International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

COVID-19 Time-Series Prediction

Dr. Matcha Venu Gopal Rao[1], Uppala Lakshmi Sai[2], Arigela Durga Sahithi[3]

[1]*Professor, Department of Electronics and Communications Engineering, K L University Vaddeswaram,* Andhra Pradesh, India

[2,3]*Student, Department of Electronics and Communications Engineering, K L University Vaddeswaram, Andhra Pradesh, India*

Abstract - *Corona virus disease known as COVID-19 is a disease caused by a novel virus called SARS-CoV-2. The virus spread rapidly across the world and several large size clusters of the spread have been observed worldwide. An essential part of minimizing the spread of the virus is to monitor, track, and estimate the outbreak. This is extremely useful for decision making against the public health crises. Predicting the spread of COVID-19 is a challenge that the world is facing now. The difficulty in predicting the COVID-19 is lack of availability of COVID-19 dataset. To predict the spread of COVID-19 we made use of the Machine learning techniques like Machine learning Prediction and Forecasting with Time Series Analysis with the available COVID-19 data set. Here in this project we are predicting the COVID-19 confirmed, death and recovery cases of different countries based on the available data set and also made a comparision study of actual and predicted covid-19 cases of different states in India using machine learning algorithms. In this project we used Linear regression and Support vector regression Random forest methods to predict the spread of COVID-19 .The output is in the form of time series where one can able to know the spread of COVID-19 cases in different countries across the world and also graphs between the actual and prediction.*

Key Words: **COVID-19, SARS-CoV-2, Time Series, Linear Regression, Random forest, Support Vector Regression.**

1. INTRODUCTION

COVID-19 Time Series Prediction is all about predicting the spread and containment of COVID-19. Coronavirus disease is a transmissible disease caused by severe acute respiratory syndrome. Now a days forecasting the spread of COVID-19 is a challenge of utmost importance. Symptoms of COVID-19 are different from person to person, but often include fever, cough, tiredness, drowsiness, breathing difficulties, and loss of smell and taste. Symptoms begin one to 14 days after exposure to the virus. The virus seems to be transmitted mostly through the minute respiratory droplets via coughing, sneezing or when people interact with one another for a few time in close proximity.

These droplets can then be inhaled, or they will land on surfaces that others may inherit touch with, who can then get contaminate once they contact their eyes, mouth, or nose. The rapid spread of coronavirus disease (COVID-19) has had harmful effects globally. The virus first started to grow significantly in China and then in South Korea around January of 2020, and then had a major spread of COVID-19 in European countries within subsequent month, and in April the US alone has over 4 lakh cases with over 12,000 deaths. It has since spread worldwide, resulting in an on going pandemic.

The approach we use to this problem is Linear/rectilinear regression model, Random Forest and Support vector regression models. So, using these methods we can know the spread of infectious diseases like covid-19 and also other parameters like deaths, active cases, recovery can be known. In this task, the Support Vector Regression (SVR) model can be used to solve the various sorts of COVID-19 related issues. The proposed method will be fitted into the dataset containing the total number of COVID19 positive cases, and the number of recoveries and other parameters for different countries. These tasks can help a country/region to know the spread of the virus, facilitate/aware people, start mitigations. It'll also help that region/country to be prepared for what will happen within the future, which can help in saving lives.

Random decision forests correct for decision trees practice of overfitting to their training set. Random forests mostly best performed decision trees. The numerical model like linear regression models hooked in to various factors and investigations are dependent upon potential inclination. Here, we presented a model that would be useful to predict the spread of COVID-2019, We have performed linear regression, Random Forest and support vector regression models. Forecasting the potential patterns of COVID-19 effects in India and worldwide hooked in to data gathered from respective datasets. With the available usual data about confirmed, death and recovered cases world-wide for over the time period of length helps us in anticipating and estimating the distant future.

2. Literature Survey

[1] As the availability of COVID-19 data is very less it is difficult to predict. Dictionary learning is one of the algorithms which requires less data to predict the future and the effect of COVID-19. Dictionary learning and Online NMF are the techniques used to predict the COVID-19 confirmed, death and recovered cases based on the available input data

set. Here a method is proposed that uses dictionary learning to predict time series data. It is then applied to analyze and predict the new daily reports of COVID-19 cases in multiple countries.

[2] Linear regression, Multilayer perceptron and Vector auto regression are used for the prediction of COVID-19 cases. The data is collected from Kaggle and the regression models are applied to predict the future. Here the forecasting is done only on the COVID19 cases in India. By seeing the predicted values and matching with cases from John Hopkins University11 data we can conclude that the MLP method is giving good 15 prediction results than that of the LR and VAR method using WEKA and Orange. The correctness of the model can be increased by introducing some more attributes

3. Methodology

The dataset of covid-19 had been collected. The code is written in a way that the total of the covid-19 confirmed, death and recovered cases has been calculated for different countries. And finally the models linear regression, Random Forest and support vector regression are applied to predict the future.

3.1 Linear regression

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Y is the predicted value θ_0 is the bias term. $\theta_1,...,\theta_n$ are the model parameters $x_1, x_2,...,x_n$ are the feature values.

The above hypothesis can also be represented by

$$Y = \theta^T x$$

where

θ is the model's parameter vector including the bias term θ_0

x is the feature vector with x_0 =1

Mathematically, we can write a linear regression equation as:

Y=ax+b

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

3.2 Support vector Regression:

Support Vector Regression is similar to Linear Regression in that the equation of the line is y= wx+b In SVR, this straight line is referred to as hyper plane. The data points on either side of the hyper plane that are closest to the hyper plane are called Support Vectors which is used to plot the boundary line.

Linear and non-linear regression in svr- Linear case:

Y=wx+b

In this equation, w is the weight vector that you want to minimize, X is the data that you're trying to classify, and b is the linear coefficient estimated from the training data. This equation defines the decision boundary that the SVM returns

The magnitude of the normal vector to the surface that is being approximated: min ½||x||^2.

3.3 Random Forest:

Random forests also called as random decision forests are a neural network method for classification, regression and other chores that operate by constructing a multitude

of decision trees at training time and outputting the category that's the way/manner of the classes(nothing but classification) or average prediction (i.e. regression) of the individual trees.

Since, forest averages the predictions of a set of m trees with individual weight functions , so its forecasting given below

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x')\, y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(x_i, x')\right) y_i.$$

4. RESULTS AND DISCUSSION

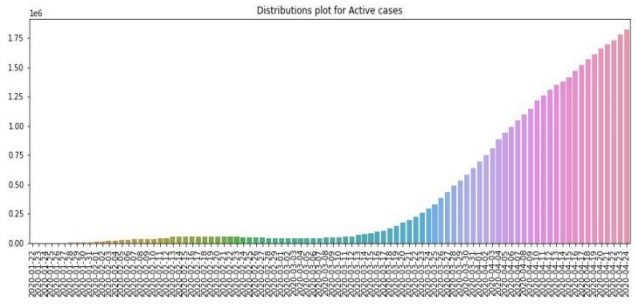

Fig-1 Distribution plot for active cases

The above graph is the analysis of COVID-19 active cases for a period of time for nearly about 4 months.

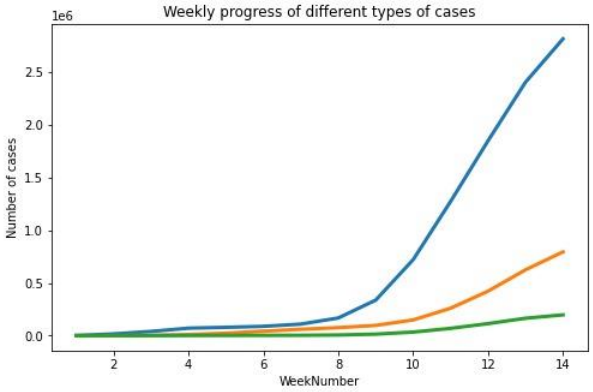

Fig-2 Weekly progress of different types of cases

A graph is drawn for week number V/s Number of cases. The blue line represents the confirmed cases, the orange line represents the recovered cases and the green line represents the death cases.

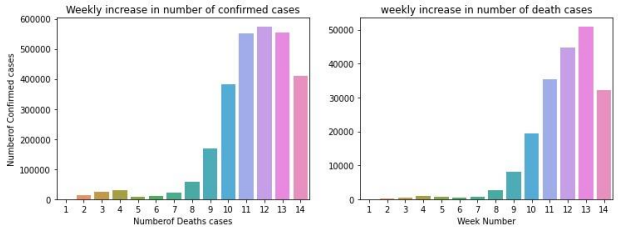

Fig-3 Weekly increase in number of confirmed, death cases

The above bar graphs represent the weekly increase in the number of confirmed cases and weekly increase in the number of death cases for a period of 14 weeks.

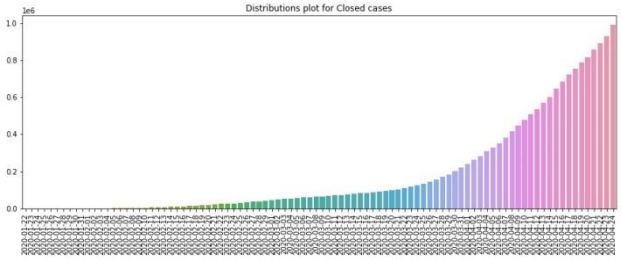

Fig-4 Distribution plot for closed cases

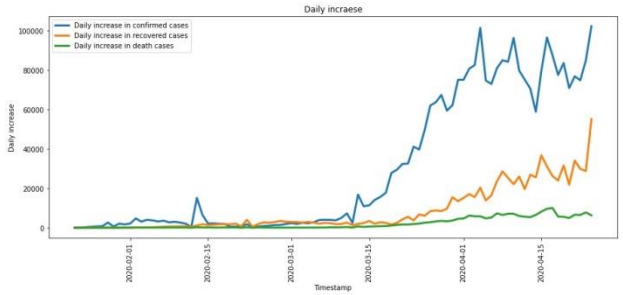

Fig-5 Daily increase in cases

The above graph represents the daily increase in confirmed, recovered and death cases.

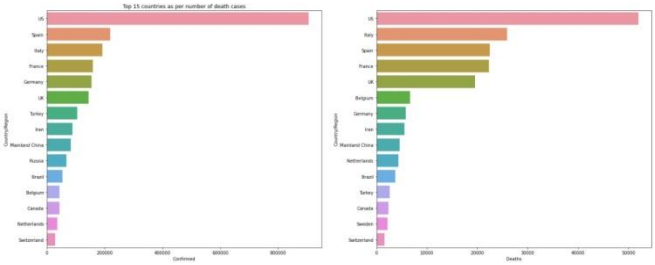

Fig-6 Top 15 countries as per confirmed, death cases

The above bar graph represents the top 15 countries of COVID-19 confirmed and death cases.

	Dates	LR	SVR
0	2020-04-25	1560529	3322586
1	2020-04-26	1582219	3500761
2	2020-04-27	1603909	3686599
3	2020-04-28	1625599	3880344
4	2020-04-29	1647289	4082245
5	2020-04-30	1668980	4292557
6	2020-05-01	1690670	4511540
7	2020-05-02	1712360	4739461
8	2020-05-03	1734050	4976588
9	2020-05-04	1755740	5223200

Fig-7 Future forecast using LR, SVR models

The forecasting of next 10 days is predicted using linear regression and support vector regression.

	Dates	LR	SVR	Holts Linear Model Prediction
0	2020-04-25	1560529	3322586	2859707
1	2020-04-26	1582219	3500761	2938730
2	2020-04-27	1603909	3686599	3017752
3	2020-04-28	1625599	3880344	3096774
4	2020-04-29	1647289	4082245	3175797
5	2020-04-30	1668980	4292557	3254819
6	2020-05-01	1690670	4511540	3333841
7	2020-05-02	1712360	4739461	3412864
8	2020-05-03	1734050	4976588	3491886
9	2020-05-04	1755740	5223200	3570908

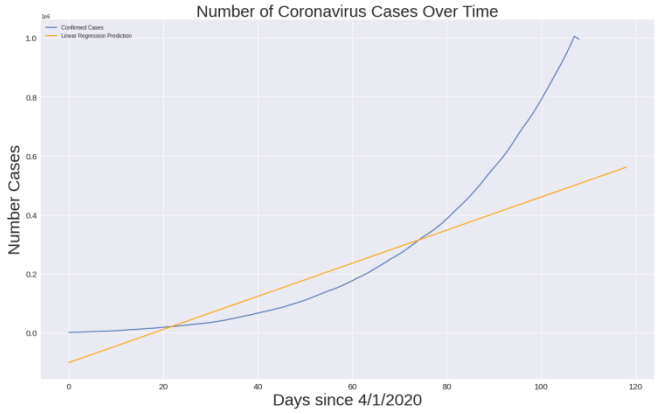

Fig-8 Actual V/S prediction using LR regression model

The above graph is taken for days v/s number of cases for India. The graph is drawn for confirmed cases and prediction. The blue line represents the confirmed cases over a period of time. The yellow line represents the prediction of confirmed cases using linear regression model. From this graph we can say that confirmed and predicted vary with a lot of difference. We observe that the cases are increasing but using linear regression the cases are less than the actual.

This indicate that linear regression is not best suitable regression algorithm for prediction.

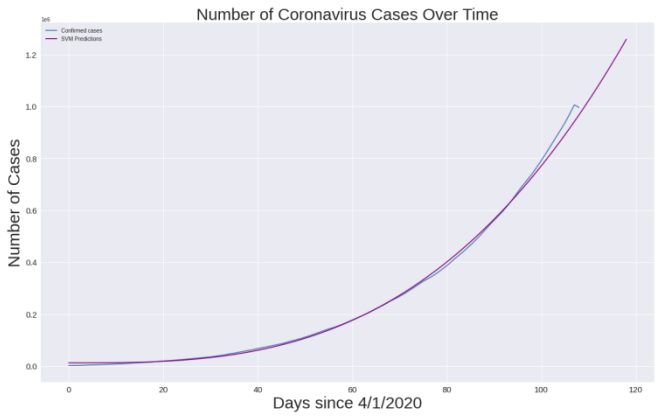

Fig-9 Actual V/S prediction using SVM regression model

The above graph is taken for days v/s number of cases for India. The graph is drawn for confirmed and prediction. The blue line represent the confirmed cases over a period of time. The purple line represent the prediction of confirmed cases using svm prediction. From this graph we can say that confirmed and predicted are almost similar with less error which proves that svm best fits for prediction.

5. CONCLUSIONS

With the quickly changing circumstance including COVID-19, it is basic to have exact and powerful techniques for foreseeing present movement. From this project we observe that Support vector regression gives more accurate results regarding the prediction than linear regression. Forecasting is a method that uses historical data as inputs and estimates its effect in the future. In our project we proposed an approach to predict time-series data. We then applied this approach to predict the new daily reports of COVID-19 cases in multiple countries. The COVID-19 confirmed, death and recovered cases of different countries are forecasted with the available data set using machine learning algorithms. The algorithms used in predicting the COVID-19 cases are linear regression, Random Forest and support vector regression. These are the basic models used for forecasting. In future it very well may be applied to anticipate the spread of infection as well as other related boundaries. These include medical and food supply shortages and demands, subgroup infections and immunity and many more.

REFERENCES

1.Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. End to end speech recognition in english and mandarin. 2016.

2. Yoshua Bengio. Deep learning of representations: Looking forward. In International Conference on Statistical Language andSpeech Processing, pages 1–37. Springer, 2013.

3. Michael W Berry and Murray Browne. Email surveillance using nonnegative matrix factorization. Computational & Mathematical Organization Theory, 11(3):249–264, 2005.

4. Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational statistics & data analysis, 52(1):155–173, 2007.

5. David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. IEEE signal processing magazine, 27(6):55, 2010.

6. Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In

Proceedings of the 27th international conference on machine learning (ICML-10), pages 111– 118, 2010.

7. Rostyslav Boutchko, Debasis Mitra, Suzanne L Baker, William J Jagust, and Grant T Gullberg. Clustering-initiated factor analysis application for tissue classification in dynamic brain positron emission tomography. Journal of Cerebral Blood Flow & Metabolism, 35(7):1104–1111, 2015.

8. Fred Brauer. Compartmental models in epidemiology. In Mathematical epidemiology, pages 19–79. Springer, 2008.

9. Yang Chen, Xiao Wang, Cong Shi, Eng Keong Lua, Xiaoming Fu, Beixing Deng, and Xing Li. Phoenix: A weight-based network coordinate system using matrix factorization. 44 IEEE Transactions on Network and Service Management, 8(4):334–347, 2011.

10. Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing, 3, 2014.

11. Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. Recent advances in deep learning for speech research at microsoft. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8604–8608. IEEE, 2013.

12. Nicolas Gillis. The why and how of nonnegative matrix factorization. Regularization, optimization, kernels, and support vector machines, 12(257):257–291, 2014.

13. Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.

14. Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of machine learning research, 5(Nov):1457–1469, 2004.

15. Matt J Keeling and Ken TD Eames. Networks and epidemic models. Journal of the Royal Society Interface, 2(4):295–307, 2005.

16. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

17. Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.

18. Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556–562, 2001.

19. H. Lyu, D. Needell, and L. Balzano. Online matrix factorization for markovian data and applications to network dictionary learning. 2019. Submitted.

BIOGRAPHIES

Dr. Matcha VenuGopal Rao

Professor of electronic and Communication Engineering at KL Univeristy,Vaddesawaram, Andhra Pradesh

Uppala. Lakshmi Sai

A UG Final Year student seeking her degree in Electronics and Communications Engineering at KL University Vaddeswaram, Andhra Pradesh

Arigela. Durga Sahithi

A UG Final Year student seeking her degree in Electronics and Communications Engineering at KL University Vaddeswaram, Andhra Pradesh