National Conference on Trends & Technologies in Multi-Disciplinary Research
Organized by A2Z Edu-LearningHub and
In Association with International Journal of Scientific Research in Science and Technology
Volume 9 | Issue 3 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

A Comparative Study of Different Machine Learning Models for COVID-19 Prediction in India

Cina Mathew, Cini Joseph, Dhannya J
Assistant Professor, Kristu Jyoti College of Management and Technology, Kottayam, Kerala, India

ABSTRACT

Machine Learning (ML) can be deployed very effectively to track the disease, predict the growth of the epidemic and design strategies and policies to manage its spread. Several prediction models for COVID-19 are being used by officials to make relevant control measures. Due to a high level of uncertainty and lack of essential data, standard models have shown low accuracy for long-term prediction .In several technology domains, ML models have been used to define and prioritize adverse threat variables. This study applies an improved mathematical model to analyse and predict the amount of forthcoming COVID-19-affected patients in India. An ML-based improved model has been used to predict the threats of COVID-19 in India. . In this paper, we have performed a comparative study of four machine learning standard models like linear regression (LR), decision tree, multi-layer perception (MLP) and random forest to predict the threatening variables of COVID-19. The prediction models such as Decision tree, MLP and Random forest are evaluated on the basis of loss functions such as R2 score.

Keywords : Machine Learning, Linear regression, Multi-layer perception (MLP)

I. INTRODUCTION

The COVID-19 pandemic in India is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2.Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The novel Coronavirus disease (COVID-19) has been reported to infect more than 2 million people, with more than 132,000 confirmed deaths worldwide. The recent global COVID-19 pandemic has exhibited a nonlinear and sophisticated nature [2]. In addition, the outbreak has differences with other recent outbreaks, which brings into question the ability of standard models to deliver accurate results. The cumulative incidence of COVID-19 is rapidly increasing day by day. Machine Learning Machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning are often used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. Recent pandemic has attracted many researchers around the globe to solve this problem.

This study applies an improved mathematical model to analyse and predict the expansion of the epidemic. An ML-based improved model has been applied to predict the potential threat of COVID-19 in countries worldwide. We tried to construct a meaningful machine learning model that is capable of performing predictions when fed with the dataset provided by the WHO to check the impact of the same and tried to perform basic analytic and visual operations to seek more a clear picture of the pandemic.

II. LITERATURE REVIEW

In this section, we tend to discuss concerning the opposite coincident researches associated with the COVID-19 international occurrence and its scientific applications. Scientists and practitioners everywhere the globe have pursued experimentation's on the worldwide real time COVID19 knowledge and have discovered valuable insights, patterns and information from the information. From the paper revealed in Journal of thoriac unwellness (2020), the authors showed however varied government policies to regulate COVID-19 affected the numbers. The authors have used the foremost updated COVID-19 knowledge with Associate in nursing merger of the amount of individuals migrating before and once Gregorian calendar month 23rd2020.The dataset was extracted from native and national government sites and therefore the UN agency web site. The modelling and approach was a four-step process:

(i) identification and process of COVID-19 data;

(ii) applied math model estimation for population death rates as a operate of your time since the death rate exceeds a threshold in a very location;

(iii) predicting time to exceed a given population death threshold in states early within the pandemic; and

(iv) Modeling health service utilization as an operate of deaths.

III. DATASET AND PREPROCESSING

The authentic datasets of COVID-19 have been gathered from https://www.mygov.in/ , the dataset is publicly available on cases from India from the first case index on January 30 2020. The datasets gathered were in a monthly format from January 2020 to February 2021. Table 1 displays the scenario of COVID-19 incidents in India from January 2020 to February 2021. As at February 2021COVID-19 dataset includes accumulated 15,075,501 total samples, confirmed cases of 11,063,491, 156,825 death cases.

IV. COMPARATIVE STUDY OF ALGORITHMS

For solving this particular regression problem, we have taken into consideration the following three algorithms:

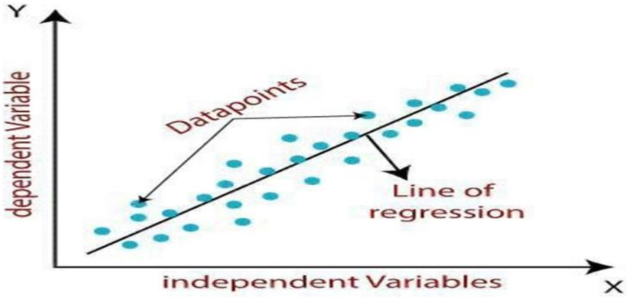
- Linear regression (LR),
- Decision tree
- Multi-layer perception (MLP)
- Random forest

These algorithms would be explained in the section below.

A. Linear Regression

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since rectilinear regression shows the linear relationship, which suggests it finds how the worth of the variable is changing consistent with the worth of the experimental variable. The rectilinear regression model provides a sloped line representing the connection between the variables.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image.



Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

Here,

y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

a0= intercept of the road

a1 = Linear regression coefficient

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

B. Decision Tree

Decision tree falls under the category of supervised machine learning algorithm. It can be used for both continuous as well as categorical output variables. It uses a decision-making flowchart like tree structure to make decisions. The branches/edges usually represent the results of the node and therefore the nodes have either:

- Conditions (Decision Nodes)
- Results (End Nodes)

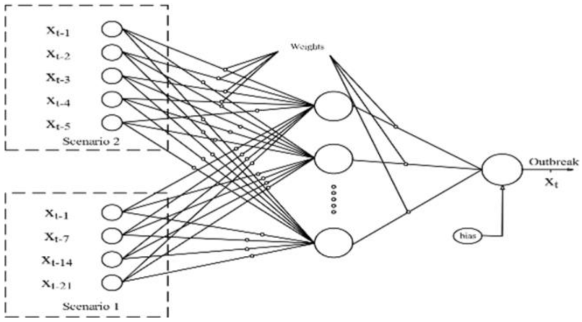
Decision tree regression observes features of an object and trains a model within the structure of a tree to predict data in the future to form meaningful continuous output. Output that is not discrete is known as continuous output, i.e., it is not represented just a discrete set of numbers or values [7].

C. Multi-layered perceptron (MLP)

ANN is a thought inspired by the biological system a nervosum, which processes information just like the brain. The key element of this concept is that the new structure of the knowledge processing system [76-78]. The neural network is about up during a learning process to perform specific tasks, like identifying patterns and categorizing information.

In the present research, one among the frequently used sorts of ANN called the MLP [76] was employed to predict the outbreak. MLP was trained using a dataset. For the training of the network, 8, 12, and 16 inner neurons were tried to realize the simplest response. Results were evaluated by R*R and correlation coefficient to reduce the cost function value.

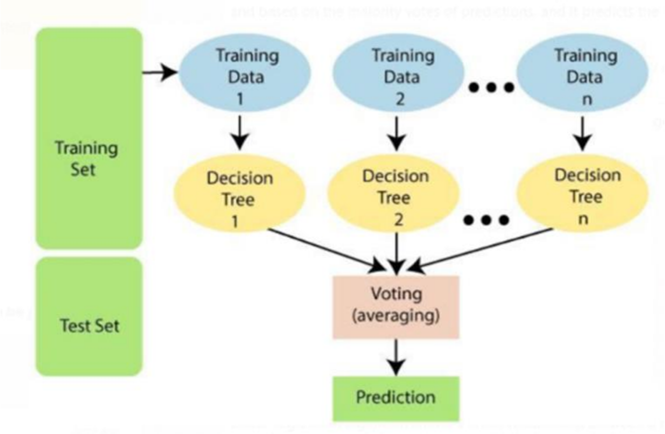
Figure 5 presents the architecture of the MLP.



D. Random Forest

Random Forest may be a popular machine learning algorithm that belongs to the supervised learning technique. It is often used for both Classification and Regression problems in ML. It is supported the concept of ensemble learning, which may be a process of mixing multiple classifiers to unravel a posh problem and to enhance the performance of the model. "Random Forest may be a classifier that contains variety of decision trees on various subsets of the given dataset and takes the typical to enhance the predictive accuracy of that dataset."

Rather than counting on one decision tree, the random forest takes the prediction from each tree and supported the bulk votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. The below diagram explains the working of the Random Forest algorithm:

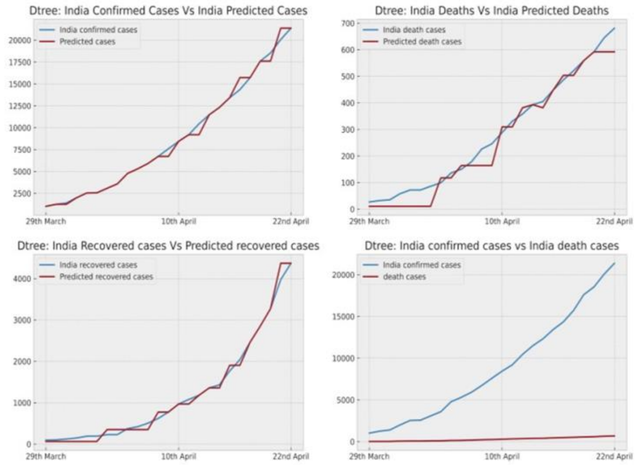


V. EXPERIMENTAL ANALYSIS

In this research, we have used MLP, Random forest and Decision tree Regressor from Python's sklearn library. The country selected for 'International COVID19 Spread Analysis' is India. All the confirmed cases, confirmed deaths, confirmed recovered cases have been analysed, compared and ran through the prediction methods for the Country.

The International COVID19 Spread comparison and analysis for India using decision tree algorithms have been shown below:

Prediction and Analysis for India Using Decision Tree:



Prediction and Analysis for India Using Decision Tree

VI. RESULTS & EVALUATION

The developed model was tested based on r2 score, which is imported using sklearn library's metrics package. The equation of r2 score is shown below:

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

Where
r = the correlation coefficient
n = number in the given dataset
x = first variable in the context
y = second variable
The Table shows the r2 score observations made when the model was executed on India's dataset:

Algorithm used and R2 Score for India's dataset

Algorithm	R2 Score
MLP	0.9988
Random Forest	0.9947
Decision Tree	-0.8215

The table clearly shows that the performance of the model is the best when the algorithm used is MLP, and its r2 score is 0.9988.

VII. CONCLUSION

In this paper, we had done the comparative study of different machine learning models for Covid 19 prediction with the collected dataset . Most popular ML models were used and was to able to achieve good result with the help of MLP. The scope of this research can be expanded by adding more country and state-level data for increasing the efficiency of this comparative study. The algorithms used for this study can be evaluated on more than one performance metric which can be mean square error (MSE), variance etc.

VIII. REFERENCES

[1]. Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, et al., “Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions,” Journal of Thoracic Disease, vol. 12, no. 3, p. 165, 2020.

[2]. I. COVID, C. J. Murray, et al., “Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months,” medRxiv, 2020.

[3]. B. Pirouz, S. Shaffiee Haghshenas, S. Shaffiee Haghshenas, and P. Piro, “Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of covid-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis,” Sustainability, vol. 12, no. 6, p. 2427, 2020.

[4]. J. T. Wu, K. Leung, and G. M. Leung, “Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study,” The Lancet, vol. 395, no. 10225, pp. 689–697, 2020.

[5]. J. Brownlee, “Time series prediction with lstm recurrent neural networks in python with keras,” Available at: machinelearningmastery.com, p. 18, 2016.

[6]. J. Brownlee, “How to develop convolutional neural network models for time series forecasting,” Available at: machinelearningmastery.com, 2018. [7] A. Das, “Python — decision tree regression using sklearn,” Available at:geeksforgeeks.org.