Volume 8 Issue 2 (2020) 119-124 ISSN 2347 - 3258
International Journal of Advance Research and Innovation

Epidemiological Covid-19 Outbreak Prediction and Analysis using Machine Learning

Manit Kakkar, Elisha Agarwal, Shaveta Arora
Department of Computer Science, The NorthCap University, Gurugram, 122001, India

Article Info
Article history:
Received 15 February 2020
Received in revised form
29 May 2020
Accepted 12 June 2020
Available online 15 June 2020

Keywords: Severe acute respiratory syndrome, coronaviruses analysis, machine learning, regression

Abstract: In December 2019, coronavirus disease (Covid-19) showed in the seafood marketplace in Wuhan, China for which no treatment has been found till date. Since this disease is emerging widely in almost all parts of the world and no medication is found so far therefore research is needed to illuminate the study of coronaviruses.

In this work, we have proposed various machine learning methods to analyse the pandemic coronavirus recovery rate and death rate in the real world. For this, polynomial regression, Decision tree regressor, Random Forest regressor and Long-short term memory (LSTM) model are utilized. The parameters are estimated, and predictions are made based on real-time dataset. Our proposed models tend to achieve high accuracy in prediction which will help to improve in-depth investigation

1. Introduction

Coronavirus was first discovered in 1930 which forced human beings to be isolated being contagion. Slowly, this became an outbreak of severe respiratory syndrome in late 2002 [1]. There were four groups of coronavirus which exist: (HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1). SARS-CoV is an outlier to these four groups. In humans, it was found that Covid-19 is included in the spectrum of viruses that cause common cold, cough, mild fever, severe acute respiratory syndrome (SARS) [2]. Some varieties of the virus also cause severe diseases including kidney failure, heart disorder, reproductive diseases, hepatitis. Since Coronavirus was spotted in mainland China on 12th December 2019, it has made its way infecting 26, 00,000 people around the globe with casualties of nearly 2, 00,000 people worldwide. Nearly, more than 170 countries are infected. Countries like Italy, Germany, USA, and Spain are some which were affected at a very high pace due to underestimating and inability to predict the effects in the country. Till date, no country can find an actual solution to this outbreak except that it is well known that the disease spreads through the physical transmission. Therefore, governments of many countries have practiced complete lockdown to stop the spread of the growth rate. Symptoms of COVID-19 are non-specific and the disease presentation can range from no symptoms (asymptomatic) to severe pneumonia and death [3–5].

The government of India has set guidelines which says that during this summer the temperature settings for rooms should be between 24 degree Celsius to 30 degree Celsius and moderate humidity should be maintained, proper hygiene of air-conditioners should be maintained as some sources state that Coronavirus transmission significantly reduces at high temperature and humidity level [6].

Covid-19 has come up as a big challenge to the medical as well as the investigation field. Technologies like Artificial Intelligence (AI) are proven to be successful to depict the outbreak consequences and impacts of such pandemic outbreaks in the past for example, in pneumonia and skin cancer. The government of various countries has used artificial intelligence such as South Korea is using mobile location, CCTV, facial scans, temperature monitor to trace people who are violating lockdowns, UAE has launched a drive-thru based Coronavirus testing facility [7]. According to an article in the Economic times which talks about two Indian origin researchers that led to the development of Covid-19 specific risk checker app to help people to keep away from confusion, fear and rumors about the infection [8]. Medical Chatbots were updated and trained to screen people and spread advice on whether they need to be quarantined. Artificial Intelligence was used to allow individuals to fill questionnaires as self-assessment [9]. In India, an AI-based application for self-assessment tests was introduced by the Apollo group of hospitality. Technology laid an immense role to spread awareness and make people conscious about how to save themselves from this deadly disease. Arogya-Setu app was also created to track the location and frequency of infected people around us. This artificial intelligence helps the government to identify hotspots and containment zones forming clusters using clustering in AI and using AI-driven Drone technology to keep a watch on the people breaking the lockdown law.

Therefore, this Covid-19 outbreak motivated us to utilize advanced AI techniques to predict future outcomes in preventing the spread of such diseases [10]. Datasets have been explored from Kaggle, GitHub and Ourworldindat.org which is the primary dataset for Coronavirus records worldwide [11]. Also, Study was done from online websites like National Health Commission of China, National Centre of Disease Prevention and Control and WHO and a related scientific study published between January and May 2020.

Our research depicts the Prediction of the rate of death, confirmation, and recovery of the infected people. The next section of our research is an outcome (Cross-validation score, Mean squared error, Mean absolute error, Standard Deviation, Z-score, T-score, R2-score, Accuracy, Precision, Recall, F1-score) that can be used further to study in this field. However, there are certain challenges faced by the researchers and the data scientists such as the uncertainty of the dataset. Our proposed models also depict error rates which will help to reduce uncertainty in the dataset. The machine learning algorithms demand a huge amount of data, but the dataset is not too big, and it is changing every hour.

2. Methodology of the proposed system

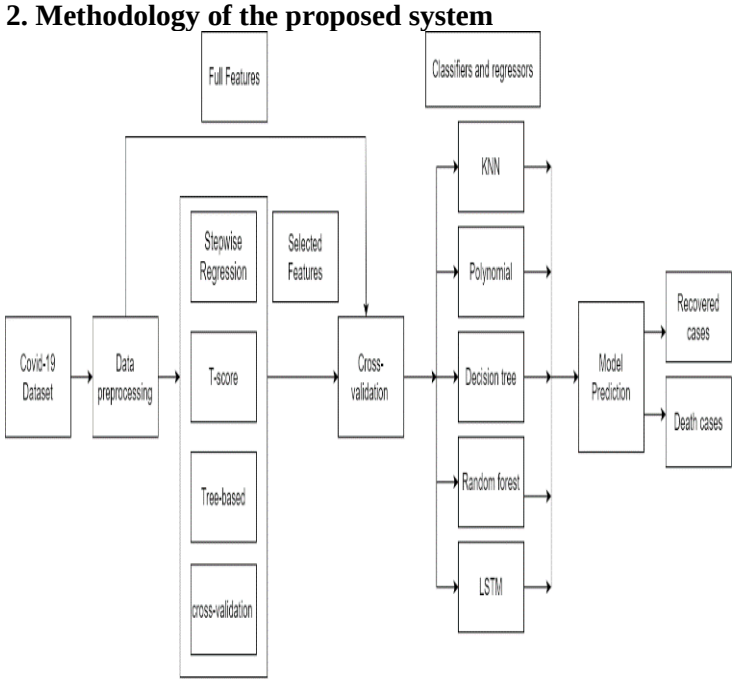


Fig. 1: The working of the prediction model through a Flow Chart

***Corresponding Author,**
E-mail address: shavetaarora@ncuindia.edu

The proposed work is represented with the help of a flowchart (Fig.1) to show all the activities, procedures and stepwise study done during the preparation of the model. We have started by selecting the most latest dataset, after which data preprocessing was done to get accurate results and better model fitting. Our data went through various feature selection procedures which resulted in cross-validation to be the best algorithm for feature selection criteria. Further various regression and classification models are applied for prediction.

3. Data Visualization

Data visualization is a pictorial or graphical representation of the results and facts. It is believed that visual understanding always helps the learner to learn better and remember things for a longer time. Therefore, we have put all outcomes in figures and graphs.

In Fig.2 the red dots on the world map represents the areas affected by the pandemic. With each passing day, the number of cases are increasing. In Fig.3 a line graph shows the death rate across the globe, china having the maximum number of deaths in April 2020.

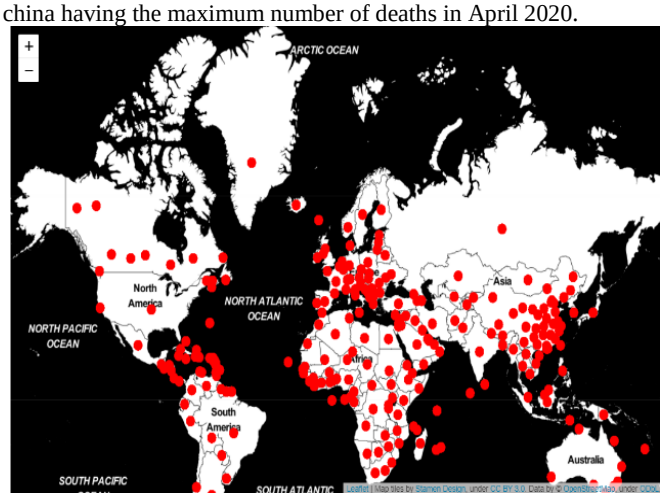


Fig. 2: Active Covid-19 cases across the world through World Map visualization

Fig. 3: Source: FT analysis of the European Centre for Disease Prevention and Control; COVID Tracking Project; FT research.]

Figure 4, gives a summary of the Confirm Vs Recovery Vs Death using a bar plot where blue colour depicts the confirmed corona cases, orange colour represents the recovery cases and yellow colour shows the deaths. Fig.5 describes the individual confirms, death and recovery numbers of all the three cases against the date.

4. Fitting and analysis of Covid-19 using Mathematical models

Data Scientists from different countries have used many mathematical models to perform analysis and prediction using Artificial Intelligence. Some majorly used models include Polynomial regression model, Exponential model, LSTM model, O-SEIHRD model. We have used a very well-known library Scikit-

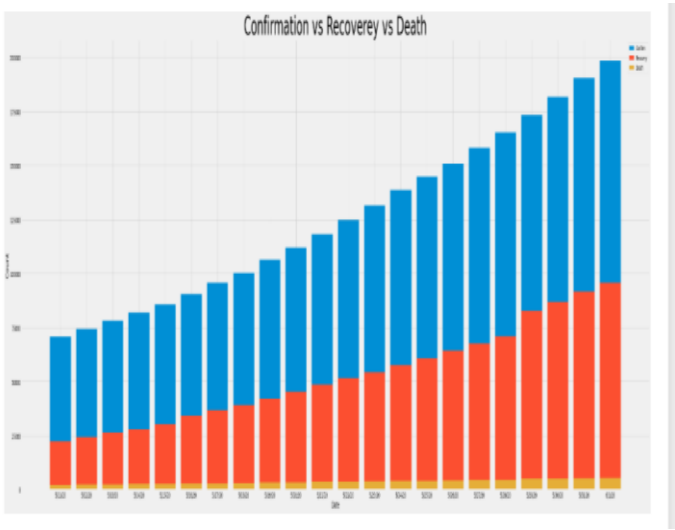


Fig. 4: Visualization between Confirm Vs Recovery Vs Death

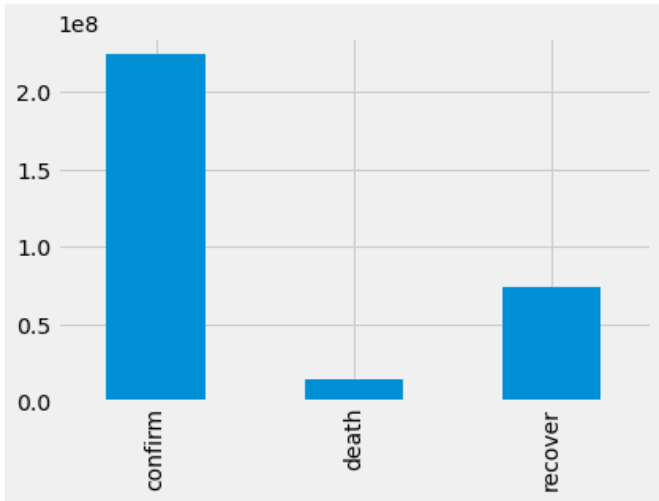


Fig. 5: Visualization between Confirm Vs Recovery Vs Death

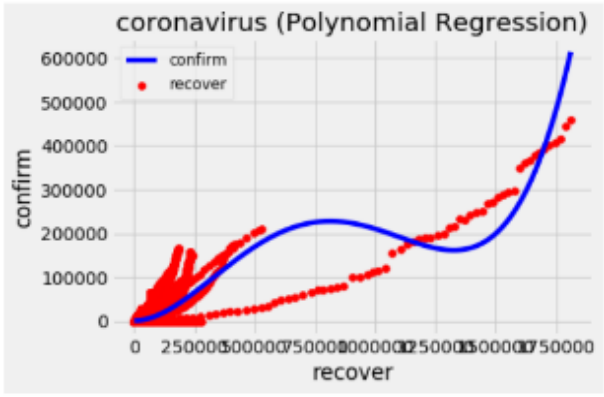
learn, the polynomial regression model, decision tree regressor model, random forest regressor model, LSTM model, KNN classification and TensorFlow at the backend. The tests are purely performed using Python language with the Pandas library. The dataset is taken from Kaggle and the information lies between 22-01-2020 to 31-05-2020. Proper pre-processing and data cleaning have been performed to get higher accuracy and precision

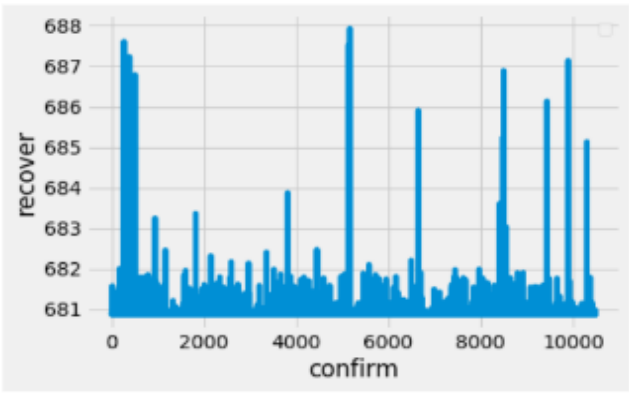
4.1 Polynomial Regression Model

Polynomial Regression falls under the category of linear regression where the relationship between the dependent and the independent variable is modelled as an nth degree polynomial. The relationships between the variables may be non-linear.

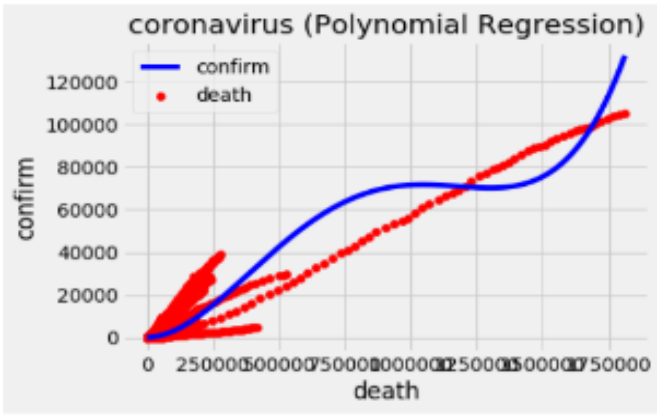
$$y = a + b_1 x + b_2 x^2 + \ldots + b_n x^n \quad (1)$$

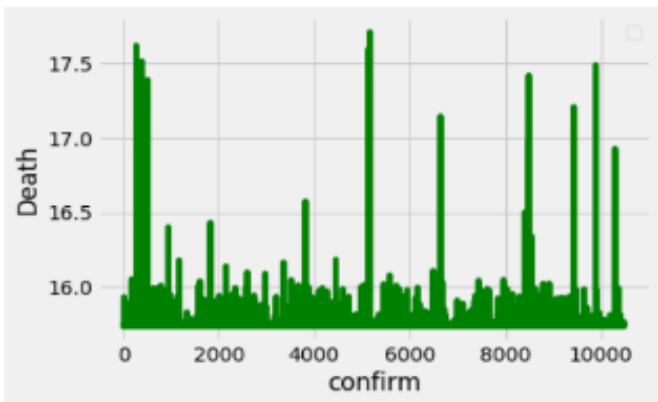
Where, x is the independent variable, y is the dependent variable. Fig 6 (a, b) shows the result from the polynomial regression model. Fig 6(a) portrays the Confirm cases in blue colour v/s Recovered cases in red colour from 22-01-2020 to 31-05-2020, similarly, Fig 6(b) reveals Confirm in blue colour v/s Death cases in red colour from 22-01-2020 to 31-05-2020.





(a) Confirm v/s Recovered cases from 22-01-2020 to 31-05-2020





(b) Confirm v/s Death cases from 22-01-2020 to 31-05-2020

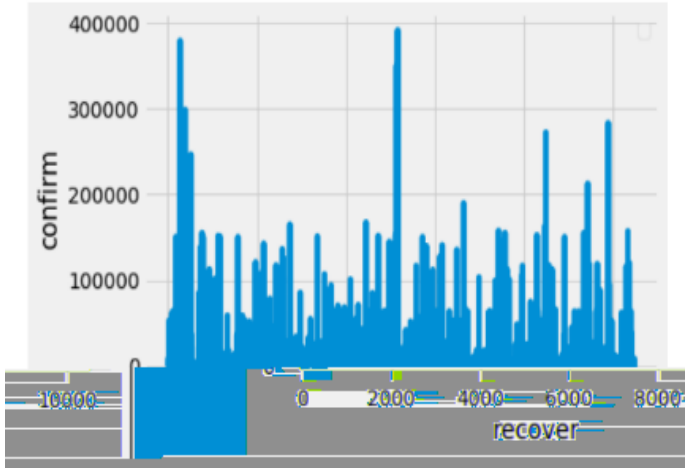
Fig. 6: Prediction using Polynomial Regression

4.2 Decision Tree Regressor

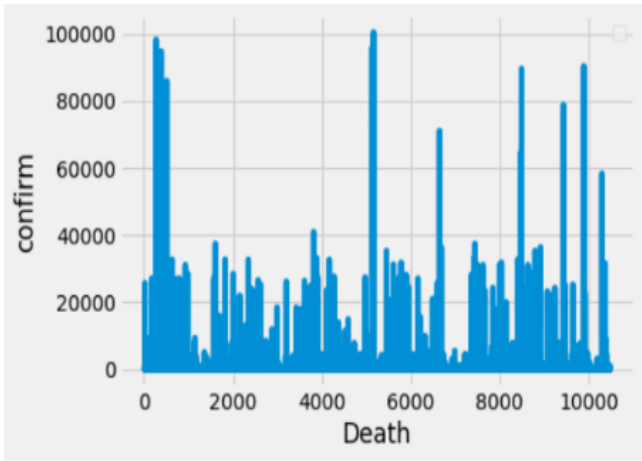
Decision tree regressor model used to form a tree structure of our dataset which breaks down further into smaller datasets. While at the same time a decision tree associated with our increasingly developed dataset. As a result, we get a final decision tree with decision nodes and leaf nodes. Our prediction mainly focusses on the death rate and the recovery rate from the confirmed cases which is calculated using equation (2) and (3) respectively.

$$\text{Death rate} = \frac{\text{number of confirmed deaths}}{\text{total number of confirmed cases}} \text{ X } 100 \quad (2)$$

$$\text{Recovery rate} = \frac{\text{number of recovered cases}}{\text{total number of confirmed cases}} \text{ X } 100 \quad (3)$$



(a)



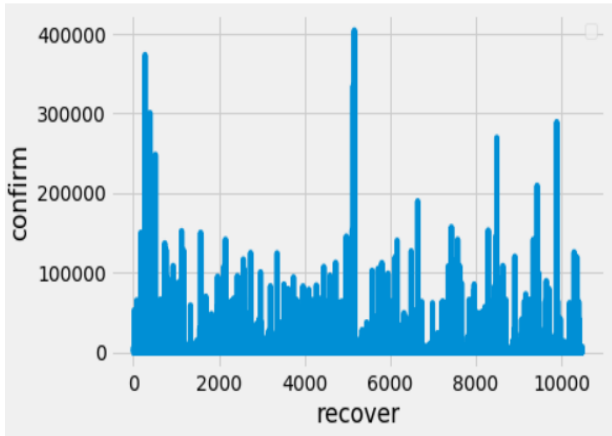
(b)

Fig. 7: Prediction using Decision tree regressor (a) Confirm v/s recovered cases from 22-01-2020 to 31-05-2020 (b) Confirm v/s death cases from 22-01-2020 to 31-05-2020

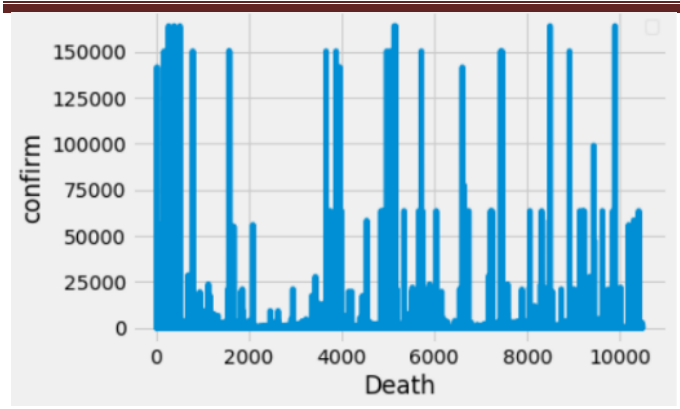
Fig 7 (a,b) describes predictions of the Decision tree regressor where Fig.7(a) depicts Confirm v/s Recovered cases from 22-01-2020 to 31-05-2020 and Fig.7(b) depicts Confirm v/s death cases from 22-01-2020 to 31-05-2020.

4.3 Random Forest Regressor

Random forest regressor is used for estimation and fits a few classifying decision trees which help in improving the accuracy and control overfitting. It takes into consideration the attributes mentioned in the dataset.



(a)



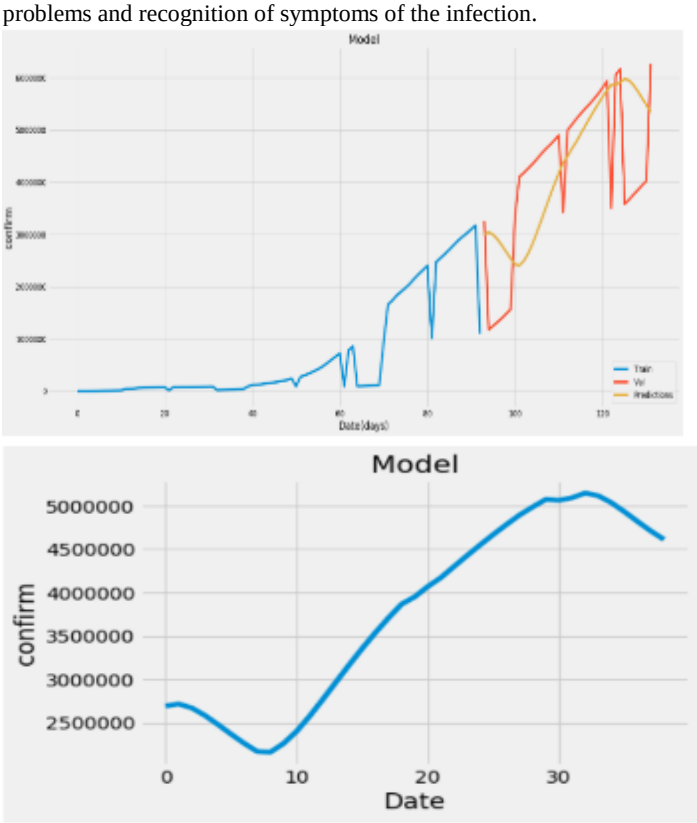
(b)

Fig. 8: Prediction using Random Forest Regressor (a) Confirm v/s Recovered cases from 22-01-2020 to 31-05-2020 (b) Confirm v/s Death cases from 22-01-2020 to 31-05-2020

Fig 8(a,b) describes predictions of the random forest regressor where Fig.8(a) depicts Confirm v/s Recovered cases from 22-01-2020 to 31-05-2020 and Fig.8(b) depicts Confirm v/s death cases from 22-01-2020 to 31-05-2020.

4.4 LSTM Model

LSTM stands for Long Short-Term Memory network that belongs to a complex area of Artificial Intelligence. It is a part of deep learning. We have used LSTM in our research for sequence prediction problems and recognition of symptoms of the infection.



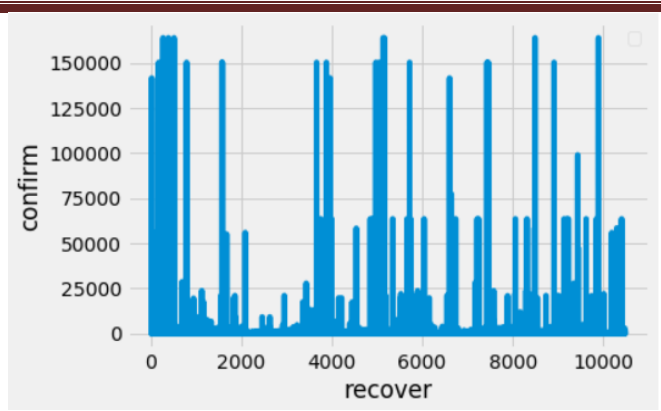
(a) (b)

Fig. 9 Prediction using LSTM Model (a) Date vs Confirmed cases graph using LSTM Model

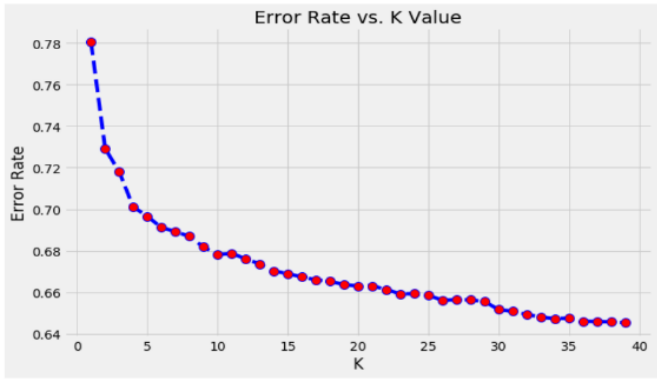
Fig 9 summarizes the predictions fetched by the LSTM Model. Where Fig 9(a) describes the relationship between data provided to the model. Blue colour depicts the trained data, orange colour represents the test data and yellow colour shows the predicted values. Fig. 9(b) displays the confirm cases against the date.

4.4 KNN Classification model

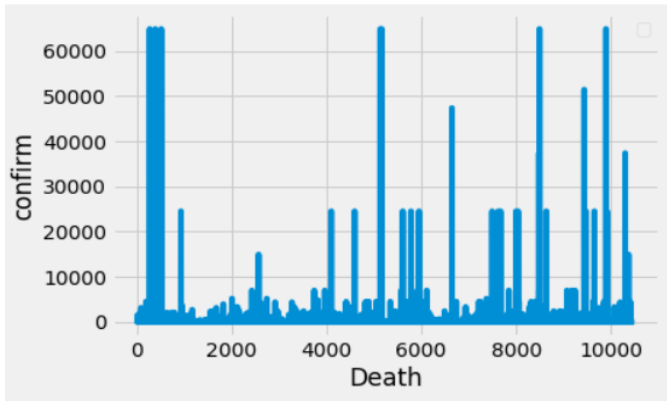
It is one of the simplest algorithms that collects all available cases and further classifies them based on similarity criteria. Pattern recognition and statistics can also be done using KNN.



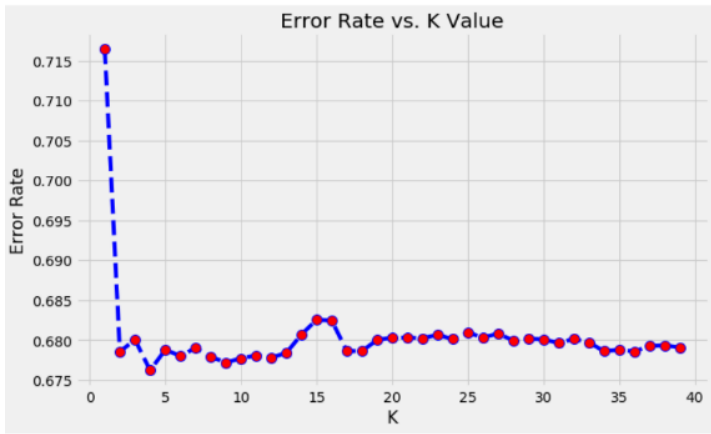
(a) Confirm v/s Recovered cases from 22-01-2020 to 31-05-2020



(b) Error rate vs K value for Confirm v/s Recovered cases



(c) Confirm v/s Death cases from 22-01-2020 to 31-05-2020



(d) Error rate vs K value for Confirm v/s Death cases

Fig. 10 Prediction using KNN Classification

Fig.10 (a,b,c,d) represents predictions of KNN Classification model where Fig 10(a) portrays the Confirm cases in blue colour v/s Recovered cases in red colour from 22-01-2020 to 31-05-2020, similarly, Fig 10(c) reveals Confirm in blue colour v/s Death cases in red colour from 22-01-2020 to 31-05-2020. Throwing light on the Fig.10 (b) draws the comparison between Error rate vs K value for Confirm v/s Recovered cases, similar to this Fig.10(d) draws the

comparison between Error rate vs K value for Confirm v/s death cases.

5. Model Evaluation Statistical Parameters

After fitting the values from the data set used, here we have calculated various statistical parameters with various regression models.

1. Accuracy tell immediately whether our model is being trained correctly which is calculated using Confusion Matrix. The actual vs. predicted classification can be framed which is called as a confusion matrix. Confusion matrix has some outputs like True positives, True negatives, False positive, False Negative that helps in calculation of F1-score, accuracy, precision, recall etc.

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{Total} \quad (4)$$

2.The mean square error (MSE) depicts the average squared difference between the estimated values(yi) and predicted values ($y`i$).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - y`i)^2 \quad (5)$$

3.The mean absolute error (MAE) is the mean/average of all the errors.

$$4.MAE = \frac{1}{n}\sum_{i=1}^{n}|xi - x| \quad (6)$$

4. Standard deviation (SD) is the square root of the variance where n in (n-1) stands for the total number of samples.

$$SD = \sqrt{\frac{\sum(x-x`)^2}{n-1}} \quad (7)$$

5. R-squared or R2-Score is a statistical measure of how close the data are to the fitted regression line. The value lies between 0 and 100%.

6.Cross-validation is done wherein the modelling process runs on different subsets of the data to get multiple measures of model quality.

7. T-score- T-score and Z-score are one of a kind, but the major difference is that T-score does not have many data points because the normal distribution of a few data points cannot be expected. It also helps in comparison of mean values of different populations for determining the similarity between them.

8. Precision- can be drawn from the confusion matrix. It is a fraction between true positive rate and true positive rate plus false positive rate. It is also called as PPV (Positive Predicted value).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

9. Recall- Recall can be drawn from the confusion matrix. It is a fraction between true positive rate and true positive rate plus false negative rate**.** It is also called as Sensitivity rate.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (9)$$

10. Z-score- The Z-score also known as the standard score is a representation of standard deviations in fractional form from the mean value of the population.

$$Z = \frac{X-\mu}{\sigma} \quad (10)$$

11. F1-score- The F1 score portrays the balance between the recall and the precision.

$$Z = \frac{\text{precision}*\text{recall}}{\text{precision+recall}} \quad (11)$$

6. Results and Discussions

In this section, we have done a comparison of various model's outcomes into numeric values. These results are the same from which all the visuals presented in the document were derived.

Table 1 and Table 2 contains the outcomes observed using the polynomial regression model wherein Table1 we have the scenario of Confirm Vs Recover cases and in Table 2 Confirm Vs Death cases. Regression model gives R2 score as an important output which helps in concluding the best model based on accuracy. Higher the R2 Score of a model higher the accuracy of the model. From this, we can say Table 2 has good outcomes.

Table 1: Results derived using Polynomial Regression for Confirm Vs Recover cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2 Score
64362203.31	1798.04	9130.16	8035.26	953.31	69.92 %

Table 2: Results derived using Polynomial Regression for Confirm Vs Death cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2-Score
1802427.13	219.74	1411.59	1293.1	137.66	86.68 %

Table 3 and Table 4 contains the outcomes observed using the Decision tree Regressor model wherein Table 3 we have the scenario of Confirm Vs Recover cases and in Table 4 Confirm Vs Death cases. Decision tree Regressor stands out from other models as it has a high R2 score.

Table 3: Results derived using Decision tree Regressor for Confirm Vs Recover cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2-Score
57547413.93	1136.49	9377.55	8431.94	978.05	73.12%

Table 4: Results derived using Decision tree Regressor for Confirm Vs Death cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2-Score
2022485.19	191.70	1392.35	1437.08	136.31	85.58%

Table 5 and Table 6 contains the outcomes observed using the Random Forest Regressor model wherein Table 5, we have the scenario of Confirm Vs Recover cases and in Table 6 Confirm Vs Death cases. It can be noted that Random Forest Regressor also has results like the Decision tree Regressor model.

Table 5: Results derived using Random forest Regressor for Confirm Vs Recover cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2-Score
41512542.29	1033.49	8055.46	7063.92	916.73	80.61%

Table 6: Results derived using Random Forest Regressor for Confirm Vs Death cases

Mean Squared Error	Mean absolute error	Cross-Validation score	Mean value	Standard Deviation	R2-Score
1591413.89	176.31	1225.76	1211.98	111.22	88.65%

Table.7 and Table 8 contains the outcomes observed using the KNN Classification model wherein Table 7 we have the scenario of Confirm Vs Recover cases and in Table 8 Confirm Vs Death cases.KNN is a classification model therefore it has accuracy measure to conclude the best model. In our case, the KNN classification model has given a low accuracy rate.

Table 7: Results derived using KNN Classification for Confirm Vs Recover cases

Precision	Recall	Cross-Validation score	Mean value	Standard Deviation	F1-Score	Accuracy
0.62	0.95	15322.05	11768.63	2222.24	0.75	50%

Table 8: Results derived using KNN Classification for Confirm Vs Death cases

Precision	Recall	Cross-Validation score	Mean value	Standard Deviation	F1-Score	Accuracy
0.86	0.88	2135.75	2297.97	257.87	0.87	54%

Table. 9 contains the results of the LSTM Model, drawing a comparison between confirm v/s Recovery and confirm v/s Death simultaneously.

Table 9: Comparison between different cases using LSTM Model

Confirm vs Recovery Cases		Confirm vs death cases	
Mean Squared Error	Mean absolute error	Mean Squared Error	Mean absolute error
1460913.26	1102126.17	1487768.22	1160331.19

Table 10 consists of a comparison of T-score and Z-Score for Confirm v/s Recovery and confirm v/s Death simultaneously.

Table 10: Comparison of T-score and Z-Score for Confirm vs Recover and Confirm vs Death cases

	T-Score	Z-score
Confirm Vs Recover cases	17951.18, 22154.51	17951.18, 22154.51
Confirm Vs Death cases	4531.31, 5858.84	4531.31, 5858.84

7. Conclusions

Among the models we have used Polynomial Regression and Random Forest Regressor has proven to be the best for fitting the data whereas Decision tree regressor are better for making predictions and have the least error rate, Whereas the LSTM model for predicting the country-specific risk of the novel coronavirus (COVID-19) and the KNN Classification predicted a very lower accuracy.

It is believed that Artificial Intelligence could be a powerful tool to fight against the pandemic. In our research, we have concluded that the cost of the pandemic in terms of human life and economic damage can be high. We also believe that Covid-19 is going to go a long way causing a high fatality rate according to our research. Our research can provide us with early warnings and alerts, tracking, and prediction.

References

[1] SR Weiss, S Navas-Martin. Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus, Microbiol. Mol. Biol. Rev., 69(4), 2005, 635–664.

[2] SA Rasool, BC Fielding. Understanding Human Coronavirus HCoV-NL63~!2009-11-13~!2010-04-09~!2010-05-25~!, Open Virol. J., 4(1), 2010, 76–84.

[3] Bruce Aylward (WHO). Wannian Liang (PRC), Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), 2020. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

[4] P Chatterjee et al. The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidence, Indian J. Med. Res., 2020, doi: 10.4103/ijmr.IJMR_519_20.

[5] SP Adhikari et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review, Infect. Dis. Poverty, 9(1), 2020, 29, doi: 10.1186/s40249-020-00646-x.

[6] J Wang, K Tang, K Feng, W Lv. High Temperature and High Humidity Reduce the Transmission of COVID-19, SSRN Electron. J., 2020, doi: 10.2139/ssrn.3551767.

[7] R Kitchin. Civil liberties or public health, or civil liberties and public health? Using surveillance technologies to tackle the spread of COVID-19, Sp. Polity, 6, 2020, 1–20, doi: 10.1080/13562576.2020.1770587.

[8] Covid-19: Apps, artificial intelligence to help tackle scare, The Economic Times, 3, 2020.

[9] A Ahaskar. How WhatsApp chatbots are helping in the fight against Covid-19, 3, 2020.

[10] S Agrebi, A Larbi. Use of artificial intelligence in infectious diseases, in Artificial Intelligence in Precision Health, 2020, 415–438.

[11] EOO, JH Max Roser, H Ritchie. Coronavirus Pandemic (COVID-19). Our World In Data. org, https://ourworldindata.org/coronaviru.