Analyzing and Predicting Covid-19 epidemic using Machine Learning Techniques

A.Lakshmanarao, Ch.Seshadri Rao, Sridevi Gadde

1Department of IT, Aditya Engineering College, Surampalem,A.P
2Department of CSE, Raghu Engineering College, Visakhapatnam, A.P, India

Email: laxman1216@gmail.com, seshadirao.chinta@raghuenggcollege.in, Sridevi.gadde@raghuenggcollege.in

Abstract. The entire world is suffering from a novel disease called covid-19 caused by a coronavirus since 2019. The main reason for the seriousness of the disease is the lack of efficient legitimate medication or vaccine. The World Health Organization (WHO) suggested several precautions to regulate the spread of disease and to reduce the contamination thereby reducing deaths. In this paper, we analysed the covid-19 dataset available in Kaggle. The previous contributions from several authors of similar work focused on covid-19 datasets having a limited number of samples. Our paper used the dataset updated till November 15th 2020. Three different aspects are considered mainly in this paper, namely the number of confirmed cases, number of recovered cases, and number of death cases. All the aspects are analysed in a daily and weekly manner. We applied linear regression, polynomial regression, and holt’s method to predict the future number of confirmed, recovered, and death cases. This analysis is useful for the health sectors and frontline workers to help reduce the contamination caused by this disease.

1. Introduction
Covid-19, a disease caused by coronavirus ringing dangerous alarms from 2019 worldwide. The number of cases at the end of January 2020 is 581, out of which 571 from china, but the number of cases worldwide in November 2020 is 58,900,547. The number of cases in India is 9,177,840[1] now. This disease spreads in several ways, but the main cause for spreading is the virus affected patient’s droplets produced through cough or sneezing. If a person is having close contact with a covid-19 patient, then there more chances of getting covid-19. The virus survival time varies from place to place. It depends on the type of surface, humidity, temperature, etc. The recovery period for covid-19 affected people is 14 days. So, if a person is tested as a covid-19 positive case, then that person must be in quarantine for 14 days. As research is going on for the covid-19 vaccine, all governments are suggesting the people take precautions to stop the spreading of the coronavirus. The general symptoms of covid-19 are cold, fever, cough, and shortness of breath. Using a face mask and sanitizers is suggested as a primary measure. Although the Russian government released a vac-cine named ‘Sputnik’, many of the countries are not accepting its complete solution for covid-19. World Health Organization (WHO) declares covid-19 as a pandemic [2]. SARS (Severe Acute Respiratory
Syndrome-2003) are MERS (Middle East Respiratory Syndrome-2012) are also another similar virus disease causes thousands of deaths.

2. Literature Survey

Machine Learning coined by Arthur Samuel is not a new concept. Although it was introduced in the 1950s, the real usage of machine learning boosts from the last decade. Today machine learning is applied to every field and the health industry is not an exclusion. S Dhamodharavadhani [3] et.al proposed Statistical Neural Network Models for COVID-19 Mortality Rate Prediction in India. They conducted different experiments on two different COVID-19 datasets and achieved good RMSE values. M. Rubaiyat Hossain Mondal[4] et.al applied polynomial regression for finding a number of cases. They also proposed classification algorithms Multilayer Perceptron, Logistic Regression, XGBoost for classification of covid-19 patients. LamiaaA. Amar [5] et.al analysed four months of covid-19 data from February 2020 to June 2020 in Egypt. After analysing the dataset with polynomial regression, they concluded that the epidemic of the virus in Egypt would not end in the near future. Lodangi Nagakrishna [6] et.al forecasted covid-19 cases using regression analysis. They have done experiments with 7 months of data in India (Jan 2020 to July 2020) and concluded that there is no chance of reducing cases quickly. Rajan Gupta [8] et.al applied the ARIMA model for predictive analysis of the covid-19 dataset and concluded that the growth rate is high worldwide (until March 2020) but there is a chance of exponential growth. Ekta Gambhir [9] et.al proposed polynomial regression, support vector regression for analysing covid-19 data from January 2020 to June 2020. Kolla Bhanu Prakash [10] proposed various classification algorithms like decision tree classifier, random forest classifier, logistic regression, naive Bayes classifier and concluded that the age group of 20 to 50 are more affected by covid-19. Bathula Preetham Kumar Reddy [11] applied Linear Regression model to predict future assumptions of covid-19 disease. Li Yan[12] et.al proposed XGBoost machine learning based model for the prediction of mortality rates of covid-19 patients and achieved an accuracy of 90%. Behrouz Pirouz[13] et.al applied binary classification techniques to identify the key factors causing covid-19. With their experiments they have shown that humidity and average temperature are having the highest impact on the disease. G.D. Barmparis [15] et.al proposed Gaussian spreading hypothesis for estimating infection horizon of covid-19. With their model, they predicted daily infections successfully.

3. Research Methodology

The covid-19 dataset is downloaded from Kaggle [14]. Later, we analysed data in different perspectives and then we applied regression models. The proposed method is shown in Figure-1.

![Figure 1. Proposed method for Covid-19 Detection](image-url)
The dataset was collected from Kaggle repository. The dataset contains 8 columns namely Date, Province/State, country, last updated date, confirmed cases, recovered cases, death cases. There are missing values in the column ‘Province/State’. We calculated total number of confirmed cases, death cases, recovered cases, active cases and closed cases from the given data. This information is tabulated in Table-1.

**Table 1. DATASET DETAILS**

| Total no. of confirmed | Total no. of recovered | Total no. of death cases | Total no. of active cases | Total no. of closed cases |
|------------------------|------------------------|--------------------------|---------------------------|--------------------------|
| 54370186               | 34955148               | 1317139                  | 18097899                  | 36272287                 |

### 3.1. Distribution plots

From the table-1, the active cases distribution plot and closed cases distribution plots are drawn and they are shown in Figure-2.

![Active cases distribution plot](image)

![Closed cases distribution plot](image)

**Figure 2. Distribution plots of active and closed cases**
3.2. Weekly progress of confirmed, recovered, death cases

Weekly progress of confirmed, recovered, death cases are shown in Figure-3. From figure-3, it is observed that death cases are increasing in a similar manner from week-1 to week-43, but the number of confirmed cases are raising high. Weekly increase in number confirmed and death cases are shown in Figure-4.

From Fig-4, average increase of confirmed, recovered, death cases are observed:
Average increase in number of confirmed cases everyday: 181838.0
Average increase in number of recovered cases everyday 116907.0
Average increase in number of death cases everyday 4405.0
3.3. Daily progress of confirmed, recovered, death cases

Daily progress of covid-19 cases is depicted in Figure-5. From Figure-5, it is observed that the gap between confirmed cases and recovered cases is increased from October 2020.

![Daily progress of cases](image)

**Figure 5.** Daily progress of cases

3.4. Country wise Mortality rate

An important chrematistic of any dangerous disease is mortality rate. It indicates how many deaths happened due to disease. Health authorities uses mortality data for evaluating their priorities/measures for control of disease. The morality data of top 15 countries are shown in Figure-6.

![Top 15 countries](image)

**Figure 6.** Confirmed and death cases for top 15 countries

4. Experimentation and Results

We applied regression algorithms for the prediction of confirmed cases, death cases, and recovered cases. We applied regression techniques in two different ways. In the first method, we considered day number as an independent variable and number of cases (confirmed/recovered/deaths) as a dependant variable. In the second method, we prepared a dataset structure where the previous 20 days’ cases are used for predicting the number of cases on a particular day. In both cases we divided the dataset into training and testing sets in 80%,20% ratio.
4.1. Applying Regression techniques with day number and number of cases

4.1.1. Applying Linear Regression

Linear Regression is a simple technique where dependent feature (y) is estimated from independent feature (x). Here number of cases is dependent feature and time is independent feature. But linear regression cannot find nonlinear relationship between features.

4.1.2. Applying Polynomial Regression

Polynomial regression is used to find nonlinear relationship between features. In polynomial regression, Polynomial features are generated by raising current features to an exponent. If X is a feature, then we raise it to X power 2, X power 3, X power 4 etc...

4.1.3. Applying Holts method

The method is an exponential smoothing method. This applicable to the time series, value, trend, and seasonality data. It forecasts the data with trend.

4.1.4. Results of Linear Regression, Polynomial Regression, Holts method

| Algorithm          | RMSE(Confirmed) | RMSE(Recovered) | RMSE(Deaths) |
|--------------------|-----------------|-----------------|--------------|
| Linear Regression  | 14810771        | 11386638        | 114504       |
| Polynomial Regression | 12965349    | 4420199         | 3728451      |
| Holts method       | 3070855         | 616482          | 32717        |

Table 2. RMSE comparison of regression algorithms

From Table-2, it is easily observed that holt’s method done better estimation, when compared to linear and polynomial regression.

Figure 7. Linear regression, polynomial regression plot for confirmed cases (world-wide)
Figure 8. Linear regression, polynomial regression plot for death cases (world-wide)

Figure 9. Holts method plot for Confirmed, death cases (world-wide)

4.2 Applying Regression techniques with previous days cases as predictor

We created a dataset with previous 15 days cases as independent variables and current cases as dependent variable. Here any particular cases on a day are predicted from its previous 15 days cases. After preparing dataset with this structure, we applied linear regression.

Table 6. Linear Regression technique with previous 10 cases as predictor

|                  | Confirmed cases | Recovered cases | Death cases |
|------------------|-----------------|-----------------|-------------|
|                  | RMSE            | R2              | RMSE        | R2          | RMSE     | R2          |
| Linear Regression| 41465           | 0.99            | 211899      | 0.97        | 1129     | 0.97        |

5. Conclusion

In this paper, we analysed the number of confirmed cases, the number of recovered cases, and the number of death cases on a daily, weekly basis in the covid-19 dataset. We applied regression techniques in two different approaches, the first one used the number of days as a predictor and the second approach used the previous days' cases as a predictor to predict the future number of confirmed, recovered, death cases. Our experimental results on the dataset shown a good R-squared value of 0.99.
References

[1] https://covid19.who.int/table
[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[3] S Dhamodharavadhani, R Rathipriya and Jyotir Moy Chatterjee 2020 COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models, *Frontiers in Public Health* Volume 8, Article 441
[4] M. Rubaiyat Hossain Mondal, Subrato Bharati, Prajoy Podder 2020 Data analytics for novel coronavirus disease, *ELESEVIER- Informatics in Medicine Unlocked* 20 (2020) 100374 https://doi.org/10.1016/j.imu.2020.100374
[5] Lamiaa A. Amar, Ashraf A. Taha, Marwa Y. Mohamed 2020 Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt, *Infectious Disease Modelling, August 2020, ELSEVIER*, https://doi.org/10.1016/j.idm.2020.08.008
[6] Lodangi Nagakrishna, Dr. Vijay Thawani, Kunda Gharpure 2020 COVID-19 Forecasting Using Regression Analysis, *VIMS Health Science Journal* Volume 7 - Issue 3 - September 2020, DOI: 10.46858/vimshsj.7307
[7] Rajan Gupta, Gaurav Pandey, Poonam Chaudhary, Saibal K. Pal 2020 Machine Learning Models for Government to Predict COVID-19 Outbreak, *Digital Government: Research and Practice* Volume 1, No. 4, Article 26
[8] Gupta, R, Pal, S. K 2020 Trend Analysis and Forecasting of COVID-19 outbreak in India, *medRxiv* March-2020, https://doi.org/10.1101/2020.03.26.20044511
[9] Ekta Gambhir, Ritika Jain, Uma Tomer, Alankrit Gupta 2020 Regression Analysis of COVID-19 using Machine Learning Algorithms, *International Conference on Smart Electronics and Communication (ICOSEC)IEEE Explore* pp 65–71
[10] Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail, TPavan Kumar, YVR Naga Pawan 2020 Prediction, Evaluation and Analysis of COVID-19 Datasets using Machine Learning Algorithms, *International Journal of Emerging Trends in Engineering Research* Volume-8(May)
[11] Bathula Preetham Kumar Reddy 2020 Prediction of Covid-19 Using Regression Analysis, *International Research Journal of Engineering and Technology (IRJET)*, Volume: 07 Issue:06
[12] Li Yan, Hai-Tao Zhang, Jorge Goncalves 2020 An interpretable mortality prediction model for COVID-19 patients, *Nature Machine Intelligence* Volume: 02 pp 283-288
[13] Behrouz Pirouz, Sina Shaffiee Haghsheen, Sami Shaffiee Haghsheen, Patrizia Piro 2020 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* Volume 2020, 12, 2427
[14] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[15] G.D. Barmparis, G.P. Tsironis 2020 Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach, *Chaos, Solitons and Fractals* 135 (2020) 109842, 2020 Elsevier