0.0100	0.0100
Multilayer Perceptron	4.3200	3.9500	3.9800
SMOreg	33.9200	24.6400	4.0300
Decision Stump	0.1000	0.0100	0.0100
M5P	0.9600	0.2000	0.1500
Random Forest	3.4400	1.4600	1.3500
REP Tree	0.2900	0.0500	0.0300

Based on data analysis and prediction, the linear regression approach returns different prediction equations. In this case, the three different equations are used to predict the future. The R2 score and other test statistics also return very strong positive correlations as well as minimum error.

Covaxin (Doses Administered) = 0.0605 * Total Doses Administered + 0.1113 * Sessions + (-32.9075) * Sites + 0.0855 * First Dose Administered + 0.1205 * Second Dose Administered +(-0.0504) * Total Individuals Vaccinated + 5514.6911

CoviShield (Doses Administered) = 0.3975 * Total Doses Administered +0.0549 * Sessions + -24.4157 * Sites + 0.5722 * First Dose Administered + 0.5855 * Second Dose Administered + (-1.1324 * Covaxin (Doses Administered) + 0.0558 * Total Individuals Vaccinated + 2563.3328

Total Individuals Vaccinated = 0.3252 * Total Doses Administered + 0.0875 * Sessions + 8.2316 * Sites + 0.5108 * First Dose Administered + -0.3911 * Second Dose Administered +0.1376 * Covaxin (Doses Administered) + 0.1387 * CoviShield (Doses Administered) + 6727.2888

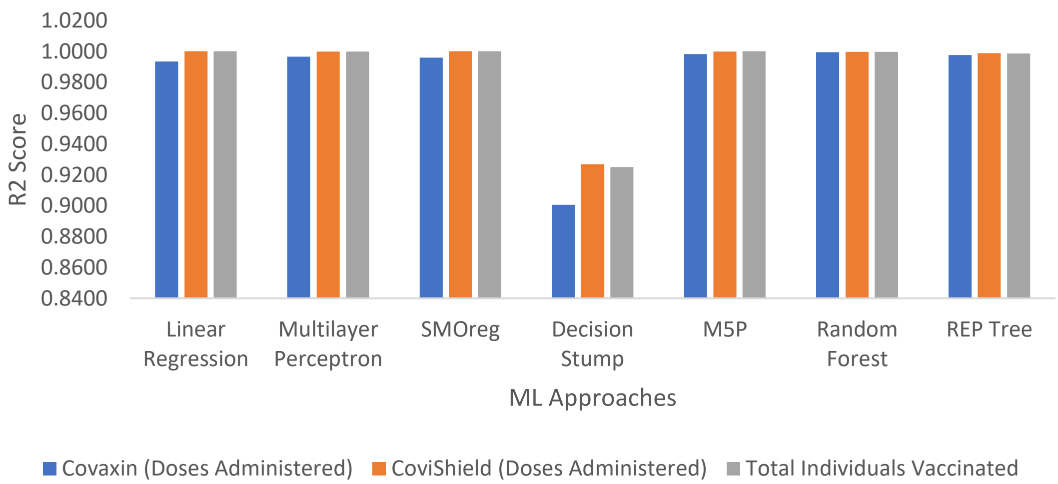

Fig. 1. R-squared for Machine Learning Approaches

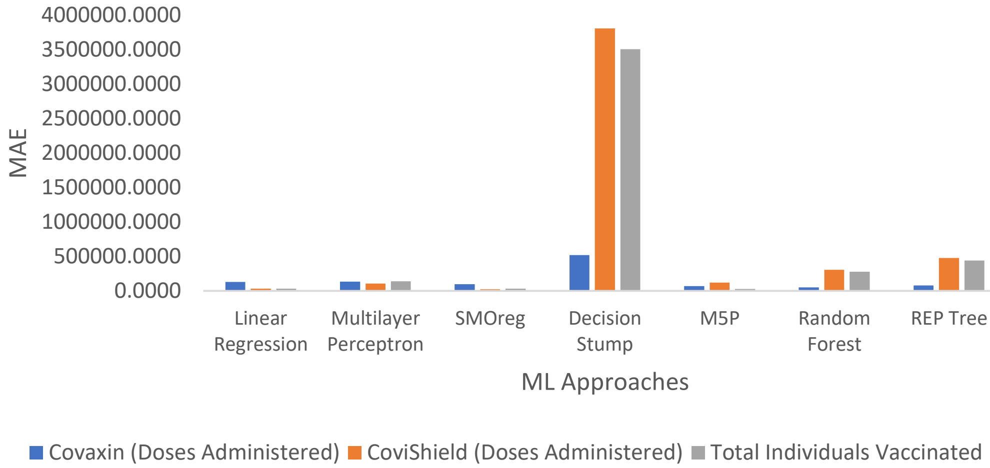

Fig. 2.Machine Learning Models with MAE

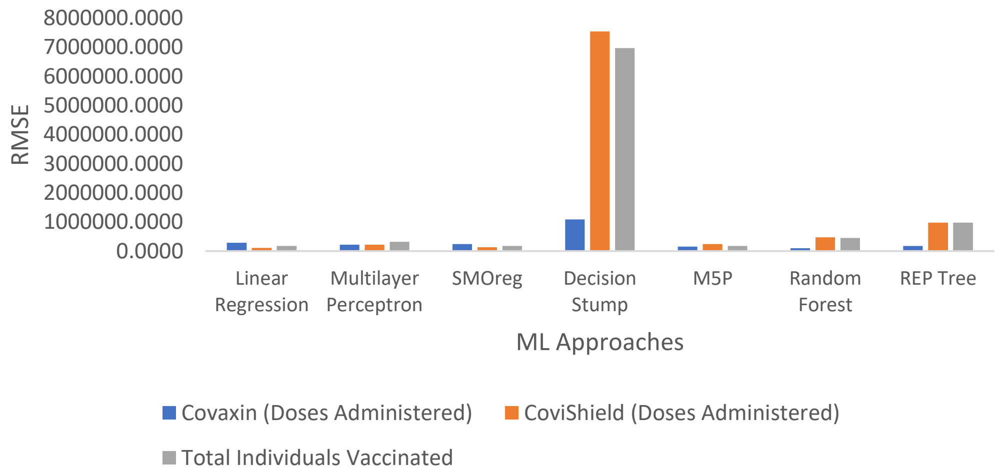

Fig. 3. Machine Learning Models with RMSE

4892

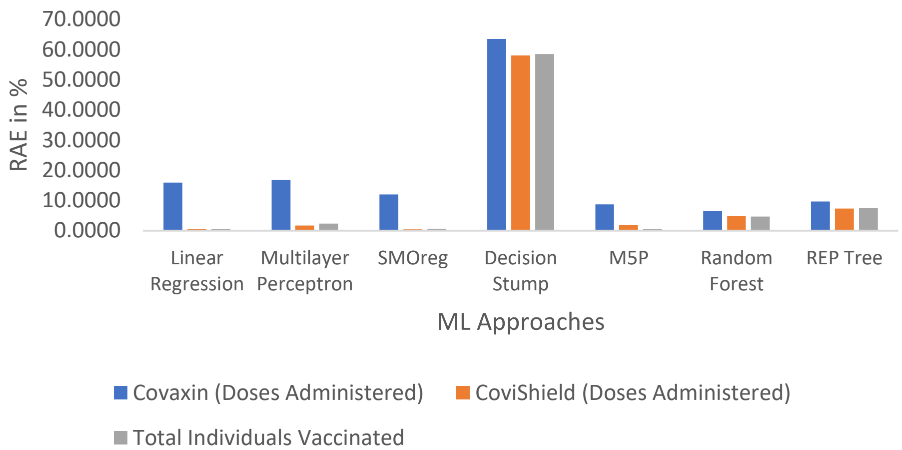

Fig. 4. Machine Learning Models with RAE (%)

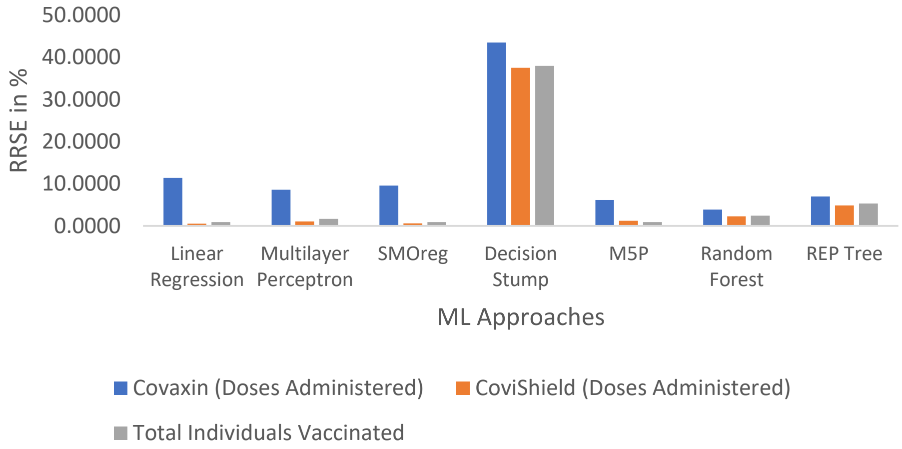

Fig. 5. Machine Learning Models with RRSE (%)

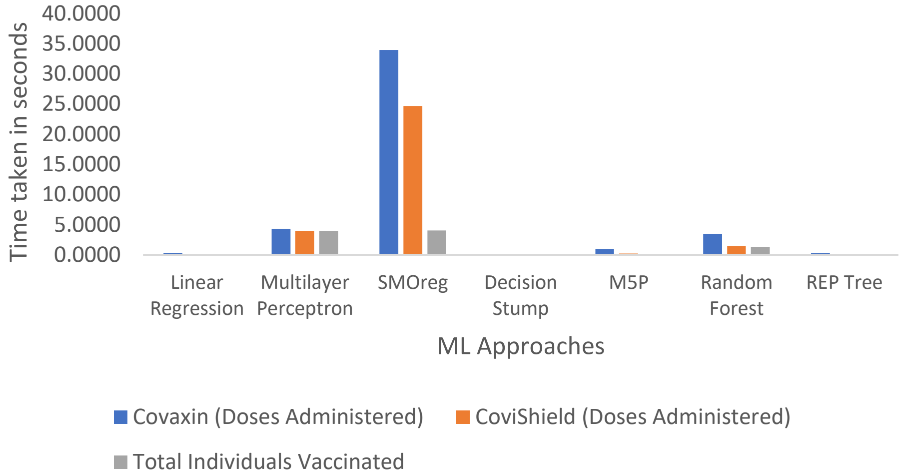

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. RESULTS AND DISCUSSION

Table 1 details ten parameters across various data categories, including timestamps, state information, total doses administered, session data, vaccination site statistics, and the administration of the first and second doses, as well as Covaxin and Covishield doses administered, and the overall count of vaccinated individuals. Our analysis of this dataset employs seven distinct machine learning techniques: Linear Regression, Multilayer Perceptron, SMOreg, Decision Stump, M5P, Random Forest, and REP Tree. The primary objective is to unveil latent patterns and identify the most influential parameter for future predictions. You can find related results and illustrative numerical data between Table 1 and Table 7, as well as Figure 1 through Figure 6. These analyses are underpinned by Equation 2, Table 2, and Figure 1, all of which contribute to the computation of the R2 score based on a comparison of these ten parameters. The numerical illustrations highlight substantial disparities among these parameters. In this context, our analysis explores combinations of three parameters using the seven distinct machine learning approaches. Notably, among these parameters - Covaxin doses administered, Covishield doses administered, and the total number of vaccinated individuals - all three exhibit a remarkably strong positive correlation. Further results and discussions can be found in Table 2 and Figure 1.

This paper employs seven diverse machine-learning algorithms. The Mean Absolute Error (MAE) is employed, as per Equation 3, to evaluate model errors. In this context, we compare these seven machine-learning approaches to identify the optimal variable for future predictions. The Decision Stump approach stands out with the least error, at approximately 522,021.0132, when considering the Covaxin doses administered parameter. Visual representations of these findings are presented in Table 3 and Figure 2.

To further scrutinize model performance, we employ the Root Mean Square Error (RMSE), defined in Equation 4, to quantify disparities between predicted and actual values. In this instance, the Random Forest approach demonstrates minimal error, nearly 96,800.1943, when utilizing the parameter of Covaxin Doses Administered. You can find visual representations of these outcomes in Table 4 and Figure 3.

Additionally, Relative Absolute Error (RAE) is used to assess accuracy, as per Equation 5, by comparing the differences between predicted and actual values in percentage terms. Both the Linear Regression and SMOreg approaches exhibit minimal error, nearly zero, when considering only two parameters - Covishield doses administered and the total number of vaccinated individuals. Visual representations of these findings are presented in Table 5 and Figure 4. Similar error assessments are reflected in Relative Root Square Error (RRSE), as detailed in Equation 6. Corresponding numerical data is available in Table 6 and Figure 5. The time taken is a critical aspect of machine learning methodologies, and based on the information in Table 7 and Figure 6, six of the machine learning approaches demonstrate minimal error in model construction.

4. CONCLUSION AND FUTURE RESEARCH

This research unequivocally establishes that the parameters of Covaxin doses administered, Covishield doses administered, and the total number of vaccinated individuals are suitable for predicting the future. Furthermore, we propose avenues for future research, including exploring additional data sources, investigating improved algorithms and hyperparameters, and fine-tuning the model to enhance its overall performance. This research provides valuable insights for the government departmentsand others seeking to handle the situation during the pandemic.

5. REFERENCE

[1]. Nopour, R., Kazemi-Arpanahi, H., Shanbehzadeh, M. and Azizifar, A., 2021. Performance analysis of data mining algorithms for diagnosing COVID-19. Journal of Education and Health Promotion, 10.

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

[3]. Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V. and Alhyari, S., 2020. COVID-19 prediction and detection using deep learning. International Journal of Computer Information Systems and Industrial Management Applications, 12(June), pp.168-181.

[4]. 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.

[5]. Atlam, M., Torkey, H., El-Fishawy, N. and Salem, H., 2021. Coronavirus disease 2019 (COVID-19): survival analysis using deep learning and Cox regression model. Pattern Analysis and Applications, 24(3), pp.993-1005.

[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]. Ali, S., Zhou, Y. and Patterson, M., 2022. Efficient analysis of covid-19 clinical data using machine learning models. Medical & Biological Engineering & Computing, pp.1-16.

[8]. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. and Gloaguen, R., 2020. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics, 8(6), p.890.

[9]. Yadav, R.S., 2020. Data analysis of COVID-2019 epidemic using machine

learning methods: a case study of India. International Journal of Information Technology, 12(4), pp.1321-1330.

[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/datasets/sudalairajkumar/covid19-in-india?select=covid_vaccine_statewise.csv