WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Prediction of Covid-19 Outbreak Using Machine Learning

Mukul Dhurkunde[1], Mohak Trivedi[2], Nayan Kadam[3], Sudarshan Kshirsagar[4]

[1,2,3,4]*Undergraduate Computer Science students of Sandip University, Nashik – 422213, India*

DOI: https://doi.org/10.55248/gengpi.2022.31291

A B S T R A C T

Coronavirus disease 2019 (COVID-19) is spreading rapidly; machine learning algorithms have been applied for a long time in many applications requiring the detection of adverse risk factors. The machine learning model proposed in this research paper uses three types of data, confirmed cases, recovered cases and deaths reported, the model can predict the spread of the virus in the next 20 days, and the data is time line series data and that is effective in predicting new cases of corona, death numbers and recovery.

Keywords: Coronavirus, Covid-19, Machine Learning, Support Vector Regressor (SVM), Linear Regression, Polynomial Regression, Randomized Search CV, Hyper-Parameter Optimization, Predictions

1. Introduction

In Wuhan, China, the Covid-19 pandemic, which is brought on by the new Extreme Acute Respiratory Syndrome Corona Virus 2 virus, occurred in December 2019. (SARS-CoV-2). The cause of Coronavirus Disease in 2019 is SARS-CoV-2 (COVID-19). On January 30, 2020, the World Health Organization (WHO) designated the outbreak a public health emergency and pandemic. Clinical signs of COVID-19 include respiratory problems, exhaustion, a dry cough, and fatigue, yet 80% of patients recover on their own. COVID-19 is more likely to affect older males, kids, men with cardiovascular disease, obesity, and men with diabetes. Clinical signs of COVID-19 include respiratory problems, exhaustion, a dry cough, and fatigue, yet 80% of patients recover on their own.[11] COVID-19 is a threat to older men, kids, men with cardiovascular disease, obesity, and men with diabetes. The greatest strategy to stop and slow down transmission is to keep your distance from others. By washing our hands or using hand sanitizers, we may protect ourselves and other people from infection. We also need to avoid touching faces. According to World-o-meter data, India will have 67,161 COVID19 cases and 2,212 fatalities by May 11th, 2020. In the entire world, 4,180305 people have been infected by viruses, and there have been 283,865 disease-related fatalities to yet. Hospitals only have a very small number of COVID-19 test kits, which is completely insufficient given the rising number of cases. So, in order to stop COVID-19 from spreading among humans, an automatic detection mechanism must be put in place. In fact, the COVID-19 crisis is being combated using artificial intelligence as the primary weapon. Deep learning and machine learning are two subfields in AI. It has numerous uses in the fields of computer vision and natural language processing. It aids in the diagnosis and forecasting of COVID-19 Tracking COVID cases, predicting, creating dashboards, diagnosing and treating patients, and creating alarms to preserve social distance as well as for other potential control mechanisms are all made possible by deep learning and machine learning techniques.

Fig. 1 - An example of how the covid passed from one person to another.

What is Machine Learning?

Machine learning (ML) is a topic of study focused on comprehending and developing "learning" methods, or methods that use data to enhance performance on a certain set of tasks. It is considered to be a component of artificial intelligence. Without being expressly taught to do so, machine learning algorithms create a model using sample data, also referred to as training data, in order to make predictions or judgements. Machine learning

algorithms are utilized in a wide range of applications, including computer vision, speech recognition, email filtering, medicine, and agriculture, where it is challenging or impractical to create traditional algorithms that can accomplish the required tasks. [4]

2. Problem Statement

The goal of this work is to develop a model that predicts the propagation of the infection over the ensuing 20 days. The aim is to create a model that forecasts the virus's spread over the following 20 days. [11]

3. Dataset

We have downloaded the dataset from the GitHub repository operated by the Johns Hopkins University Centre of System Science and Engineering. We have used data from three csv files [3]:

1) Confirmed cases
2) Recovered cases
3) Deaths reported

4. Machine Learning Libraries

We have used:

1) NumPy: for performing numerical computations.
2) Pandas: for data analysis and data cleaning.
3) Matplotlib: for data visualization.
4) Scikit-learn: for ML models and their optimization.[2]

5. Data Preparation

1) Extracted data regarding world cases, total deaths, mortality rate, recovery rate, total recovered and total active cases. Stored each in a separate Python list where each element is for a particular date. [12]
2) Extracted data regarding confirmed cases, deaths, recoveries for each country. Each country has its own Python list to store this data, where each element is for a particular date.
3) Extracted data regarding daily confirmed cases, daily deaths, daily recoveries for each country. Each country has its own list to store this data, where each element is for a particular date.
4) Extracted data regarding unique countries and stored them in a Python list.
5) Eliminate all those countries from the unique countries list, which have no confirmed cases at all.
6) Extracted data regarding total confirmed cases, total deaths, total recoveries, mortality rate, total active cases, all of these for each separate unique country. Stored each in a Python list where each element is for a unique country.
7) Create a data-frame using all of the above created lists, such that all elements are sorted in descending order of confirmed cases in a country.[7]
8) Repeat all of the above given steps for Provinces/States as well.
9) Eliminate Provinces having NaN values and their corresponding confirmed cases values.
10) future_forecast_dates: Added 20 days (predict outbreak for 20 days ahead) and then converted dates for all of these days into actual date-time format and stored these dates into a list called future_forecast_dates

6. Exploratory Data Analysis

In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. [10][8]

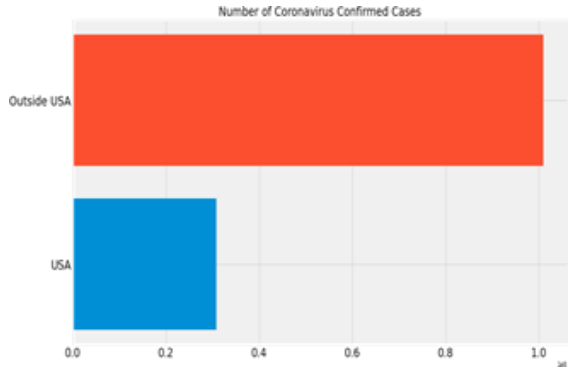


Fig.2 - Number of confirmed cases in the USA and outside USA

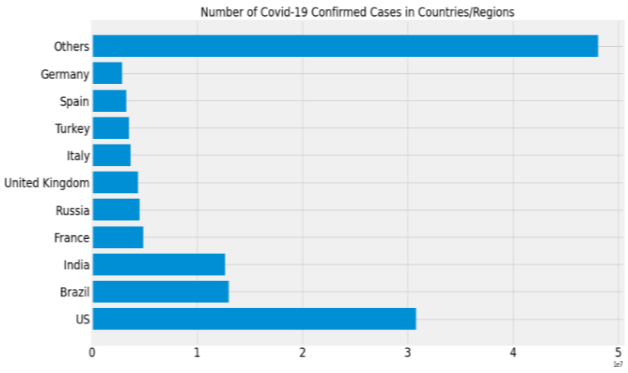


Fig.3 - Number of confirmed cases in the USA and outside USA

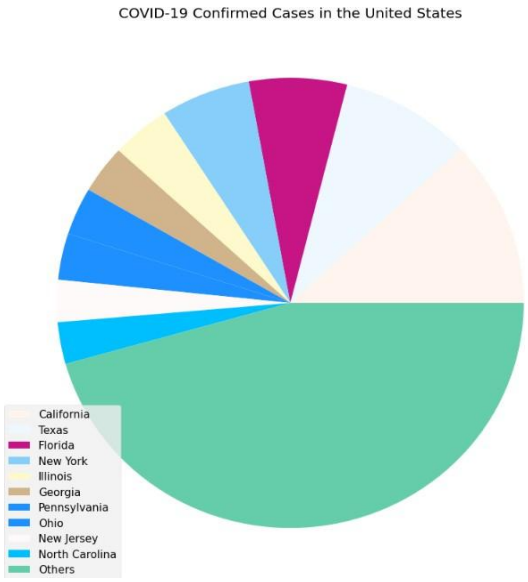


Fig.4 - Covid-19 confirmed cases in USA-states

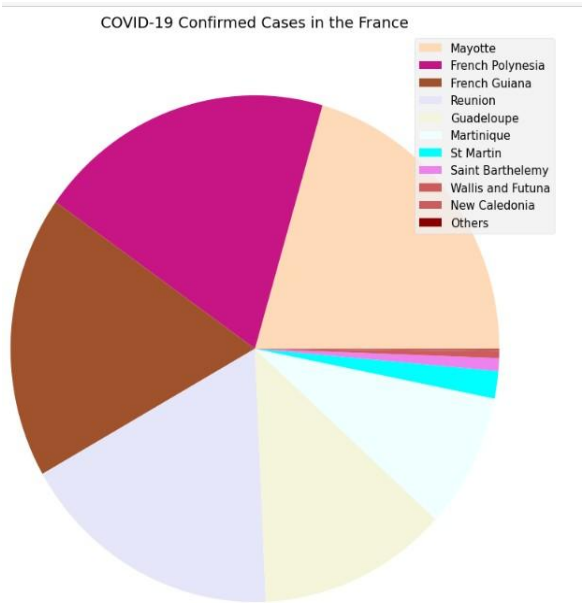


Fig.5 - Covid-19 confirmed cases in states within the USA and France.

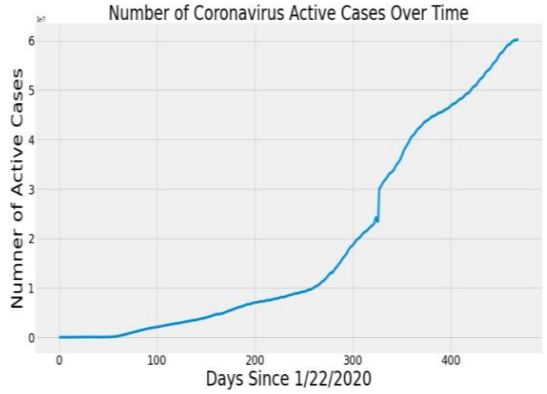


Fig.6 - Top 10 states/provinces all over the world with the highest number of confirmed casesas time passes, covid cases increasing steadily.

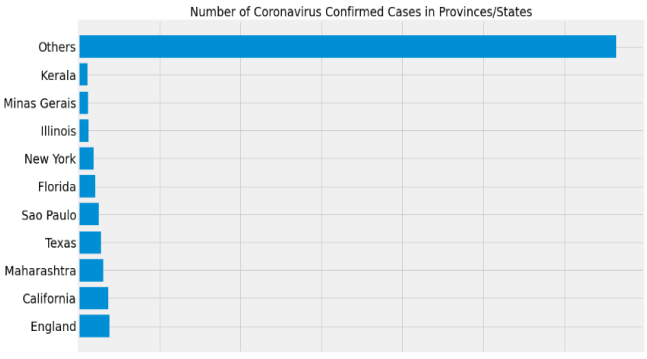
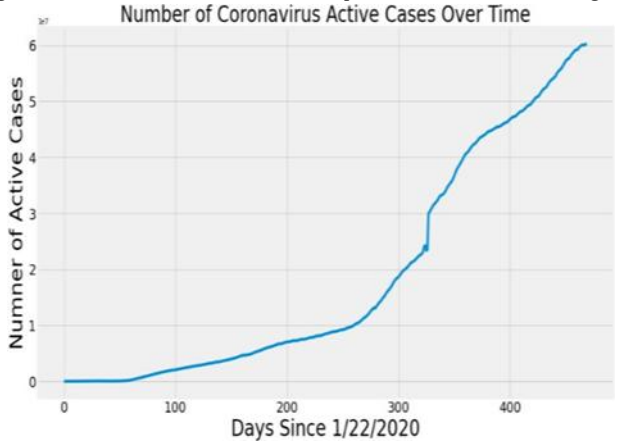


Fig.7 - Time-series analysis of world-wide coronavirus positive cases over time,



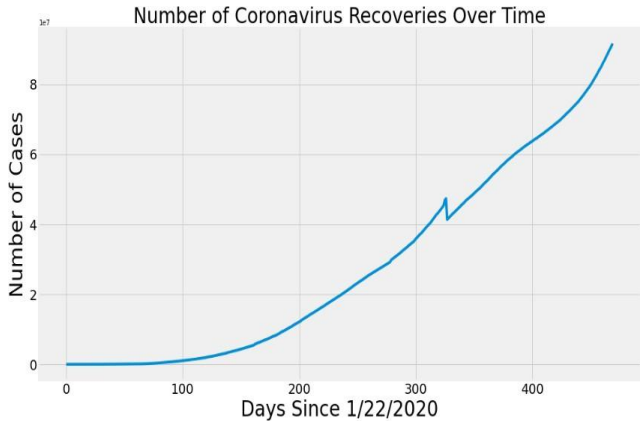


Fig.8 - Time-series analysis of world-wide confirmed coronavirus active cases over time

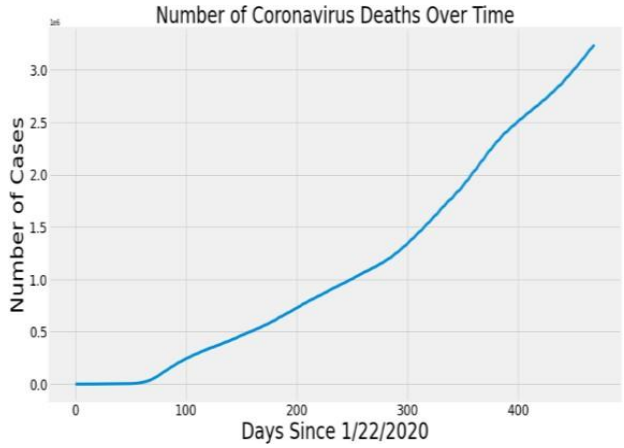


Fig.9 - There is slight drop in recoveries in between 300 to 400, as shown in graph

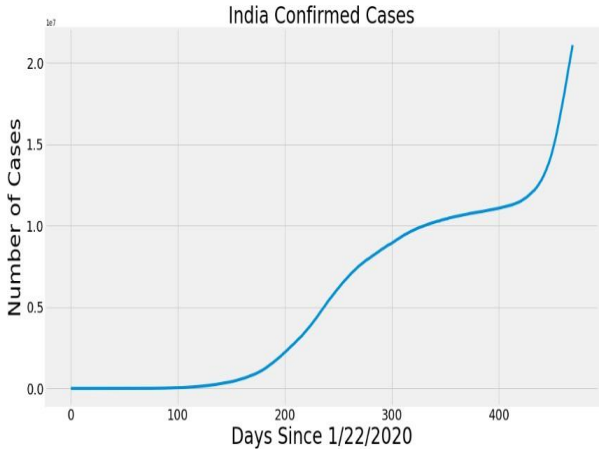
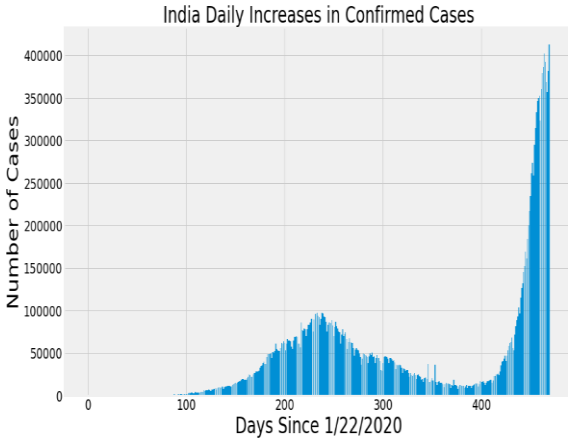


Fig.10 - Time-series analysis of world-wide confirmed, death, recovered and active cases of Covid-19

Fig.11 - Since 1/22/2020, confirmed cases started increasing over time in India.



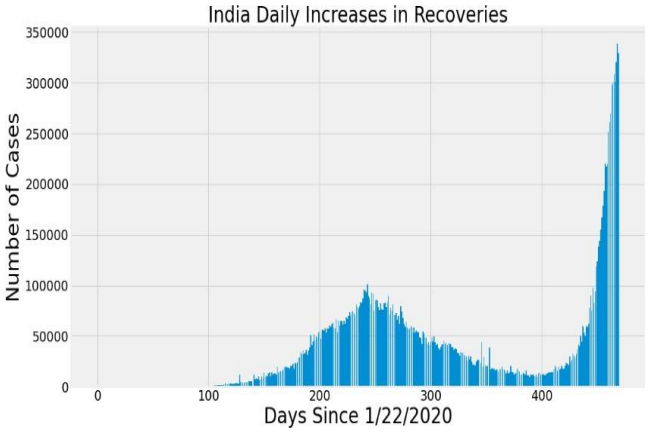


Fig.12 - Drawing critical analysis of COVID-19 hockey curve, few months back curve was flattening almost but later on curve began to increase

Fig.13 - As the world continues to battle against the coronavirus pandemic, India so far has recorded over 4M positive cases daily, out of which 3.5M have successfully recovered.

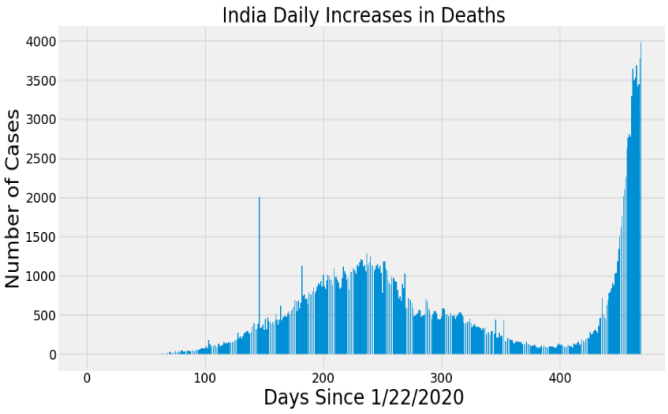


Fig.14 - Time series analysis of confirmed cases, daily increase in confirmed cases, daily increase in deaths and daily increase in recoveries in India.

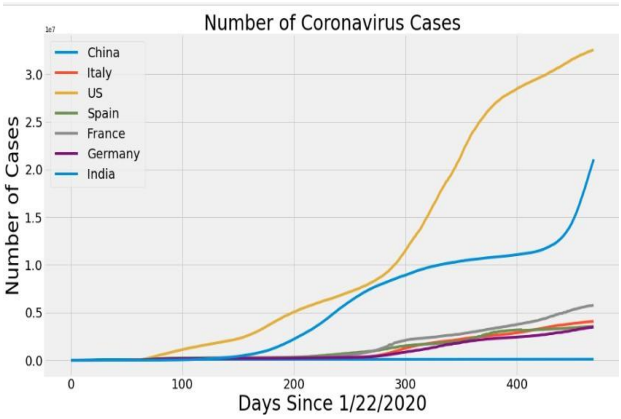


Fig.15 - Time series analysis for top 7 countries indicating their confirmed cases.

6.

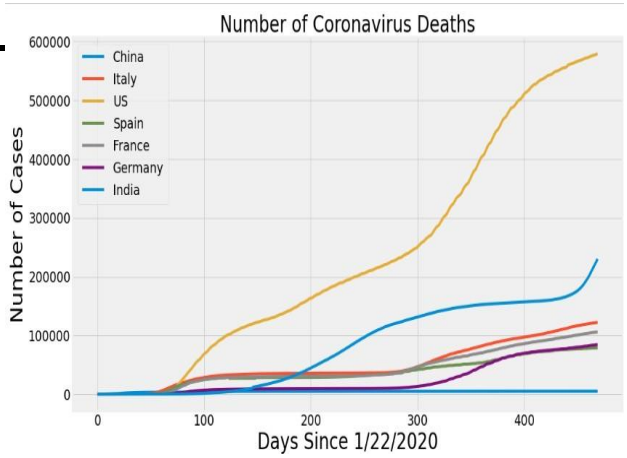


Fig.16 - Time series analysis for top 7 countries indicating their death cases.

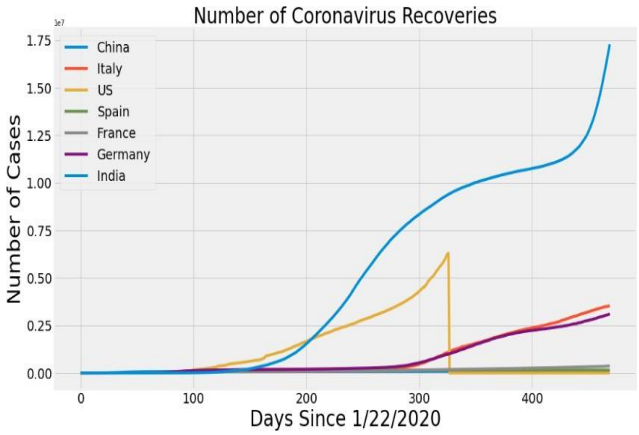


Fig.17 - Time series analysis for top 7 countries indicating their recovered cases.

Building ML Models:

6.1 Train – Test split:

A machine learning model aims to make good predictions on new, previously unseen data. But if you are building a model from your data set, how would you get the previously unseen data? Well, one way is to divide your data set into two subsets [9]:

- **Training set** - a subset to train a model.
- **Test set** - a subset to test the model.

Good performance on the test set is a useful indicator of good performance on the new data in general, assuming that

- The test set is large enough.
- Using the same test set repeatedly is not considered cheating.

Fig.18 - Data were used for testing in 25% of the cases, and training in 75%.

6.2 Support Vector Regressor (SVM Regressor):

Support vectors can also be applied to regression scenarios – where you estimate a real number instead of assigning classes to inputs. Support Vector Regression maintains all the interesting properties from Support Vector Machines. Given data points, it attempts to find a curve. However, rather than having the curve act as a decision boundary in a classification problem, in SVR, a match is found between some vector and the *position* on the curve. *It's a regression scenario after all*. And support vectors participate in finding the closest match between the data points and the actual function that is represented by them. We should get the closest to the actual curve, it seems, by maximising the distance between the support vectors and the regressed curve (because there is always some noise present in the statistical samples). It also follows that we can discard all the vectors that are no support vectors, for the simple reason that they are likely statistical outliers. The result is a regressed function[5]. Support vector regression is a special kind of regression that gives you some sort of buffer or flexibility with the error. How does it do that? I'm going to explain it to you in simple terms by showing 2 different graphs

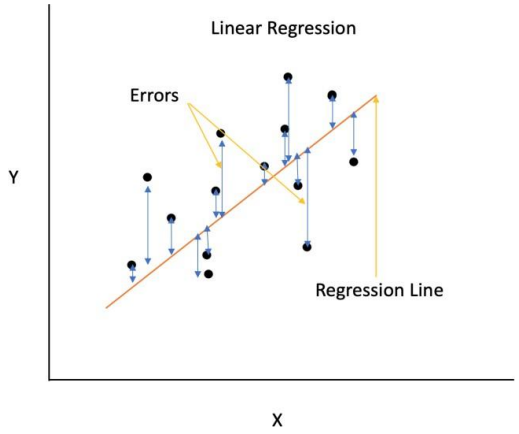
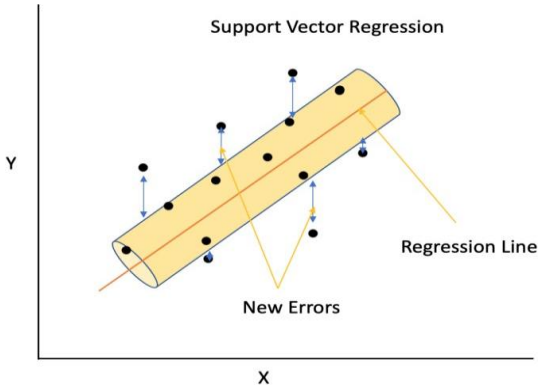


Fig.19 - hypothetical linear regression graph

The above is a hypothetical linear regression graph. You can see that the regression line is drawn at a position with minimum squared errors. Errors are basically the squares of difference in distance between the original data point (points in black) and the regression line (predicted values).



The setting above is identical to SVR (Support Vector Regression). You can see that the regression line is surrounded by two boundaries. This is a tube with the vertical distance of epsilon above and below the regression line. In reality, it is known as an epsilon insensitive tube. The role of this tube is that it creates a buffer for the error. To be specific, all the data points within this tube are considered to have zero error from the regression line. Only

the points outside of this tube are considered for calculating the errors[5]. The error is calculated as the distance from the data point to the boundary of the tube rather than data point to the regression line (as seen in Linear Regression)

How come support vectors?

Slack points, which are all the points outside of the tube, are essentially vectors in a two-dimensional space. Imagine drawing vectors from the origin to the individual slack points, then you can see all the vectors in the graph. Support vector regression is the term used to describe how these vectors are supporting the formation or structure of this tube. You can comprehend it by looking at the graph above. Kernels can also be applied in SVR. Hence, it is possible to regress a *nonlinear function*, or a curve, using SVR. Similarly, the non-linear data is mapped onto a space that makes the data *linear*. However, in the case of SVR, it is not necessary for it to be linear in order to distinguish between two groups; instead, it only needs to represent a straight line in order to calculate the contribution of support vectors to the regression issue.

6.3 Hyper – Parameter optimization:

The challenge of selecting a set of ideal hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning in machine learning. A parameter whose value is utilised to regulate the learning process is known as a hyperparameter. The values of other parameters, such as node weights, are often learned. To generalise various data patterns, the same machine learning model may need different constraints, weights, or learning rates. Hyperparameters are the measurements that need to be adjusted in order for the model to work best when solving a machine learning problem. An ideal model is produced through hyperparameter optimization, which identifies a tuple of hyperparameters that minimises a predetermined loss function on given independent data. The associated loss is returned by the goal function after receiving a tuple of hyperparameters. This generalisation performance is frequently estimated using cross-validation.

6.4 Randomized Search CV:

Set up a grid of hyperparameter values and select *random* combinations to train the model and score. The number of search iterations is set based on time/resources. Every time the search pattern is iterated, random parameter combinations are taken into account. Because of the random search pattern, which increases the likelihood that the model will be trained using the optimal parameters without any aliasing, the chances of finding the optimal parameter are noticeably higher in random search. An illustration of the random search's search pattern is provided below.

Fig.20 - Visual representation of random search

6.5 Polynomial Regression:

When fitting a polynomial equation to the data with a curvilinear relationship between the target variable and the independent variables, we are using polynomial regression, a special case of linear regression. In a curvilinear relationship, the value of the target variable changes in a non-uniform manner with respect to the predictor (s). In Linear Regression, with a single predictor, we have the following equation [1]:

$$Y = \theta_0 + \theta_1 x$$

Where:

- Y is the target ,
- x is predictor,
- $\theta 0$ is the bias and $\theta 1$ is the weight in the regression equation. This linear equation can be used to represent a linear relationship. But, in polynomial regression, we have a polynomial equation of degree n represented as:

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_n x^n$$

Here:

- $\theta 0$ is the bias,

- $\theta 1$, $\theta 2$, …. θn are the weights in the equation of the polynomial regression and n is the degree of the polynomial.

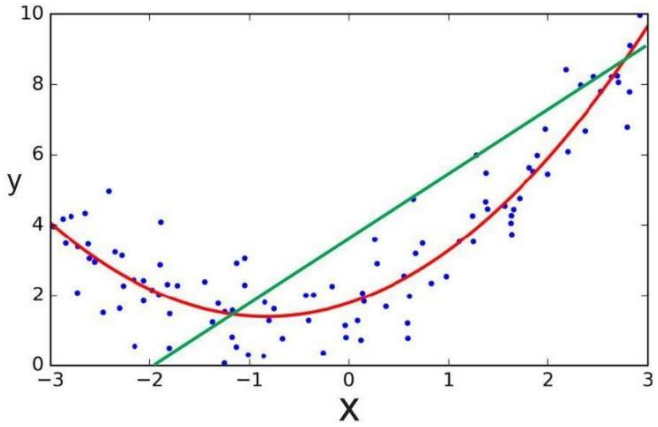


Fig.21 - Polynomial regression

In the above image you can see the example of a polynomial curve. Most of the real time data isn't really linear in nature, but non-linear. All the dots in blue in the above graph represent the data points and the green line is the linear regression line whereas the red line is the polynomial regression line. You can instantly see which line fits perfectly. It is the red line, i.e., it represents the general trend in the data within the given range.

7. Predictions:

Following are the predictions made by our Support Vector Regressor ML model, regarding worldwide Covid-19 confirmed cases[6]:

	Date	SVM Predicted of Confirmed Cases Worldwide
0	05/06/2021	211687459.0
1	05/07/2021	213032553.0
2	05/08/2021	214383372.0
3	05/09/2021	215739926.0
4	05/10/2021	217102229.0
5	05/11/2021	218470292.0
6	05/12/2021	219844127.0
7	05/1 3/2021	221223747.0
8	05/1 4/2021	222609163.0
9	05/1 5/2021	224000388 o
10	05/1 6/2021	225397435 o
11	05/1 7/2021	226800314.0
12	05/18/2021	228209039.0
13	05/19/2021	229623622.0
14	05/20/2021	231044074.0
15	05/21/2021	232470407.0
16	05/22/2021	233902635.0
17	05/23/2021	235340769.0
18	05/24/2021	236784821.0
19	05/25/2021	238234803.0

	Date	Predicted number of Confirmed Cases Worldwide
0	05/06/2021	205903754 0
1	05/07/2021	207194207 0
2	05/08/2021	208490141 0
3	05/09/2021	209791567 .0
4	05/10/2021	211098496.0
5	05/11/2021	212410941.0
6	05/12/2021	213728914.0
7	05/13/2021	215052426.0
8	05/14/2021	216381490.0
9	05/15/2021	217716117.0
10	05/16/2021	219056319.0
11	05/17/2021	220402107.0
12	05/18/2021	221753495.0
13	05/19/2021	223110494.0
14	05/20/2021	224473115.0
15	05/21/2021	225841370.0
16	05/22/2021	227215272.0
17	05/23/2021	228594832.0
18	05/24/2021	229980062.0
19	05/25/2021	231370974.0

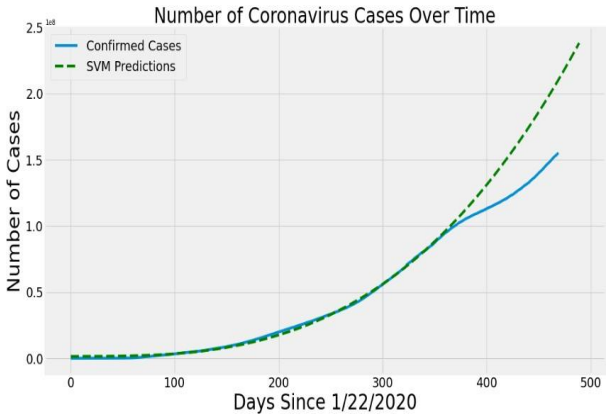


Fig.22 - SVR model's predictions as compared to actual values

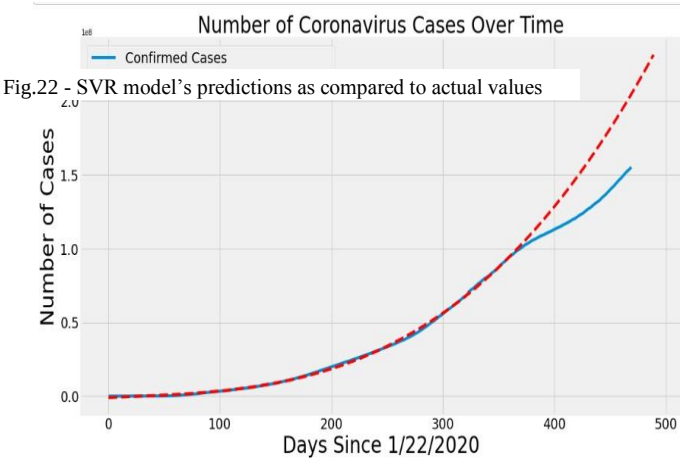


Fig.22 - Our Polynomial Regression model's predictions as compared to actual values.

Conclusion

The coronavirus disease continues to spread across the world, following a trajectory that is difficult to predict. The health, humanitarian, and socio-economic policies adopted by countries will determine the speed and strength of the recovery. Thus, we have successfully analysed the COVID-19 dataset as well as built ML models (SVM (Support Vector Machine) Regressor and Polynomial Regression) that predict the outbreak of COVID-19 cases. Our short-term forecasting model can aid real-time decision making that will be needed to prepare, such as anticipating the amount of equipment, beds for treatment in the hospital, and other medical resources that will be needed.

References

1) AlphA Bradley and Sushil S Srivastava. Correlation in polynomial regression. The American Statistician, 33(1):11–14, 1979.

2) AurelienGeron. ´ Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

3) Meg Miller. 2019 novel coronavirus covid-19 (2019-ncov) data repository: Johns hopkins university center for systems science and engineering. Bulletin-Association of Canadian Map Libraries and Archives (ACMLA), (164):47–51, 2020.

4) Tom M Mitchell and Tom M Mitchell. Machine learning, volume 1. McGraw-hill New York, 1997.

5) William S Noble. What is a support vector machine? Nature biotechnology, 24(12):1565–1567, 2006.

6) Kolla Bhanu Prakash, S Sagar Imambi, Mohammed Ismail, T Pavan Kumar, and YVRN Pawan. Analysis, prediction and evaluation of covid19 datasets using machine learning algorithms. International Journal, 8(5):2199–2204, 2020.

7) Dorian Pyle. Data preparation for data mining. morgankaufmann, 1999.

8) Ali Hassan Sial, Syed Yahya Shah Rashdi, and Abdul Hafeez Khan. Comparative analysis of data visualization libraries matplotlib and seaborn in python. International Journal, 10(1), 2021.

9) Patrice Y Simard, SaleemaAmershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. Machine teaching: A new paradigm for building machine learning systems. arXiv preprint arXiv:1707.06742, 2017.

10) Sandro Tosi. Matplotlib for Python developers. Packt Publishing Ltd, 2009.

11) Li Yang, Shasha Liu, Jinyan Liu, Zhixin Zhang, Xiaochun Wan, Bo Huang, Youhai Chen, and Yi Zhang. Covid-19: immunopathogenesis and immunotherapeutics. Signal transduction and targeted therapy, 5(1):1–8, 2020.

12) Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. Applied artificial intelligence, 17(5-6):375–381, 2003