Forecasting COVID-19 using Polynomial Regression and Support Vector Machine

Javeriya Bano Altaf Hussain[1], S.S.Hatkar[2]

[1]Dept. of Computer Science and Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, Maharashtra, India
[2]Associate Professor, Dept. of Computer Science and Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, Maharashtra, India

---***---

Abstract - *Many lives are getting affected by COVID-19 daily. Machine learning always plays an important role in health care sectors. Many researchers have used different machine learning models for prediction of COVID-19. This paper uses two supervised machine learning models i.e., polynomial regression and support vector machine. These models can forecast for the next 20 days COVID-19 cases. The efficiency of polynomial regression is more than support vector machine.*

***Key Words*: Forecasting, Covid-19, Supervised machine learning, Support vector machine, Polynomial regression**

1.INTRODUCTION

COVID-19 is a disease which spreads from infected person to healthy person. In December 2019, the first case was found in Wuhan, China. Schools, colleges, markets, offices, parks, gyms, etc. had been shut down. The disease had spread in every corner of the world. Therefore, it has been declared a pandemic. Many people died. Many people migrated because they lost their jobs. To prevent spread of disease various prevention methods are implemented. They are face coverings(usually done by taking help of masks), hand washing, social distancing(keeping distance between two people), Older people are at a higher risk of getting infected. Many complications have been observed in people post recovery i.e., kidney failure, pneumonia, etc. To understand long term effects many studies, need to be done [1].

Many COVID-19 vaccines have been developed, tested, approved, and distributed all around the world. World's largest population is believed to be vaccinated very soon. Vaccines do not confirm that vaccinated people will not get infected. There had been cases where people who had taken 2 doses of vaccines also got infected. Other measures like physical distancing, use of face masks is also necessary even if a person gets vaccinated.

Tiredness, fever, loss of taste , and cough are the most repeated indications of COVID-19. Diarrhea, sore throat, red eyes, headache , rashes on skin, pains are less repeated indications [2]. One must drink a lot of fluids so that the body does not get dehydrated and take proper rest.

This paper includes four sections. Section 1 is about the introduction. In section 1 covid-19 has been explained. Its damaging effects have been discussed. Symptoms of COVID-19 are listed. Preventive measures have been suggested. COVID-19 vaccines' importance is discussed. Section 2 is all about training, testing and models used. Section 3 mentions information obtained from dataset and result. Section 4 ends the paper. At the end it was revealed that polynomial regression is better than support vector machine.

2. TRAINING, TESTING AND MODELS USED

2.1 Supervised Machine Learning Model

In supervised machine learning, models or machines are trained firstly. For training of machines, "labelled" training data is used. When training a machine is done, the machine is now able to predict the output [3]. This paper uses two models for prediction of COVID-19. They are Support Vector Machine model and Polynomial Regression model.

2.1.1 Support Vector Machine (SVM)

The full form of SVM is Support Vector Machine. SVM can put two different kinds of data into its respective category by taking help of a line. With help of this line, we can easily put data into its respective category. The best line is also known as a hyperplane.

When SVM creates a hyperplane, it also selects nearest points. These nearest cases are called support vectors. Therefore, the algorithm's name is Support Vector Machine. There are two different categories of data that are classified using a hyperplane [4].

To understand SVM, we can take help from the following case. Consider a situation, a strange horse looks very similar to a donkey. A model can be created that can tell whether it is a donkey or horse. Firstly, the model is trained by taking thousands of images of donkeys and horses. Thus, the model understands different properties of donkeys and horses. After that, a strange horse can be tested using the created model. So, the created model draws a decision boundary between the data of donkey and horse and decides support vectors. With the help of support vectors, it will decide whether it is donkey or horse.

There can be thousands of decision boundaries to divide the categories. But only one the finest decision boundary is selected to divide the categories. The finest decision boundary is called a hyperplane. The hyperplane which has the largest border is selected.

The points which are the nearest to the hyperplane as well as influence the whole location of the hyperplane are called the support vector. The vectors assist the whole hyperplane. Therefore, they are known as support vectors.

The functioning of the SVM algorithm can be understood by taking help of the following case. There is a dataset that has two labels(red and yellow). The dataset has two features y1 and y2. A classifier needs to be created that can classify the pair (y1,y2) whether the point is red or yellow. There can be thousands of lines that divide the red and yellow. The SVM role is to search for the best line, or it can be called a decision boundary. Hyperplane is nothing but that decision boundary. SVM algorithm searches the nearest dots from both the categories. These dots are nothing but support vectors. Margin is nothing but the gap between hyperplanes and vectors. SVM aims to make the margin as broad as possible. Optimal hyperplane is nothing but a hyperplane with the widest margin.

2.1.1 Polynomial Regression (PR)

PR stands for polynomial Regression. In polynomial regression, there are two variables. The variable y is dependent, and x is independent. Polynomial regression is nth degree polynomial; therefore, it is called polynomial regression. Polynomial Regression's job is to build relationships between x and y.

The difference between linear regression and polynomial regression is that some terms are inserted into polynomial regression. By converting multiple linear regression, we get polynomial regression. Polynomial regression is an advanced version of linear regression which gives more accuracy as compared to linear regression. Non-linear dataset is given in polynomial regression [5].

If data is linear in nature, then linear regression is best. If data is non-linear, linear regression is the worst option because accuracy is reduced, loss function is increased, and error rate will be too high. When data is non-linear, Polynomial regression is the best option.

Linear models are not able to handle non-linear data. But a polynomial model can cover every point. Thus, it is concluded that if the dataset is non-linear in nature at that time instead of linear model polynomial model should be preferred . Polynomial regression is also known as polynomial linear regression. Polynomial regression is independent of variables but dependent on coefficients.

2.2 Training and Testing

This research paper has been written to study about COVID-19 predictions. COVID-19 is also known as novel coronavirus. COVID-19 has been declared as pandemic . Lakhs of people have died each day. At this time , it becomes very necessary to forecast COVID-19 cases. This forecast will help in several ways.75% of data is used for training purpose. Remaining 25% of data is used for testing the model. Learning models used are SVM and Polynomial Regression.

3. OUTPUT

3.1 Dataset and Information Obtained from The Dataset

Aim is to forecast confirmed upcoming cases. John Hopkins University's provided data is used in this paper [6]. Four different datasets have been taken for this research. They are confirmed cases, deaths reported, recovered cases and latest data. The labels used for training the models are world cases, total deaths, total recovered. Datasets contain information of mortality rate, confirmed cases, active cases, recoveries, and deaths of 184 countries. Top 10 countries with the highest number of confirmed cases are US, Spain , Italy, France, Germany, United Kingdom, Turkey, Iran, China, Russia, and Brazil respectively. India is listed at 15th number. Top 10 countries with the lowest number of confirmed cases are Yemen, Sao Tome and Principe, South Sudan, Western Sahara, Mauritania, Bhutan, Papua New Guinea, MS Zaandam, Holy See, and Suriname respectively. Number of confirmed cases in USA as well as outside USA have been shown in figure 1. US confirmed cases have been shown by taking help of blue horizontal bar. Outside US confirmed cases have been shown by taking help of red horizontal bar. Figure 2 shows confirmed cases in different countries like US, Spain, Italy, France, Germany, United Kingdom, Turkey, Iran, China, Russia. Other countries are shown in others labelled horizontal bar [7].

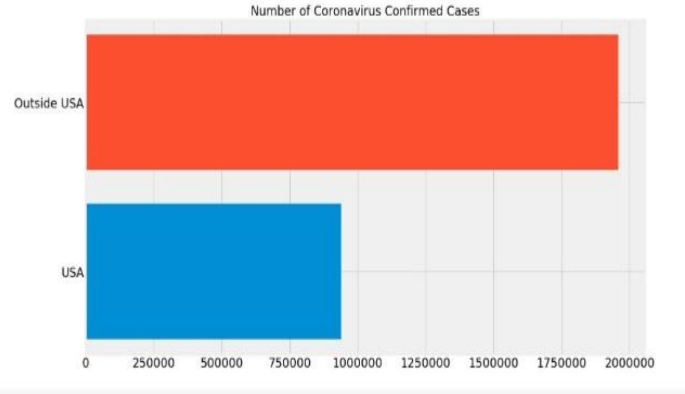

Fig -1: Confirmed cases in the US vs outside US

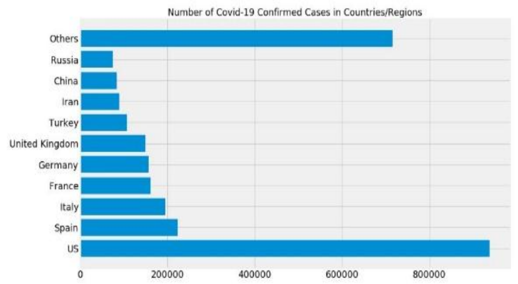

Fig -2: Confirmed cases in different countries

3.2 Result

This paper aims to predict upcoming cases of coronavirus. Upcoming 20 days prediction have been done. This prediction will help government, hospitals, doctors, people, and NGO take necessary steps to reduce the damage caused by coronavirus as much as possible. Two machine learning models SVM and poly LR have been used. The result is shown in the Table 1 given below. Predictions done by the model SVM and poly LR are shown in figure 3 and figure 4.

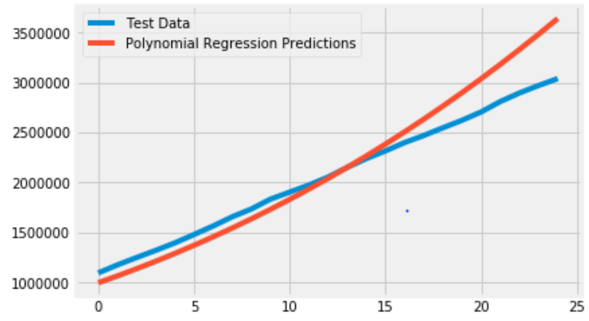

Fig -3: Polynomial regression prediction

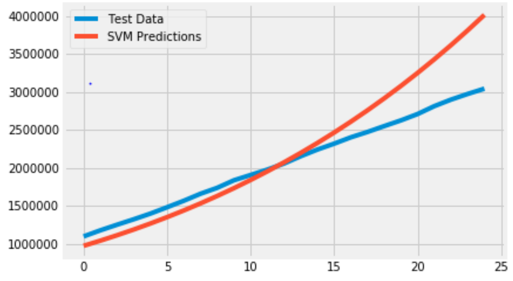

Fig -4: Support vector machine prediction

Table -1: Efficiency of models during the prediction

Models	**Mean Absolute Error**	**Mean Square Error**
Support Vector Machine	265489.6552454628	139718600180.29932
Polynomial Regression	172343.87626768733	54097652301.452614

4. CONCLUSIONS

After COVID_19 arrival, many scientists have developed COVID-19 vaccines in a very less amount of time [8]. Many researchers have tried to find out COVID-19 upcoming cases. Their research has helped government to take necessary actions on time. It is true that world's most population has been vaccinated but it is also necessary to predict for upcoming days. The dataset contain data for each day . Two ML models polynomial regression and support vector machine have been used for prediction. Both models provide good result, but polynomial regression shows more accuracy over support vector machine. MAE for polynomial regression is 172343.87626768733. MSE for polynomial regression is 54097652301.452614. MAE for SVM is 265489.6552454628. MSE for SVM is 139718600180.29932.The model is providing good results . Thus, this model can be used for future prediction.

REFERENCES

[1] WHO, "Novel Coronavirus (2019-nCoV)," WHO Bull., 2020.

[2] A. Waris, U. K. Atta, M. Ali, A. Asmat, and A. Baset, "COVID-19outbreak: current scenario of Pakistan," New Microbes and New Infections.2020,doi:10.1016/j.nmni.2020.100681.

[3] Balika J. Chelliah, S. Kalaiarasi, Apoorva Anand, Janakiram G, Bhaghi Rathi, Nakul K. Warrier, Classification of Mushrooms using Supervised Learning Models, IJETER, Vol 6(4), April 2018, ISSN 2454-6410, pp 229-232.

[4] P.Rivas-Perea, J. Cota-Ruiz, D. Chaparro, J. Venzor, A. Carreón and J.Rosiles, "Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations," International Journal of Intelligence Science, Vol. 3 No. 1, 2013, pp. 5-14.

[5] Documentation of SciKitLearn package, scikit-learn.org

[6] Dataset from official github website of John Hopkins University https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

[7] Google Collaboratory documentation by Google, https://colab.research.google.com/notebooks/intro.ipynb

[8] N. Noreen et al., "Coronavirus disease (COVID-19) Pandemic and Pakistan; Limitations and Gaps," Limitations Gaps. Glob. Biosecurity,2020.

[9] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends inCOVID-19 with logistic model and

machine learning technics," Chaos,Solitons & Feactals, vol. 139, 2020.

[10] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A novel biclustering approach to association rule mining for predicting hiv-1–human protein interactions," PLoS One, vol. 7, no. 4, p. e32289, 2012.

[11] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[12] H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang, and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): status and future perspectives, "International Journal of Antimicrobial Agents, p. 105951, 2020.

[13] J. Wu et al., "Rapid and accurate identification of COVID-19infection through machine learning based on clinical available blood testresults,"2020,doi: 10.1101/2020.04.02.20051136.

[14] Furqan Rustum, Aijaz Ahmad Reshi, Arif Mehmood ,Saleem Ullah ,Byung-Won On, Waqar Aslam, Gtu Sang Choi, Covid-19 future forecasting using supervised machine learning models,2020.