Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 816-821
Received 15 April 2021; Accepted 05 May 2021.

Covid-19 Analysis Using Machine Learning Models

V. Rohith, P. Rishi Kumar and P. Mahalakshmi
Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Chennai, India.

Abstract- The outbreak of the novel coronavirus (Covid-19) is a deadly disease which has taken the lives of millions all around the globe which has since been deemed as a pandemic by the World Health Organization (WHO). Every nation is forced to take action to control the situation. The fundamental objective of this paper is to analyze the development rate of the virus and help in minimizing the spread of this disease. This paper also aims to find out whether the mitigation strategies implemented by the government have been effective. Using various machine learning models to forecast the transmission pace and provide a helpful insight of the upcoming days. Various models with the latest data have been assessed and contrasted with each other and its outcome shows its prevalence in both effectiveness and precision.

Key-terms- CoronaVirus, India, Data Science, Data Analysis, Predictions, Trend Analysis, Forecasting, Lockdown, Social Distancing, Support Vector Machines, Regression, Time-Space Analysis

I. INTRODUCTION

AI and Machine Learning has been substantiated as an unmistakable domain throughout the most recent years by settling numerous extremely mind boggling and modern genuine issues. The application territories included practically all these present reality areas like medical services, autonomous vehicles (AV), natural language processing (NLP), business applications, intelligent robots, gaming, environment modeling, voice, and image processing. ML algorithms' learning is regularly founded on experimentation technique, very inverse of customary calculations, which adheres to the programming directions dependent on decision statements. The main focus of ML is predicting, various standard ML calculations have been utilized around there to control the future path of activities required in numerous application regions namely climate forecasting, stock market forecasting, disease forecasting. Different regression and neural network models have broad applicability in anticipating the states of sick people later on with a particular illness [1]. Many studies have been performed to forecast the various illnesses utilizing AI methods like coronary artery disease, breast cancer prediction and cardiovascular disease prediction. Specifically, the investigation is focused on real-time prediction of COVID-19 affirmed cases and study is likewise centered around the prediction of COVID-19 breakout. These forecast frameworks can be useful in dynamic ways to deal with the current situation to direct mediations to deal with these infections viably. This examination intends to give an estimated model to the spread of novel Covid, otherwise called SARS-CoV-2, formally given the name as COVID-19 by the World Health Organization (WHO). Coronavirus is by and by an intense danger to human existence everywhere in the world. Toward late 2019, the first reported infection was in Wuhan, City in China, when an enormous number of individuals created side effects like pneumonia. It diversely affects the human body, including extreme intense respiratory disorder and multi-organ failure which can at last prompt passing in a brief span. Countless individuals are influenced by this pandemic all through the world with a large number of passing each coming day. “A large number of new individuals are accounted for to be infected consistently from nations around the world. The infection spreads principally by individual to individual actual close contacts, by respiratory drops, or by contacting the contaminated surfaces. The most difficult part of its spread is that an individual can have the infection for a long time without showing indications. The reasons for its spread and thinking about its peril, practically all the nations have announced either halfway or exacting lockdowns all through the influenced areas and urban communities.”[1] Clinical specialists all through the globe are presently included to find a suitable immunization and drugs for the illness. As there are no endorsed prescriptions as of now for arresting the infection the legislatures of all nations are zeroing in on the safety measures that can help mediate propagation. Of all the safeguards, "being educated" pretty much all the parts of COVID-19 are advised to be critical. To add to this part of data, various specialists are contemplating the various components of the outbreak and construct the outcomes to aid mankind.

II. RELATED WORK

In this section we will discuss some previous works in this domain, with their merits and unexplored fields.

Adaptive Phase-Space Approach: This proposes a new approach using data-guided detection and aggregation of infection waves, these waves are generated by the Riccati equation(I) and therefore are called "Riccati modules". This approach is applied to daily data of confirmed cases of coronavirus in the US, resulting in the epidemic time-period to break down into five Riccati modules representing major infection waves till date. This approach provides robust estimation as the concept of concatenating infection waves adds to the adaptability of the model to some extent.

$$dx(t)/dt = Ax(t) - Bx^2(t) + R(t)$$

Polynomial Regression: It is a supervised ML approach that is used when the two variables are correlated in a non-linear relationship. Therefore, a polynomial function that fits our regression pattern is taken and the dataset is trained accordingly. Due to the daily escalations of reported covid cases fluctuating erratically, finding a polynomial expression to best fit the curve becomes highly inefficient and demands heavier computation.

$$y= b_0+b_1x + b_2x^2+ b_3x^3+....+ b_nx^n$$

Logistical growth model: The rate of cases fits in a polynomial or exponential curve only in the beginning stages of a pandemic; therefore, a more rigorous model is required to map a proper curve in the later stages of a pandemic. Logistical growth models (II) can more accurately draw the growth and fall of the data values when compared to previous approaches.

$$dX(t)/dt = r(1 - X(t)/c)X(t)$$

Support Vector Regression: The SVR model provides accurate results but only is stable only for smaller datasets. With larger datasets, the accuracy goes down significantly therefore the option of scalability is limited. Selecting the appropriate kernel can be tricky and model performance is very slow for a large dataset.

III. PROPOSED WORK

To obtain the latest dataset on Covid-19 in India, BeautifulSoup will be employed. BeautifulSoup is a web scraping tool in python that parses HTML and XML files and converts them to our desired .csv file. This python library along with PrettyTables, another python library to work with data, is used to scrape data from the Ministry of Health and Family Welfare website (MOHFW) (www.mohfw.gov.in/) and (www.covid19india.org/) to obtain the latest data. Resources from Kaggle will also be used to improve the quality of the dataset. The dataset should be cleaned and normalized which is removing the Nan values and filling in values that are left empty. The empty data are either filled with 0 or 1 or average from the column to provide a uniform dataset which is both easy to work with and provides more accurate results. This paper proposes using all previous performed machine learning models like Linear regression, Polynomial regression, Exponential regression, Bayesian Ridge regression, Support Vector regressor (SVR) and comparing them with newly implemented machine learning models. This paper proposes to use time series forecasting which use model such as Auto-Regression model (AR), Moving Average model (MA), Auto-Regressive Integrated Moving Average (ARIMA), Seasonal Auto-Regressive Integrated Moving Average (SARIMA), FBProphet and Random Forest Regressor. These models are input with training and test data which are selectively picked for each model to provide the best accuracy and lowest Root Mean Square Error. All these data models are finally compared with each other and the model with the least Root Mean Square Error is used for the forecasting. From research, Random Forest Regressor has promised to be the best model for forecasting with the least Root Mean Square Error for the data set and will be used to to protect spread of the virus for the next 21 days. The remaining models will be used for visualising the data for better understanding about the situation in India and the spread of the virus.

IV. IMPLEMENTATION

BeautifulSoup is a web scraping tool in python that parses HTML and XML files and converts them to our desired .csv file. This python library along with PrettyTables, another python library to work with data, is used to scrape data from the Ministry of Health and Family Welfare website (MOHFW) (www.mohfw.gov.in/) and (www.covid19india.org/) to obtain the latest data. Resources from Kaggle will also be used to improve the quality of the dataset. Send a HTTP request to MOHFW site and the server reacts to the request by returning the mentioned HTML content. For this task , we will utilize an outsider HTTP library for python-requests. Once we have gotten to the HTML content the information is parsed. Since a large portion of the HTML information is nested, we cannot extract information just through string processing. we need a parser which can make a tree design of the HTML information. Html5lib is utilized for this undertaking. We should simply navigate and look through the parse tree that has been made. This is where BeautifulSoup is utilized.

The data has been visualized to provide a better understanding of the situation using python packages such as seaborn and matplotlib.

Linear regression, Polynomial regression exponential regression Bayesian ridge regression is implemented with confirmed cases count in our dataset. Root Mean Square Error has been calculated and added to a table for comparison later.

Previously obtained data has been converted to a time series forecasting compatible dataset using the operation called shift() in Python Pandas library. This data contains total confirmed cases, total deaths, total recovered between intervals of time. Data has been collected from March 20th 2020 to March 20th 2021. This time series forecasting compatible data set has been used to train supervised machine learning models such as Autoregressive model(AR), Moving Average model(MA) Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). Holt's Linear Model and Holt’s Exponential Smoothing.

Prophet, or FBProphet is an open-source library created by Facebook for time series forecasting. It implements a summative time series forecasting model, and the exertion supports trends, seasonality, and holidays. It is designed to be easy to use and completely automatic in nature. The time series dataset which we obtained is fit in FBProphet and the confirmed cases has been forecasted for the next 21 days. The RMSE is calculated and added to a table to compare the models at the end.

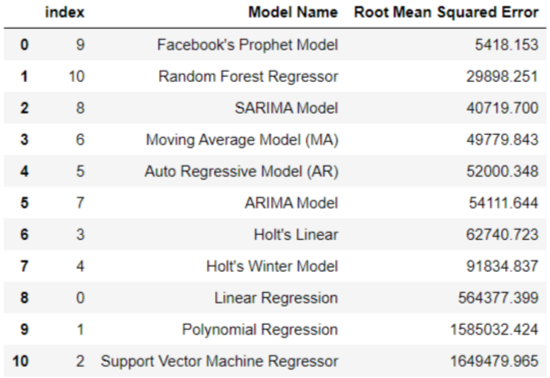
Received 15 April 2021; Accepted 05 May 2021.

Random Forest Regressor is an estimator which fits numerous classifying decision trees on multiple subsets of the data and uses the average to improve its predictive accuracy. This model works well with large datasets unlike Support Vector regressors and has very little impact on outliers. It handles missing data very well and has no problem of overfitting. The data is subsetted into 100 for training the model and tested with double the size of training data. A max depth of 6 has been chosen and 25 trees per subset has been employed to provide the best results. The parameters are tweaked even further to get the best possible result.

V. RESULTS AND DISCUSSIONS

The errors of all the models have been stored in a table and compared side by side. It has been found that FBProphet has the least RMSE followed by Random Forest Regressor and SARIMA model.

Random Forest Regressor has been used to forecast the next 21 days on the spread of coronavirus remaining models have been used to visualise the dataset to provide a better understanding of the situation in our country and how the various mitigation strategies have been effective. The data visualisation should aid the government and the concerning bodies understand the situation and take precautionary actions. The active cases, death counts, recovered counts have been predicted by the model for the next three weeks.



	index	Model Name	Root Mean Squared Error
0	9	Facebook's Prophet Model	5418.153
1	10	Random Forest Regressor	29898.251
2	8	SARIMA Model	40719.700
3	6	Moving Average Model (MA)	49779.843
4	5	Auto Regressive Model (AR)	52000.348
5	7	ARIMA Model	54111.644
6	3	Holt's Linear	62740.723
7	4	Holt's Winter Model	91834.837
8	0	Linear Regression	564377.399
9	1	Polynomial Regression	1585032.424
10	2	Support Vector Machine Regressor	1649479.965

VI. CONCLUSION

This investigation has provided an extensive examination of the Covid-19 outbreak in India. The infected cases are rising quickly, and effective control techniques need to be implemented by the government of India. The growth patterns of infected cases in India have been illustrated and the prediction of the number of coronavirus cases for the following days have been showcased, the result of lockdown and social distancing on the residents of India can also be seen on the graph. This study will be useful for the Government of India and different states of India, Frontline wellbeing workforce of India, analysts and researchers. This study will likewise be ideal for the governing bodies of different nations to consider different views identified with the control of coronavirus spread in their respective nations.

VII. REFERENCES

1. F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.1.
2. P. Ghosh, R. Ghosh, B. Chakraborty, “COVID-19 in India: State-wise Analysis and Prediction”, medRxiv, May 19 (2020).
3. P.Mahalakshmi and N.Sabiyath Fatima, “An art of review on Conceptual based Information Retrieval”, Webology Journal, volume 18, issue no. 1, pp. 51-61, 2021.
4. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al., China Novel Coronavirus Investigating and Research Team. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb,382(8):727–33.
5. Almeida JD, Tyrrell DA. The morphology of three previously uncharacterized human respiratory viruses that grow in organ culture. J Gen Virol. 1967 Apr,1(2):175–8.
6. Kapikian AZ, James HD Jr, Kelly SJ, Dees JH, Turner HC, McIntosh K, et al. Isolation from man of “avian infectious bronchitis virus like” viruses (coronaviruses) similar to 229E virus, with some epidemiological observations. J Infect Dis. 1969 Mar,119(3):282–90.
7. P. Mahalakshmi and N. S. Fatima, "Collaborative Text and Image based Information Retrieval Model using BiLSTM and Residual Networks," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 958-964, 2020.
8. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. Nat Med. 2004 Dec, 10(12 Suppl):S88–97.
9. van der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, et al. Identification of a new human coronavirus. Nat Med. 2004 Apr,10(4):368–73
10. S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi Koczy, U.Reuter, T. Rabczuk, P. M. Atkinson, “COVID-19 Outbreak Prediction with Machine Learning”, April 19 (2020).
11. S. Zhao, H. Chen, “Modeling the epidemic dynamics and control of COVID-19 outbreak in China”. Quantitative Biology, vol. 8, Issue 1, March (2020).
12. Mahalakshmi, P., Fatima, N.S. Ensembling of text and images using Deep Convolutional Neural Networks for Intelligent Information Retrieval. Wireless Pers Commun (2021)
13. W.C.Roda, M. B.Varughese, D. Han, M. Y. Lia, “Why is it difficult to accurately predict the COVID-19 epidemic?”, Infectious Disease Modelling, vol. 5, (2020).
14. R. Gupta, S.K. Pal, G. Pandey, “A Comprehensive Analysis of COVID-19 Outbreak Situation in India”, Published: April (2020).
15. F. Petropoulos, S. Makridakis, “Forecasting the novel coronavirus COVID-19”, Published: March 31 (2020).
16. N. S Punn, S. K. Sonbhadra, S. Agarwal, “COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms”, medRxiv,, June 1 (2020).
17. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep. 1985 Oct;69(10):1071-77. PMID: 4042087.