International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT4217255

Indian State/City Covid-19 Cases Outbreak Forecast utilizing Machine Learning Models

Mr. Brijesh Patel [1], Dr. Sheshang Degadwala[2]

[1] P.G. Students, Sigma Institute of Engineering, Vadodara, Gujarat, India

[2]Associate Professor, Sigma Institute of Engineering, Vadodara, Gujarat, India

Article Info
Volume 7, Issue 2
Page Number: 286-293

Publication Issue :
March-April-2021

Article History
Accepted : 12 April 2021
Published : 17 April 2021

ABSTRACT

Several episode expectation models for COVID-19 are being used by officials all over the world to make informed decisions and maintain necessary control steps. AI (ML)-based deciding elements have proven their worth in forecasting perioperative outcomes in order to enhance the dynamic of the predicted course of activities. For a long time, ML models have been used in a variety of application areas that needed identifiable evidence and prioritization of unfavorable factors for a danger. To cope with expecting problems, a few anticipation strategies are commonly used. This study demonstrates the ability of ML models to predict the number of future patients affected by COVID-19, which is now regarded as a potential threat to humanity. In particular, four standard evaluating models, such as Linear Regression, Support Vector Machine, LASSO, Exponential Smoothing, and Decision Tree, were used in this investigation to hypothesis the compromising variables of COVID-19. Any one of the models makes three types of predictions, for example, the number of recently Positive cases after and before preliminary vexing, the amount of passing's after and before preliminary lockdown, and the number of recuperations after and before lockdown. The outcomes demonstrate with parameters like R2 Score, Adjust R2 score, MSE, MAE and RMSE on Indian datasets.

Keywords: Linear Regression, Support vector machine, LASSO, Exponential Smoothing, and Decision Tree

I. INTRODUCTION

The healthcare market is massive and necessitates the collecting and distribution of medical evidence in real time. Furthermore, at the heart of this industry is the issue of data handling, which necessitates real-time prediction and distribution of information to clinicians in order to provide prompt medical care. Major players in this field, such as doctors, manufacturers, hospitals, and health-care providers, have sought to capture, manage, and revitalise data with the aim of using it to improve patient procedures and spur technical advancement. Dealing with healthcare data, on the other hand, has recently

become a complex challenge due to the large amount of data, security problems, cellular network device stupidity, and the rate at which it is rising. As a result, healthcare companies need data processing systems to handle such dynamic data in order to improve performance, accuracy, and workflow.

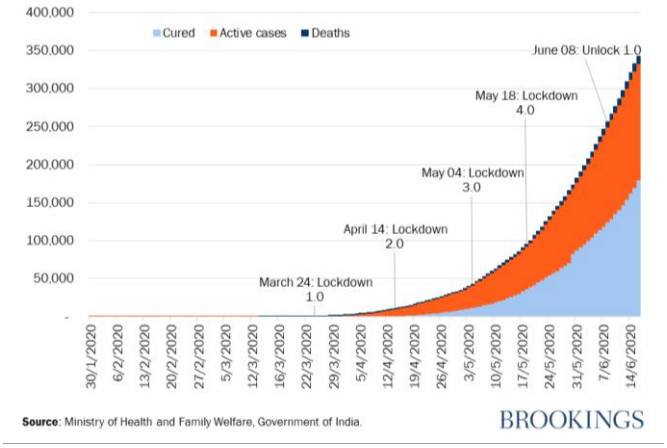

Figure 1: COVID-19 Indian cases

The entire planet is suffering as a result of Corona disease. We are referring to the forecast of daily find, death, and recovery cases of Covid-19 here. We have a large amount of data in which we would search right and reliable data in order to forecast those cases more precisely. We forecast these cases using data mining and machine learning techniques. I want to use machine learning's Regression Tree model to more correctly forecast a situation.

One of the statistical modelling methods used in analytics, data processing, and machine learning is decision tree learning. It employs a decision tree (as a statistical model) to progress from assumptions about an item (represented by branches) to judgments about the item's target value (represented in the leaves). Classification trees are tree models in which the target variable may take a distinct collection of values; in these tree systems, leaves represent class labels and branches represent function conjunctions that correspond to certain class labels. Regression trees are decision trees in which the target variable may take continuous values (typically real numbers). In this research paper, we summarize different types of regression models. For that Total positive and Negative cases are trained using LR, SVM, ES, LASSO and Regression tree. After training we will predict the data using different model and compare it with parameters.

II. RELATED WORKS

In [1] Ahmad Reshi, Mehmood Saleem Ullah, Won ON, Waqar Aslam, Furqan Rustam, and Gyu Sang Choi made ES function admirably when the time arrangement dataset has a restricted arrangement. ML-based forecasts can be exceptionally useful for chiefs to contain a pandemic like COVID-19. Live expectation progressively.

In [2] Amir Mosavi, Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, Sina F. Ardabili, Peter M. Atkinson clarified that MLP and ANFIS detailed the capacity to rehearse ordinary long haul estimating. Breaking point of this Modeling the death rate. The improvement of worldwide models with standard execution would not be conceivable. In [3] Atharva Peshkar, R. Sujatha, Jyotir Moy Chatterjee, Celestine Iwendi, Ali Kashif Bashir, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai and Ohyun Jo made expectations about the patient's potential results. The limit is to fabricate a pipeline that incorporates C * R filtering PC seeing models with these sorts of evaluation and medical care information preparing that underpins portable medical services.

In [4] Asmita B. Kalamkar, Parikshit N. Mahalle, Nilanjan Dey, Aboul Ella Hassanien, Gitanjali R. Shinde, Jyotismita Chaki, Methods Calculation Methods, Various ML Algorithms, Deep Learning. Its Benefit The forecast of the spread and generation

number should be examined in different information bases. It is restricted to limiting the disturbing effect of this pandemic. In [5] Imre Felde, Amir Mosavi, Pedram Ghamisi, Gergo Pinter, and Richard Gloaguen use ANFIS (Adaptive Network-based Fuzzy Inference System), MLP-ICA (Multi-Layered Perceptron - Imperialist Competitive Algorithm). Preferences the Gaussian MF (Membership work) gave low blunder and high precision contrasted with other MF sorts of Prediction and Mortality Rate. Yet, it is restricted to Priority to guarantee results and improve forecast quality. Progressed inside and out learning and top to bottom support learning model is enthusiastically suggested in relative examinations on different ML models in every nation.

In [6] Jyotir Moy Chatterjee and S Dhamo dharavadhani, R Rathipriya utilizing the Hybrid SNN-NAR-NN model intended to anticipate transient death rate the prescient model of the Deep learning Recurrent Neural Network time arrangement won't be utilized. In [7] Wie Kiang H. applied for Machine Learning and top to bottom perusing theory will take into consideration the assurance of day-by-day appropriations and resources. The PR model, thusly, prompts huge death toll.

In [8] Ibrahim A. Aljamaan, Salah I. Alzahrani, Ebrahim A. Al-Fakih utilized the ARIMA (Auto-Regressive Integrated Moving Average), model. Day by day passing and restoration estimates were not performed. In [9] Dhahi Alshammari, Nourah Alqahtani, Dabiah Alboaneen, Bernardi Pranggono, and Raja Alyaffer utilize the Logistic Growth model with high expectations contrasted with others. Test Plan isn't endorsed.

In [10] Chellai Fatih, Deepa Rawat, Saswati Sahu, Sagar Anand Pandey, Pradeep Mishra, M. Beam, Anurag Dubey, and Olawale Monsur Sanusi utilized ARIMA and FTS were discovered to be appropriate for foreseeing viral contaminations. Long haul direction with numerous subtleties. In [11] Manju Bala, Sukhvinder Singh Bamber, Rajbir Kaur, Mohit Angurala, Prabhdeep Singh accomplished the work effectively when the city was utilized astutely (in an incorporated way) in the different locale. New self-guideline methodologies should be utilized to conquer the issue.

In [12] Rajiv Gupta, and Farhan Mohammad Khan, utilized the ARIMA (Auto-Regressive Integrated Moving Average), the NAR (Non-straight Auto-Regressive) model Almost made the two model's equivalent. anticipating the status of confirmed, decreased, and accessible instances of Coronavirus have not been made.

III. METHODOLOGY

A. Dataset [22]

The purpose of this assessment is the future assessing of COVID-19 spread focusing in on the amount of new certain cases, the amount of passing's, and the amount of recovery. The dataset used in the examination has been gotten from the GitHub store [22].

TABLE I: BEFORE VACCINE SAMPLE DATA

No	Date	Time	State/Union Territory	Cured	Deaths	Confirmed
1	30/01/20	6:00 PM	Kerala	0	0	1
2	31/01/20	6:00 PM	Kerala	0	0	1
3	1/2/2020	6:00 PM	Kerala	0	0	2
4	2/2/2020	6:00 PM	Kerala	0	0	3
5	3/2/2020	6:00 PM	Kerala	0	0	3

The vault was essentially made open for the visual dashboard of the 2019 Novel Coronavirus by the school and was maintained by the ESRI Living Atlas

Team. The coordinator contains ordinary time game plan layout tables, including the number of avowed cases, passing's, and recoveries. All data is from the ordinary case report and the refreshed repeat of data is one day. Data tests from the records show up in Tables I, II, separately.

TABLE II: AFTER VACCINE SAMPLE DATA

No	Date	Time	State/Union Territory	Cured	Deaths	Confirmed
1	15/07/20	8:00 AM	Andaman and Nicobar Islands	109	0	166
2	15/07/20	8:00 AM	Andhra Pradesh	17467	408	33019
3	15/07/20	8:00 AM	Arunachal Pradesh	153	3	462
4	15/07/20	8:00 AM	Assam	11416	40	17807
5	15/07/20	8:00 AM	Bihar	12849	174	19284

B. Pre-Processing

Restrictive Random Field is a potential system for naming and arranging organized information, for example, successions, trees and grids. The essential thought is to characterize the restrictive prospects in a name arrangement given to a specific survey grouping, instead of dispersed all in all in both the mark succession and the view. The primary preferred position of Conditional Random Field over Markov's shrouded models is its restrictive nature, which prompts the unwinding of the autonomous intuition needed by Hidden Marko Models to guarantee an unmistakable pattern. What's more, Conditional Random Fields evades the issue of name inclination, shortcomings distinguished by very good quality entropy models Markov and other contingent Markov Model dependent on focused displaying models. Contingent Random Field outperforms both MEMMs and Hidden Marko Models in some true exercises including bioinformatics, computational etymological, and discourse acknowledgment.

C. Machine Learning

1) Linear Regression [2,10]:

In backslide illustrating, a target class is predicated on the free features. This methodology can be therefore used to remove off the relationship among free and ward factors and moreover for deciding. Direct backslide such a backslide showing is the most usable authentic procedure for perceptive examination in AI. Each discernment in straight backslide depends upon two characteristics, one is the penniless variable and the second is the free factor. Straight backslide chooses an immediate association between these poor and free factors. There are two factors (x; y) that are locked in with a straight backslide examination. The condition under shows how y is related to x known as backslide.

$$Y=b0+b1x+e \quad (1)$$

$$E(y)= b0+b1x \quad (2)$$

Here, e is the blunder term of straight relapse. The mistake terms here use to account the changeability between both x and y, b0 speaks to y-capture, b1 speaks to slant.

2) LASSO [2]

Tether is a backslide model that has a spot with the straight backslide procedure which uses shrinkage. Shrinkage in this setting insinuates the contracting of unprecedented assessments of a data test towards central characteristics. The shrinkage cycle subsequently improves LASSO and steadier and moreover diminishes the misstep. Tie is seen as a more suitable model for multicollinearity circumstances. Since the model performs L1 regularization and the discipline remembered for this case is comparable to the significance of coefficients. Thusly, LASSO makes the backslide less unpredictable to the extent the quantity of features it is using. It uses a regularization procedure for

normally rebuffing the extra features. That is the features that can't help the backslide results enough can be set to a small worth potentially zero.

3) SVM [3]:

A Support vector machine (SVM) is a coordinated AI system that is utilized for demands. SVM develops a hyperplane or set of hyperplanes in a high dimensional space, which can be utilized for demands or different undertakings like an affirmation of uncommon cases from information. A good arrangement is refined by the hyperplane that has the best parcel to the closest preparing information inspiration driving any class.

5) Exponential Smoothing [5]:

Exponential regression is the process of finding the equation of the exponential function (y=abx form where a≠0) that fits best for a set of data. In linear regression, we try to find y=b+mx that fits best data. So, exponential regression is non-linear. In both cases, though, the best fitted equation is computed in such a way that the sum of squares of distances between observed and predicted values are minimized.

Exponential smoothing a forecasting technique. The method of forecasting compares your prior forecast with your prior actual and then applies the difference between the two to the next forecast. If A is actual demand, F foretasted demand and α smoothing factor, then forecast for a period, Ft, in terms of most recent actual and forecast is:

$$Ft=\alpha At-1+(1-\alpha)Ft-1 \quad (3)$$

Here α, expressed in decimal and limited within $0<\alpha<1$, is the weighting of the most recent period's demand. Understanding of exponential smoothing should be a lot easier if you have clear concept of moving average and weighted moving average. Give the terms a look.

In short, to predict future, you use past predictions and actual data for exponential smoothing whereas you use only past data for regression.

6) Regression Tree

Each relapse methodology contains one variable (reaction) what's more, in any event one components (pointer). Yield assortment is a number. A standard tree advancement methodology that licenses foundation versatility to be a blend of tireless and stage flexibility. The decision tree is made when each decision center point in a tree contains a test in the assessment of an assortment of a particular data. The end center points in the tree contain the foreseen yield regards. Backslide tree can be considered as a variety of decision trees, expected to balance practices with real worth, instead of being used for gathering strategies. The backslide tree is outlined by a cycle known as twofold division, which is a dull cycle that isolates data into zones or branches, and a while later continues detaching each divided into more humble social affairs as the route progresses with each branch.

D. K-fold Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. The general procedure is as follows:

Step 1: Shuffle the dataset randomly.

Step 2: Split the dataset into k groups

Step 3: For each unique group:

Step 3.1: Take the group as a hold out or test data set

Step 3.2: Take the remaining groups as a training data set

Step 3.3: Fit a model on the training set and evaluate it on the test set

Step 3.4: Retain the evaluation score and discard the model

Step 4: Summarize the skill of the model using the sample of model evaluation scores.

Example: Data sample with 6 observations: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]
for k=3 fold the data is divided as follow:
Fold1: [0.5, 0.2]
Fold2: [0.1, 0.3]
Fold3: [0.4, 0.6]

IV. REGRESSION TREE BASED PREDICTION STRATEGY

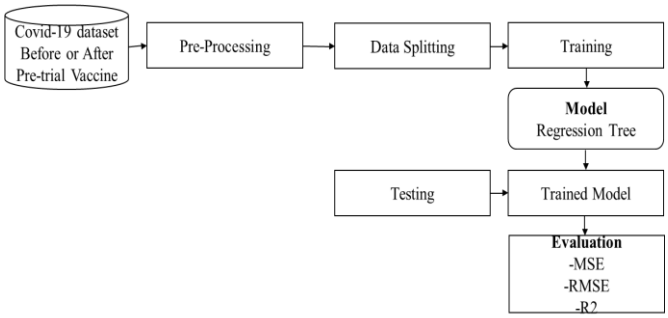

Figure 2: Prediction Strategy

Step 1: Read datasets
Step 2: Apply Pre-Processing
-Remove Null data
-Check datatypes
-Convert to time stamp ascending
Step 3: Splitting data into Train 75% training and 25% testing.
Step 4: Train Regression tree model.
Step 5: Test using Trained Model.
Step 6: Calculate Parameters R2 score, R2 score adjust, MSE, MAE, RMS and K-fold Validation.

V. RESULTS AND ANALYSIS

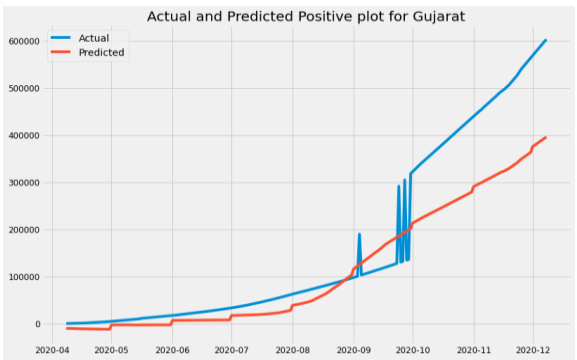

Figure 3: Linear Regression Model Prediction

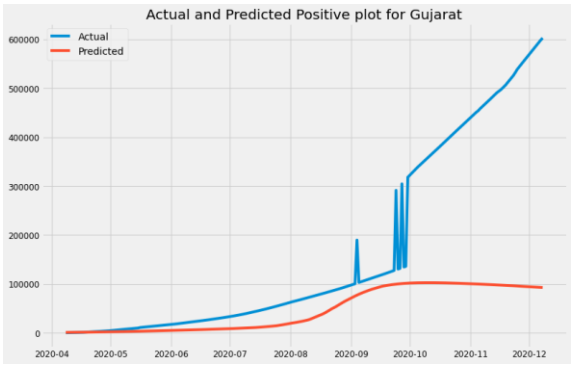

Figure 4: SVM Regression Model Prediction

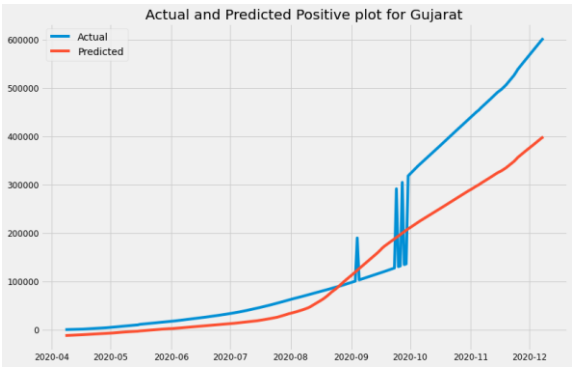

Figure 5: LASSO Regression Model Prediction

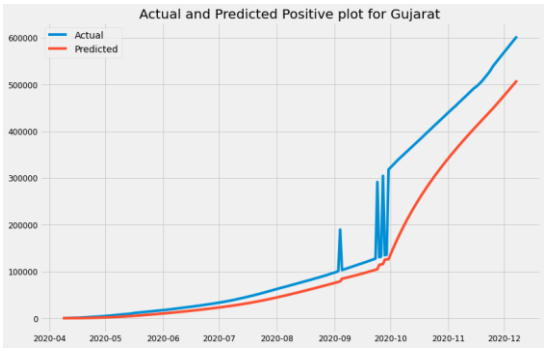

Figure 6: ES Regression Model Prediction

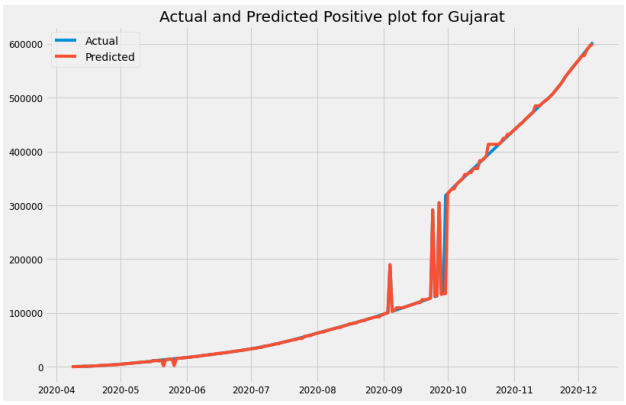

Figure 7: Regression Tree Model Prediction

TABLE III

COMPARATIVE STUDY

Model	R2 Score	R2 Adjust	MSE	MAE	RMSE	K-fold Score
LR	0.43	0.35	2.21E+10	6.06E+04	1.25E+05	0.43
SVM	0.11	0.54	3.48E+10	5.68E+04	1.87E+05	0.26
LASSO	0.43	0.35	2.33E+10	6.13E+04	1.53E+05	0.68
ES (Existing)	0.88	0.71	4.23E+09	4.21E+04	6.50E+04	0.88
Regression Tree (Proposed)	0.99	≈1	≈0	≈0	≈0	0.98

VI. CONCLUSION

Because of the ups and downs in the dataset values, we can infer that SVM and LR show bad results in all scenarios. It was extremely difficult to create an accurate hyperplane between the dataset's specified values. Other models' LASSO, ES, and Regression Tree forecasts based on the current scenario could be right, which will help us understand what is going to happen. The study predictions will therefore be of great assistance to authorities in taking prompt steps and making decisions to contain the COVID-19 crisis. In the future, we want to investigate the prediction approach with the revised dataset (via vixen) and use the most reliable and effective Machine learning methods LASSO, RF, DT, Gradient Bossing, Regression tree, etc for forecasting.

VII. REFERENCES

[1] S. F. Ardabili et al., "COVID-19 Outbreak Prediction with Machine Learning," SSRN Electron. J., 2020, doi: 10.2139/ssrn.3580188.

[2] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," IEEE Access, vol. 8, pp. 101489–101499, 2020, doi: 10.1109/ACCESS.2020.2997311.

[3] C. Iwendi et al., "COVID-19 patient health prediction using boosted random forest algorithm," Front. Public Heal., vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[4] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India," Stoch. Environ. Res. Risk Assess., vol. 34, no. 7, pp. 959–972, 2020, doi: 10.1007/s00477-020-01827-8.

[5] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, J. Chaki, and A. E. Hassanien, "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art," SN Comput. Sci., vol. 1, no. 4, pp. 1–15, 2020, doi: 10.1007/s42979-020-00209-9.

[6] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach," SSRN Electron. J., 2020, doi: 10.2139/ssrn.3590821.

[7] S. Dhamodharavadhani, R. Rathipriya, and J. M. Chatterjee, "COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models," Front. Public Heal., vol. 8, no. August, pp. 1–12, 2020, doi: 10.3389/fpubh.2020.00441.

[8] D. Gaglione et al., "Adaptive Bayesian Learning and Forecasting of Epidemic Evolution—Data Analysis of the COVID-19 Outbreak," IEEE Access, vol. 8, no. March, pp. 175244–175264, 2020, doi: 10.1109/access.2020.3019922.

[9] M. Jain, "Pandemic in India," no. Icces, pp. 784–789, 2020.

[10] A. Andreas, C. X. Mavromoustakis, G. Mastorakis, S. Mumtaz, J. M. Batalla, and E. Pallis, "Modified Machine Learning Techique for Curve Fitting on Regression Models for COVID-19 projections," IEEE Int. Work. Comput. Aided Model. Des. Commun. Links Networks, CAMAD,

vol. 2020-September, no. December 2019, 2020, doi: 10.1109/CAMAD50429.2020.9209264.

[11] S. Singh, P. Raj, R. Kumar, and R. Chaujar, "Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models," Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020, pp. 93–97, 2020, doi: 10.1109/ICIRCA48905.2020.9183126.

[12] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," Sci. Total Environ., vol. 728, p. 138762, 2020, doi: 10.1016/j.scitotenv.2020.138762.

[13] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming," Chaos, Solitons and Fractals, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109945.

[14] K. Roosa et al., "Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020," Infect. Dis. Model., vol. 5, pp. 256–263, 2020, doi: 10.1016/j.idm.2020.02.002.

[15] L. Li et al., "Propagation analysis and prediction of the COVID-19," Infect. Dis. Model., vol. 5, pp. 282–292, 2020, doi: 10.1016/j.idm.2020.03.002.

[16] J. P. A. Ioannidis, S. Cripps, and M. A. Tanner, "Forecasting for COVID-19 has failed," Int. J. Forecast., no. xxxx, 2020, doi: 10.1016/j.ijforecast.2020.08.004.

[17] M. Yousaf, S. Zahir, M. Riaz, S. M. Hussain, and K. Shah, "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan," Chaos, Solitons and Fractals, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109926.

[18] K. Sarkar, S. Khajanchi, and J. J. Nieto, "Modeling and forecasting the COVID-19 pandemic in India," Chaos, Solitons and Fractals, vol. 139, pp. 1–16, 2020, doi: 10.1016/j.chaos.2020.110049.

[19] K. N. Nabi, "Forecasting COVID-19 pandemic: A data-driven analysis," Chaos, Solitons and Fractals, vol. 139, p. 110046, 2020, doi: 10.1016/j.chaos.2020.110046.

[20] M. Wieczorek, J. Siłka, and M. Woźniak, "Neural network powered COVID-19 spread forecasting model," Chaos, Solitons and Fractals, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110203.

[21] B. Malavika, S. Marimuthu, M. Joy, A. Nadaraj, E. S. Asirvatham, and L. Jeyaseelan, "Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models," Clin. Epidemiol. Glob. Heal., no. May, pp. 1–8, 2020, doi: 10.1016/j.cegh.2020.06.006.

[22] COVID-19 in India [Internet]. Kaggle.com. 2020 [cited 30 December 2020]. Available from: https://www.kaggle.com/sudalairajkumar/covid19-in-india

[23] Dhaya, R. "Deep net model for detection of covid-19 using radiographs based on roc analysis." Journal of Innovative Image Processing (JIIP) 2, no. 03 (2020): 135-140.

[24] Muthukumar, Vignesh, and N. Bhalaji. "MOOCVERSITY-Deep Learning Based Dropout Prediction in MOOCs over Weeks." Journal of Soft Computing Paradigm (JSCP) 2, no. 03 (2020): 140-152

Cite this article as :

Brijesh Patel, Dr. Sheshang Degadwala, "Indian State or City Covid-19 Cases Outbreak Forecast utilizing Machine Learning Models", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 286-293, March-April 2021. Available at
doi : https://doi.org/10.32628/CSEIT4217255
Journal URL : https://ijsrcseit.com/CSEIT4217255