International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217213

Predicting and Analysing the Behaviour of COVID-19

Gaurav Singh*, Shivam Rai, Himanshu Mishra, Manoj Kumar
Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Article Info
Volume 7, Issue 2
Page Number: 40-46

Publication Issue :
March-April-2021

Article History
Accepted : 02 March 2021
Published : 09 March 2021

ABSTRACT

The prime objective of this work is to predicting and analysing the Covid-19 pandemic around the world using Machine Learning algorithms like Polynomial Regression, Support Vector Machine and Ridge Regression. And furthermore, assess and compare the performance of the varied regression algorithms as far as parameters like R squared, Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. In this work, we have used the dataset available on Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at John Hopkins University. We have analyzed the covid19 cases from 22/1/2020 till now. We applied a supervised machine learning prediction model to forecast the possible confirmed cases for the next ten days.

Keywords: Covid-19 ,Prediction, Polynomial Regression, Ridge Regression, Support Vector Machine.

I. INTRODUCTION

This Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus[1]. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment[1]. Older people and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness[1]. The best way to prevent and slow down transmission is to be well informed about the COVID-19 virus, the disease it causes and how it spreads. Protect yourself and others from infection by washing your hands or using an alcohol-based rub frequently and not touching your face[1]. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes, so it's important that you also practice respiratory etiquette (for example, by coughing into a flexed elbow)[1].

On 31 December 2019, WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China. A novel coronavirus was identified as the cause by Chinese authorities on 7 January 2020 and was temporarily named "2019-nCoV"[2]. Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases[2]. A novel coronavirus (nCoV) is a new strain that has not been previously identified in humans. The new virus was subsequently named the "COVID-19 virus"[2].

Since the first cases were reported, WHO has worked around the clock to support countries to prepare and

respond to the COVID-19 pandemic[2]. In the words of Dr Hans Henri P. Kluge, WHO Regional Director for Europe, "Through transparent knowledge-sharing, tailored support on the ground, and steadfast solidarity, we will beat COVID-19"[2].

In this work, we are predicting and analysing the coronavirus disease (COVID-19) pandemic around the world using supervised machine learning regression models for comparative analysis of various parameters like r squared, mean absolute error, mean squared error and root mean squared error. In this work, we have applied supervised machine learning prediction models to forecast the possible confirmed cases for the next ten days and comparing accuracy.

II. PROPOSED WORK

Figure 1 shows the flow of supervised learning models with machine learning calculations, where the Covid-19 dataset is loaded, features need to be extracted and therefore the regression model is often trained and used for prediction of confirmed, deaths, and recovered cases.

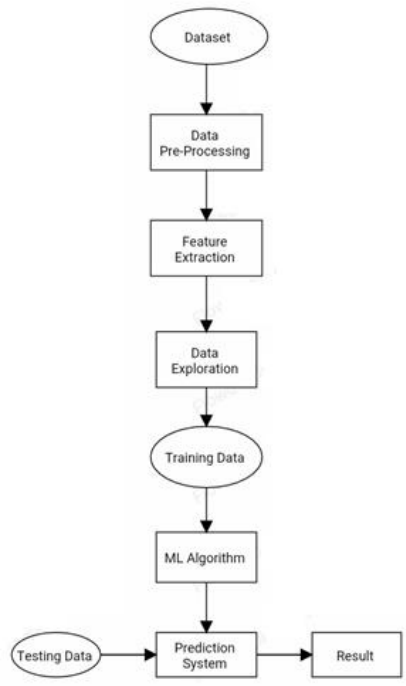

Figure 1. Flow Diagram

A. Polynomial Regression

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial In addition to linear terms.

B. Support Vector Machine

Support vector machine are powerful yet flexible supervised machine learning algorithms that are used both for classification and regression. But generally, they are used in classification problems. Support Vector Machine have their unique way of implementation as compared to other machine learning algorithms.

C. Ridge Regression

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

If the work Machine learning algorithm is done the we are calculating various parameters are as follows:

D. R Squared

It is also known as the **coefficient of determination**. This metric gives an indication of how good a model fits a given dataset. It indicates how close the regression line (i.e the predicted values plotted) is to the actual data values. The **R squared value lies between 0 and 1** where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

E. Mean Absolute Error (MAE)

We know that an error basically is the absolute difference between the actual or true values and the values that are predicted. Absolute difference means

that if the result has a negative sign, it is ignored. MAE takes the **average** of this error from every sample in a dataset and gives the output.

It is not very sensitive to outliers in comparison to MSE since it doesn't punish huge errors. It is usually used when the performance is measured on continuous variable data. It gives a linear value, which averages the weighted individual differences equally. The lower the value, better is the model's performance.

F. Mean Squared Error (MSE)

MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

It is one of the most commonly used metrics, but least useful when a single bad prediction would ruin the entire model's predicting abilities, i.e when the dataset contains a lot of noise. It is most useful when the dataset contains outliers, or unexpected values (too high or too low values).

G. Root Mean Squared Error (RMSE)

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

In RMSE, the errors are squared before they are averaged. This basically implies that RMSE assigns a higher weight to larger errors. This indicates that RMSE is much more useful when large errors are present and they drastically affect the model's performance. It avoids taking the absolute value of the error and this trait is useful in many mathematical calculations. In this metric also, lower the value, better is the performance of the model.

III. DATA COLLECTION

As W.H.O declared Coronavirus pandemic as Health Emergency. The researchers and hospitals give open access to the data regarding this pandemic. We have collected from an open-source data repository by the Center for Systems Science and engineering (CSSE) at John Hopkins University- This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL) [3].

IV. LITERATURE REVIEW

Akib et al. [4] using traditional machine learning algorithm in various features are being extracted from clinical records for detecting Covid-19 in which Naive Bayes classifier gives excellent results.

Furqan et al.[5] uses the Machine Learning methods for Covid-19 forecasting in which the results prove that Exponential Smoothing method performs best.

Hameed et al. [6] using the Long Short Term Memory Recurrent Neural Network Approach to detect meaningful sentiment-comment-classification on Covid-19 related issues from healthcare forums, such as subreddits.

Mohammad et al. [7] using some deep learning methods namely Generative Adversarial Networks, Extreme Learning Machine, and Long Short Term Memory to geographical issues, high-risk people and recognizing were the main problems with Covid-19 and have been studied and discussed in this work.

V. OBSERVATION

Figure 2 shows the graphical representation of rate of confirmed cases top five countries around the world.

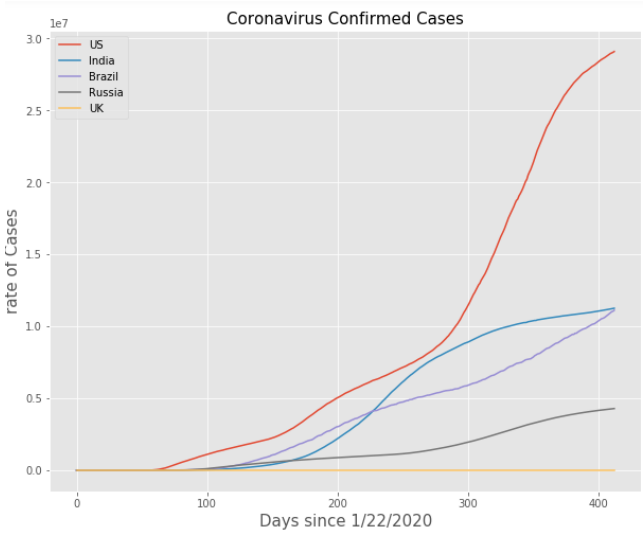

Figure 2. Top 5 Countries Confirmed Cases

Figure 3 shows the graphical representation of rate of death cases top five countries around the world.

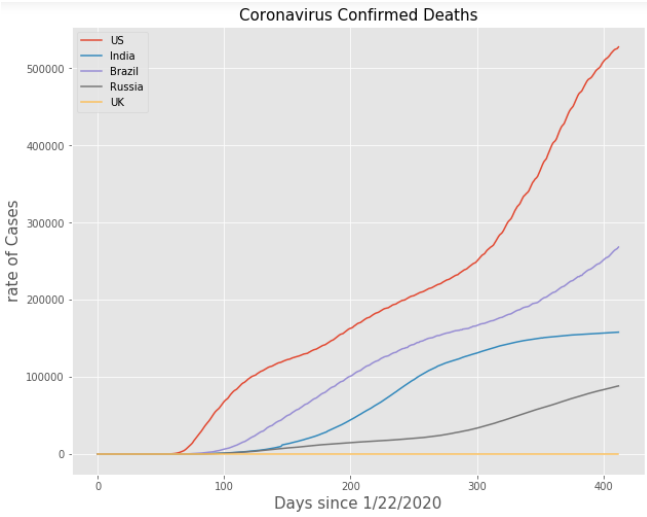

Figure 3. Top 5 Countries Death Cases

Figure 4 shows the graphical representation of rate of recovered cases top five countries around the world.

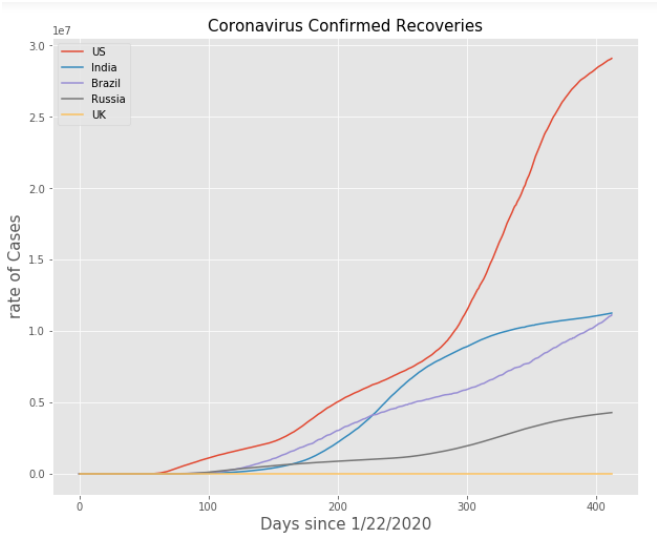

Figure 4. Top 5 Countries Recovered Cases

Figure 5 shows the rate of confirmed cases vs polynomial regression predicted confirmed cases worldwide over time.

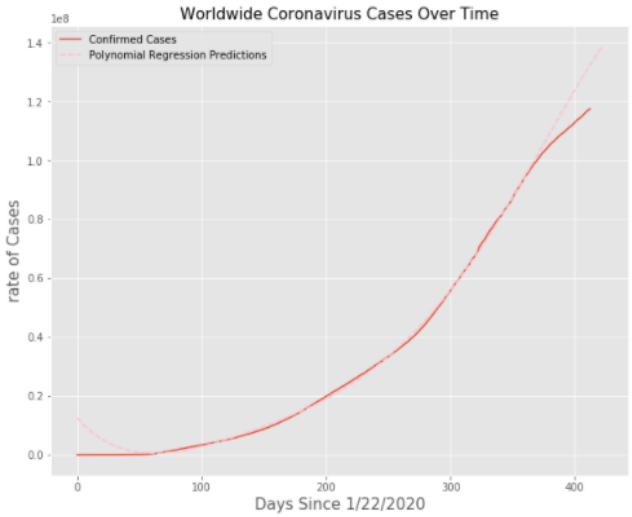

Figure 5. Confirmed vs Polynomial Prediction

Figure 6 shows the rate of confirmed cases vs support vector machine predicted confirmed cases worldwide over time.

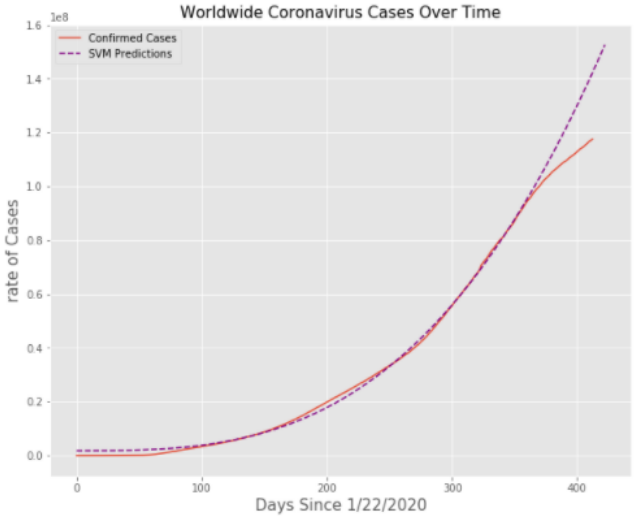

Figure 6. Confirmed vs SVM Prediction

Figure 7 shows the rate of confirmed cases vs ridge regression predicted confirmed cases worldwide over time.

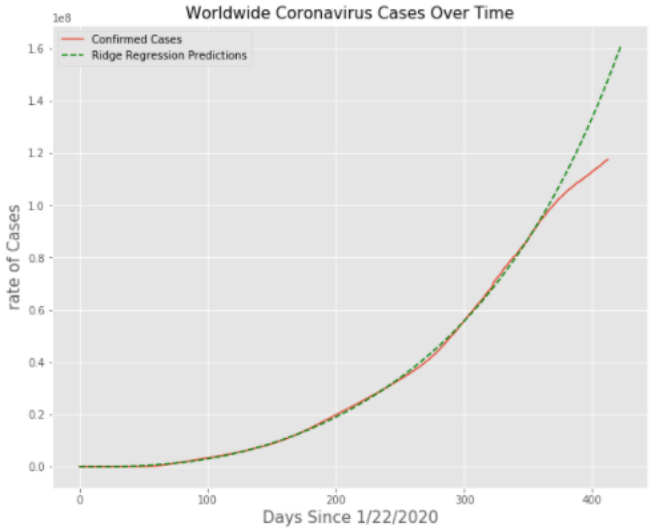

Figure 7. Confirmed vs Ridge Regression Prediction

Table 1 shows the R squared values of supervised machine learning regression algorithms.

Table 1:- R Squared Values

Algorithms	R Squared
Polynomial Regression	76.40%
Support Vector Machine	86.44%
Ridge Regression	40.08%

Table 2 shows the Mean Absolute Error values of supervised machine learning regression algorithms.

Table 2:- MAE Values

Algorithms	MAE
Polynomial Regression	2224323.6536729196
Support Vector Machine	279532.8692821376
Ridge Regression	3598182.0896637505

Table 3 shows the Mean Squared Error values of supervised machine learning regression algorithms.

Table 4:- MSE Values

Algorithms	MSE
Polynomial Regression	9219465117160.574
Support Vector Machine	183519597828.3253
Ridge Regression	14420416763240.494

Table 4 shows the Root Mean Squared Error values of supervised machine learning regression algorithms.

Table 4:- RMSE Values

Algorithms	RMSE
Polynomial Regression	3036357.2117194273
Support Vector Machine	428391.87413900055
Ridge Regression	3797422.384096941

VI. RESULTS AND DISCUSSION

Table 5 shows that the predicted upcoming worldwide confirmed cases for next ten days using Polynomial Regression.

Table 5 :- Polynomial Predicted Confirmed Cases

	Date	Polynomial Predicted Number of Confirmed Cases Worldwide
0	03/10/2021	133010925.0
1	03/11/2021	133677581.0
2	03/12/2021	134339554.0
3	03/13/2021	134996679.0
4	03/14/2021	135648788.0
5	03/15/2021	136295714.0
6	03/16/2021	136937284.0
7	03/17/2021	137573327.0
8	03/18/2021	138203667.0
9	03/19/2021	138828127.0

Table 6 shows that the predicted upcoming worldwide confirmed cases for next ten days using Support Vector Machine.

Table 6 :- SVM Predicted Confirmed Cases

	Date	SVM Predicted Number of Confirmed Cases Worldwide
0	03/10/2021	143155228.0
1	03/11/2021	144184334.0
2	03/12/2021	145218424.0
3	03/13/2021	146257510.0
4	03/14/2021	147301603.0
5	03/15/2021	148350716.0
6	03/16/2021	149404860.0
7	03/17/2021	150464049.0
8	03/18/2021	151528293.0
9	03/19/2021	152597605.0

Table 7 shows that the predicted upcoming worldwide confirmed cases for next ten days using Ridge Regression.

Table 7 :- Ridge Predicted Confirmed Cases

	Date	Ridge Predicted Nuber of Confirmed Cases Worldwide
0	03/10/2021	149214028.0
1	03/11/2021	150470332.0
2	03/12/2021	151736771.0
3	03/13/2021	153013429.0
4	03/14/2021	154300390.0
5	03/15/2021	155597738.0
6	03/16/2021	156905556.0
7	03/17/2021	158223931.0
8	03/18/2021	159552947.0
9	03/19/2021	160892692.0

The above result shows the predicted result of confirmed cases using machine learning algorithms namely Polynomial regression, Support vector machine regression and Bayesian ridge regression for upcoming ten days.

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we have using three different types of machine learning regression algorithm namely Polynomial Regression, Support Vector Machine, and Ridge Regression for predicting Covid-19 confirmed cases for next ten days the target of this comparative analysis was to search out the foremost accurate machine learning algorithm which will act as a tool for predicting Covid-19 confirmed cases worldwide. consistent with the prediction results, Support Vector Machine has the very best accuracy for this predictive analysis.

In future we are using Deep Learning techniques like LSTM model for improving results and accuracy. Also we are are comparing Covid-19 with other epidemic like influenza in future.

VIII. ACKNOWLEDGEMENT

I have completed this work under the guidance of Ms. Sapna Yadav(Assistant Professor), Department of Computer Science and Engineering at IMS Engineering College, Ghaziabad, Uttar Pradesh. I am doing an online Summer Internship on Machine Learning where I have learn various Machine Learning algorithm from both of my mentor as a course instructor. This paper has been assigned as a project assignment for us.

I would like to express my special thanks of my mentor for inspiring us to complete work and write a paper. Without their active guidance, help cooperation & encouragement, I would not lead way in writing the paper. I am extremely thankful for

their valuable guidance and support on completion of this paper.

I extend my gratitude to "IMS Engineering College, Ghaziabad, Uttar Pradesh" for giving me this opportunity. I also acknowledge with a deep sense of reverence, my gratitude towards my friends, parents and member of my family, who always supported me morally, mentally as well as economically.

Any omission in this brief acknowledgement does not mean a lack of gratitude.

IX. REFERENCES

[1]. World Health Organization- https://www.who.int/
[2]. World Health Organization- https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov
[3]. COVID-19 Data Repository by the Center for Systems Science and engineering (CSSE) at John Hopkins University: https://github.com/CSSEGISandData/COVID-19
[4]. Akib Mohi Ud Din Khanday,"Machine learning based approaches for detecting COVID-19 using clinical text data" 2020, Springer
[5]. Furqann Rustam,"COVID-19 Future Forecasting Using Supervised Machine Learning Models", 2020,IEEE
[6]. Hamed Jelodar,"Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach", 2020, IEEE
[7]. Mohammad (Behdad) Jamshidi," Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment" 2020, IEEE

About Author's

Gaurav Singh is B.Tech student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. His areas of interest is Programming in Python, Data Science and Machine Learning.

Shivam Rai is B.Tech student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. His areas of interest is Programming in Python, Data Science and Machine Learning.

Himanshu Mishra is B.Tech student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. His areas of interest is Programming in Python, Data Science and Machine Learning.

Manoj Kumar is B.Tech student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, UP, India. His areas of interest is Programming in Python, Data Science and Machine Learning.

Cite this article as :

Gaurav Singh, Shivam Rai, Himanshu Mishra, Manoj Kumar, "Predicting and Analysing the Behaviour of COVID-19", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 40-46, March-April 2021.
Available at
doi : https://doi.org/10.32628/CSEIT217213
Journal URL : https://ijsrcseit.com/CSEIT217213