An Efficient Covid19 Epidemic Analysis and Prediction Model Using Machine Learning Algorithms

https://doi.org/10.3991/ijoe.v17i11.25209

A.Lakshmanarao[1](✉), M. Raja Babu[1], T.Srinivasa Ravi Kiran[2]
[1] Aditya Engineering College, Surampalem, India
[2] P.B. Siddhartha College of Arts & Science, Vijayawada, India
laxman1216@gmail.com

Abstract—The whole world is experiencing a novel infection called Coronavirus brought about by a Covid since 2019. The main concern about this disease is the absence of proficient authentic medicine The World Health Organization (WHO) proposed a few precautionary measures to manage the spread of illness and to lessen the defilement in this manner decreasing cases. In this paper, we analyzed the Coronavirus dataset accessible in Kaggle. The past contributions from a few researchers of comparative work covered a limited number of days. Our paper used the covid19 data till May 2021. The number of confirmed cases, recovered cases, and death cases are considered for analysis. The corona cases are analyzed in a daily, weekly manner to get insight into the dataset. After extensive analysis, we proposed machine learning regressors for covid 19 predictions. We applied linear regression, polynomial regression, Decision Tree Regressor, Random Forest Regressor. Decision Tree and Random Forest given an r-square value of 0.99. We also predicted future cases with these four algorithms. We can able to predict future cases better with the polynomial regression technique. This prediction can help to take preventive measures to control covid19 in near future. All the experiments are conducted with python language.

Keywords—Covid19, Kaggle, machine learning, regression

1 Introduction

The novel corona Covid 2019 (COVID-19) pandemic began in Wuhan, China, in December 2019 and is a real broad clinical issue throughout the world. Corona Viruses are a colossal class of contaminations that cause afflictions achieved by cold, for instance, the Middle East respiratory condition Covid and serious in-tense respiratory disorder Covid. The COVID-19 is another type of Coronavirus family found in Wuhan in the year 2019. Studies show that the SARS-CoV dis-ease defiles civet individuals and the MERS-CoV contamination pollute dromedary individuals. The COVID-19 disease is acknowledged to be shrunk by individuals from bats. The disease spreads very fastly from one person to another person. There are also some studies saying that this virus can be transmitted through the air also. Some variants of corona also affecting the animals [1]. Although most of the countries suffered a lot in the first wave of covid-19,

the transmission rate and death rate of the second wave of covid-19 is very dangerous when compared to the first phase. New coronavirus variants (like delta variants) are generated and struggling the world.

2 Literature review

Machine Learning and Deep learning playing a vital role in the health sector [2]. Applying ML models for disease prediction is not new. Several authors also applied ML models to covid-19. Degadwala, S [3]. et al applied Convolutional Neural Networks for the classification of covid-19 cases. They collected X-ray chest images of 1560 covid patients and with their model they achieved an accuracy of 90%. Prathyusha K. [4] et.al applied various machine learning regression algorithms like linear regressor, polynomial regressor and achieved the best results with the polynomial regression technique. A. Lakshmanarao [5] et.al applied various regression techniques for analyzing and predicting corona disease and achieved good results with linear regression. Sumayh S. Aljameel [6] et.al applied three classification algorithms random forest, Gradient Boosting, and Logistic Regression. As they have taken an unbalanced dataset, first they applied SMOTE sampling technique, later they achieved an accuracy of 99% with random forest classification. Yazeed Zoabi [7] et.al applied a machine learning model for predicting covid-19 with eight binary features and achieved good accuracy.S. Dhamodharavadhani [8] et.al applied a Neural Network-based method for the prediction of the mortality rate of corona disease and achieved good results Archana Kalidindi[9] et.al applied deep learning for classification of human brain and achieved good results. Malki Z. [10] et.al proposed a machine learning oriented covid19 prediction model and predicted that this pandemic decline in September 2021. They applied four classifiers namely K-NN, support vector machine, random forest, decision trees. Nishitha [11] et.al applied machine learning techniques for chest im-ages and achieved good results. Sanjay Kumar [12] et.al applied machine learning for analyzing the vaccination process in India.Barmparis G.D. [13] et.al applied the Gaussian spreading model for estimating the infection horizon of novel corona disease. They successfully predicted the daily cases with their model. Meryem Fakhouri Amr [14] implemented a model for transforming CIM (Computing Independent Model) model to the PIM (Platform Independent Model) model according to the MDA (Model Driven Architecture) approach for covid-19 patient management.

3 Proposed methodology

The proposed methodology was depicted in Figure 1. First, we collected a dataset from Kaggle. Then analyze the dataset to find daily cases, weekly cases country-wise. Later, we applied regression techniques, holt's method to predict future trends.

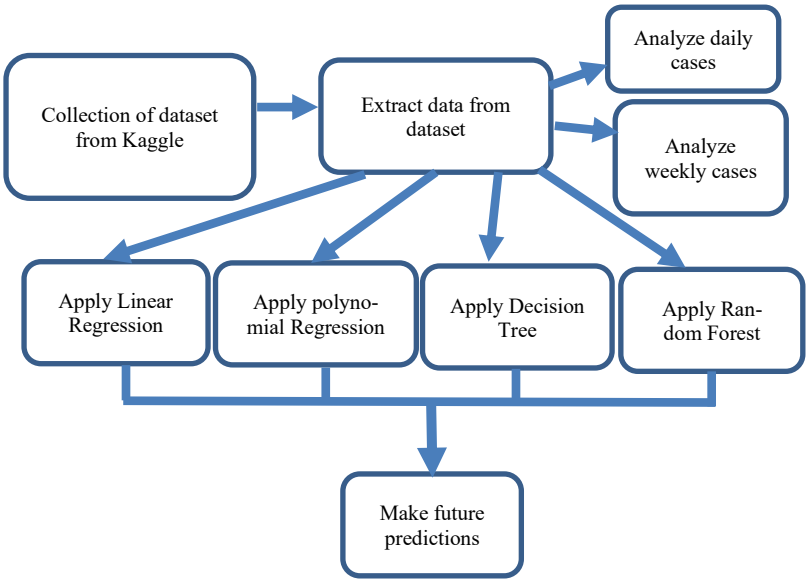

Fig. 1. The proposed methodology

3.1 Dataset

The covid dataset was collected from Kaggle [15]. The dataset contains features namely "country", "State", "Date", "Confirmed cases", "Recovered cases", and "Death cases". The data was arranged in the date-wise style. From that, we extracted the number of confirmed, death, and recovered cases, active & closed cases. (Shown in Table 1).

Table 1. Details of cases extracted from dataset

Item (Time period: Jan2020-May2021)	**Total no. of confirmed**
Total no. of confirmed	169951560
Total no. of recovered	107140669
Total no. of death cases	3533619
Total no. of active cases	59277272
Total no. of closed cases	110674288

3.2 Analysis of active cases

The active cases number is always increasing drastically from the identifying first few cases. The distribution plot for active cases was shown in Figure 2.

From Figure 2, it is observed that active covid cases are always increasing. From Jana 2020 to May 2021, the cases are increasing drastically even though precautions are taken.

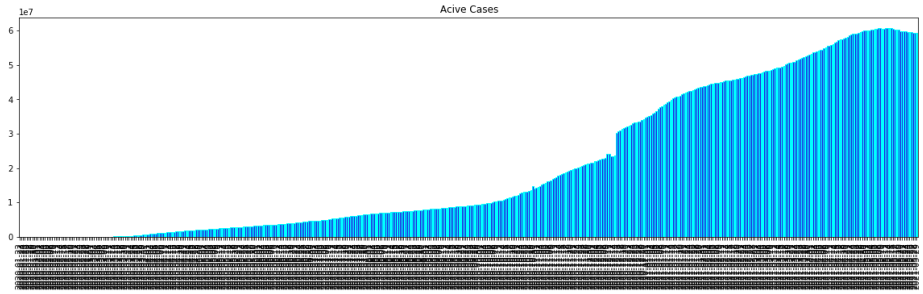

Fig. 2. Distribution plots of active cases

3.3 Analysis of weekly cases

Weekly progress of cases is depicted in Figure 3. (Red-Death cases, blue-confirmed cases, green-recovered cases.)

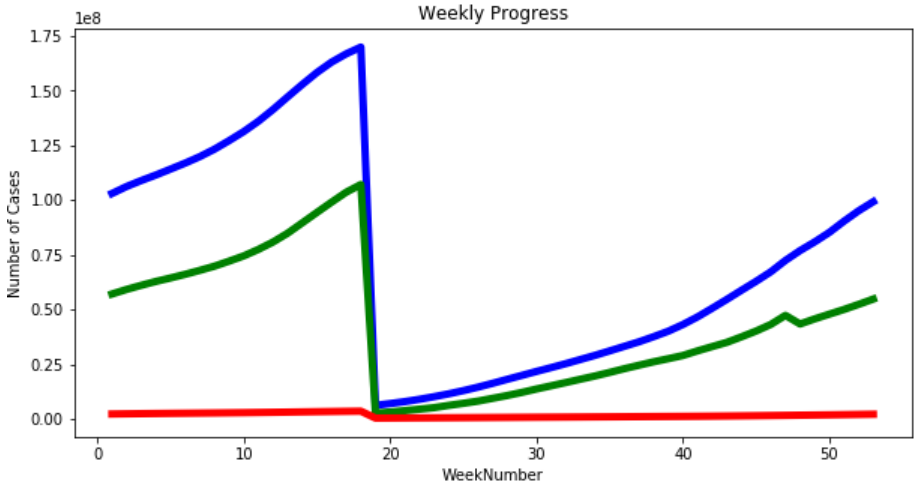

Fig. 3. Weekly progress of cases worldwide

From Figure 3, it is observed that the number of death cases is proportional to total cases. Total confirmed cases are very high around the 20th week (after jan2020). After that, confirmed and recovered cases are following the same trend.

3.4 Analyzing mortality rate

The mortality data of the top 10 countries are shown in Figure 4 (Until May 29th, 2021). The top three countries in confirmed cases are the US, India, Brazil. The top three countries in death cases are the US, Brazil, India. So, these three countries facing a health emergency with covid-19.

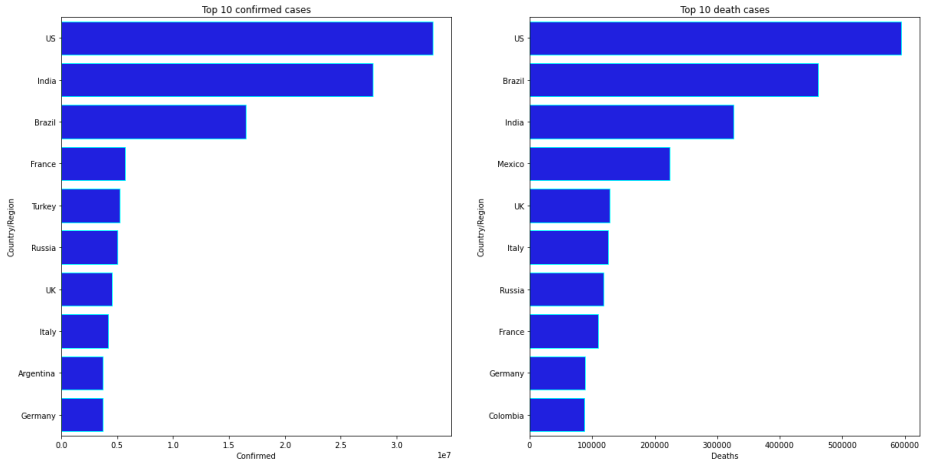

Fig. 4. Top 10 countries in terms of confirmed(left)cases, death(right)cases

4 Experimentation and results

Predicting number of cases is a regression problem. We applied ML regression algorithms for covid19 prediction. We applied several regression algorithms on the dataset to predict covid19.The day number is considered as the independent variable and the number of cases is considered as the dependent variable. We applied several regression models, but only four of them done well for this covid prediction.

4.1 Linear regression

Linear Regression is a basic algorithm where output variable is predicted based on the input variable. Here day number is the input variable and the number of cases is the output variable. The dataset contains 494 samples (494 days). The dataset is divided into training and testing sets in a 70%:30% split. Training set contains 345 days cases and the testing set contains 149 days cases. Later we applied linear regression and achieved an r-square value of 0.90. Figure 5 shows test set results after applying linear regression.

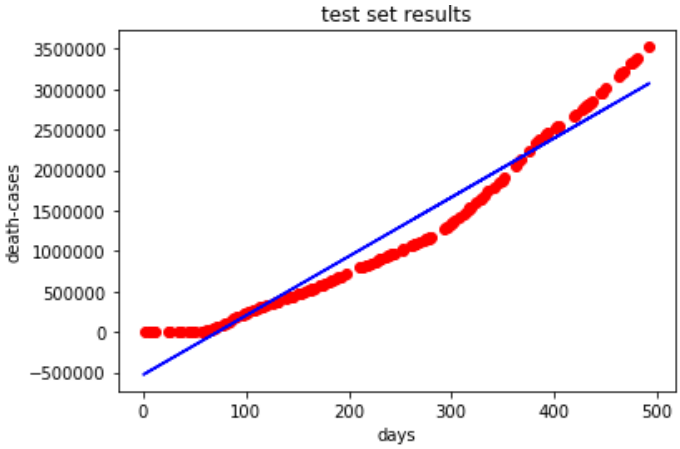

Fig. 5. Linear Regression plot for testing data

4.2 Polynomial regression

In polynomial regression, nth degree polynomial relation is established between independent and dependent variables. As the covid cases are not increasing linearly, polynomial regression applicable to predict cases. We applied polynomial regression with several degrees and find best solution with a degree of 4. Figure 6 shows test set results after applying polynomial regression. With polynomial regression,.098 r-squared value is achieved.

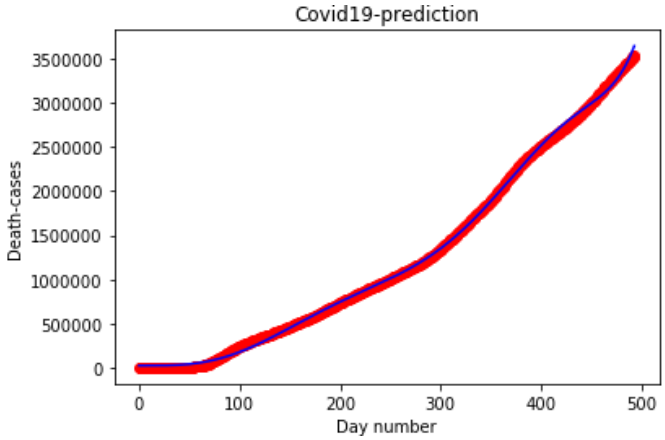

Fig. 6. Polynomial Regression plot for testing data

4.3 Decision tree regression

Decision Tree regression is a tree-based ML model. Based on the given input, output value at leaf node is predicted as output. We applied decision tree regression with entropy and achieved an r-squared value of 0.99.

4.4 Random forest regression

Random Forest is an ensemble model. It combines several decision trees. We applied random forest with 60 decision trees and achieved 0.99 r-squared value.

4.5 Comparison of regression algorithms

We applied four regression methods namely linear regression, polynomial regression, decision tree regression, random forest regression for three different cases namely confirmed cases, recovered cases, death cases. The results of the performance are given in Table 2. From Table 2, it is observed that Random Forest and Decision Tree performed well for covid-19 prediction. Although, four algorithms given good r-squared values, we further compared the regressors with respect to future confirmed cases. The dataset contains the number of confirmed cases up to 29-5-2021.So, we predicted number of cases on 1-7-2021(as on date) and checking which algorithm is doing good. For this we collected number of confirmed cases from [16] as on date. The comparison of all these algorithms for predicting number of cases on 1-7-2021 are shown in Table 3.

Table 2. Comparison of regression algorithms

Algorithm	**R-squared value**		
	Confirmed Cases	***Recovered Cases***	***Death Cases***
Linear Reg	0.90	0.91	0.94
Polynomial Reg	0.98	0.99	0.99
DTR	0.98	0.99	0.99
RF	0.98	0.99	0.99

Table 3. Comparison of regression algorithms for future predictions

Actual confirmed cases [16] (as on date 1-7-21)	181,722,790
Linear Regression predicted cases	152,430,298
Polynomial Regression predicted cases	181,723,467
Decision Tree Regression predicted cases	169,951,560
Random Forest Regression predicted cases	169,579,302

From Table 3, it is observed that polynomial regression performed well for future predictions. Although DTR, RF given good r-squared values, they are unable to predict future cases.

5 Conclusion

In this paper, we collected a covid-19 dataset from Kaggle and analyzed the number of confirmed, recovered, death cases in a daily and weekly manner. Later we applied four regression algorithms on the dataset and achieved a good r-squared of 0.99 with decision tree and random forest. Later, we tried to predict the number of future cases with all four algorithms. Polynomial Regression achieved good results while predicting future cases.

6 References

[1] Pourhomayoun, Shakibi.M.M, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making", Smart Health 20 (2021) ,100178, ELSEVIER,2021. https://doi.org/10.1016/j.smhl.2020.100178

[2] George. A. Arjun Vijayanatha Kurup, Parthasarathy Balachandran, Manjusha Nair, Siby Gopinath, Anand Kumar, Harilal Parasuram "Predicting Autonomic Dysfunction in Anxiety Disorder from ECG and Respiratory Signals Using Machine Learning Models", International Journal of Online and Biomedical Engineering, (iJOE), 2021 volume-17, No.7,2021223-224. https://doi.org/10.3991/ijoe.v17i07.22581

[3] Degadwala, S,Vyas.D,Dave.H, "Classification of COVID-19 cases using Fine-Tune Convolution Neural Network (FT-CNN)", In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 609-613, https://doi.org/10.1109/icais50930.2021.9395864

[4] Prathyusha.K, Helini.K,Raghavendran,C.V.,Kumar Kurumeti, N.S.L,"COVID-19 in India: Lockdown analysis and future predictions using Regression models",11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021. https://doi.org/10.1109/confluence51648.2021.9377052

Lakshmanarao. A, Seshadrirao. Ch,Sridevi.G, "Analyzing and Predicting Covid-19 epidemic using Machine Learning Techniques", In 2021 IOP Conference. Ser.: Mater. Sci. Eng. 1074 012018, International Conference on Computer Vision, High Performance Computing, Smart Devices and Networks (CHSN 2020). https://doi.org/10.1088/1757-899x/1074/1/012018

[5] Sumayh. S, Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri, Eman S. Alsulmi,"Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID19 Patients", Scientific Programming, vol. 2021, Art-ID-5587188,2021. https://doi.org/10.1155/2021/5587188

Zoabi.Y,Deri Rozov.S,Shomron.N, "Machine learning-based prediction of COVID-19, diagnosis based on symptoms", npj Digital Medicine.4,3 (2021), https://doi.org/10.1038/s41746-020-00372-6

[6] Dhamodharavadhani.S,Rathipriya.R,Jyotir Moy Chatterjee, "COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models", Frontiers in Public HealthVolume-8, Article-441. https://doi.org/10.3389/fpubh.2020.00441

[7] Sumayh. Archana.K,Prasanna Lakshmi.K,Sairam.B,Raagh Rao.A, "CT Image Classification of Human Brain Using Deep Learning", International Journal of Online and Biomedical Engineering (iJOE),2021, volume-17, No.1,2021.

[8] Malki.Z,Atlam.ES,Ewis.A. et al, "The COVID-19 pandemic: prediction study based on machine learning models", Environ Sci Pollut Res (2021). https://doi.org/10.1007/s11356-021-13824-7
[9] Nishitha, K.,Shiny Gracy,C., Priyadharshini,B., Sowmiya,M.,Kadhar Basha,N, "Covid Prediction using Machine Learning", International Journal Of Engineering Research & Technology (IJERT) Volume 10, Issue 05 (May 2021).
[10] Sanjay Kumar, Sumant Kumar, Aditya Singh, Anand Raj, "COVID-19 Data Analysis and Prediction Using (Machine Learning) and Vaccination Update of India", (May 17, 2021), https://doi.org/10.2139/ssrn.3847564
[11] Barmparis, G.D., Tsironis G.P, "Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach", Chaos, Solitons and Fractals 135 (2020) 109842, 2020 Elsevier. https://doi.org/10.1016/j.chaos.2020.109842
[12] Fakhouri Amr M.,Benmoussa,N.,Mansouri, K.,Mohammed Qbadou "Transformation of the CIM Model into A PIM Model According to The MDA Approach for Application Interoperability: Case of the "COVID-19 Patient Management" Business Process", International Journal of Online and Biomedical Engineering (iJOE),2021, volume-17, No.5,2021. https://doi.org/10.3991/ijoe.v17i05.21419
[13] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[14] https://covid19.who.int/table

7 Authors

A.Lakshmanarao is currently working as Associate Professor in Aditya Engineering College, Surampalem. He completed his B.Tech in CSIT and M.Tech in Software Engineering. He is pursuing Ph.D. in Andhra University, Vishakapatnam. His areas of interest are Machine Learning, Cyber Security, Deep Learning. He is a life member of Computer Society of India (CSI).

M.Raja Babu is Currently working as Associate Professor in Aditya Engineering College, Surampalem. He completed his B.Tech in CSE from MVGR College of Engineering, Vizianagaram and M.Tech in Information Technology from NCET, Vijayawada. He is pursuing Ph.D. from JNTUH, Hyderabad. He is having nearly 12 years of teaching experience. His research interest includes Image Processing, Information Retrieval and Pattern Recognition. He is a life member of Computer Society of India (CSI)and Indian Science Congress Association (ISCA). He has published research papers in various National, Inter National conference proceedings and Journals.

T.Srinivasa Ravi Kiran currently working as Assistant Professor & HOD in Department of Computer Science, PB Siddhartha College of Arts & Science, Vijayawada. He completed his Ph.D. in Acharya Nagarjuna University. His areas of interest are machine learning, databases, cyber security. He has published research papers in various conferences and journals.

Article submitted 2021-07-02. Resubmitted 2021-08-07. Final acceptance 2021-08-07. Final version published as submitted by the authors.