International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

# Prediction of COVID-19 Outbreak using Machine Learning

**Richa Tamhane[1], Sumeet Mulge[2]**

[1,2]*Student, CSE, SVERI's College of Engineering, Pandharpur, Maharashtra, India*

---

**Abstract -** *Artificial intelligence has been proven to be a dormant and powerful tool in the fight against COVID-19 pandemic. The purposes needed of Artificial Intelligence are Machine Learning, Natural Language Processing (NLP) and Computer Vision Applications. These models are used to teach computers to recognize patterns, to do predictions, etc. The goal of this paper is to predict the outbreak of COVID-19 using two techniques of Machine Learning viz. Data Visualization and Predictions using Polynomial Regression and Support Vector Regression. Predictions would help to get the future count of the COVID-19 cases and thus to establish precautions before the situation goes out of control. AI still needs a lot of accurate data and continuous human-AI interaction in order to build models against COVID-19 still it has become a ray of hope for innovation in technology so as to help mankind from this global pandemic.*

***Key Words***: **machine learning, artificial intelligence, prediction, COVID-19, COVID-19 outbreak, data visualization, polynomial regression, support vector regression.**

## 1.INTRODUCTION

The COVID-19 pandemic upon which the world is growling right now has been one of the most hilarious life destruction and has been defined as the global health crisis of our time. Coronaviruses are a family of viruses that cause illness such as respiratory disease of gastrointestinal diseases. The COVID-19 disease is caused by the virus called SARS-CoV-2 virus.

The COVID-19 patient first emerged in Wuhan in December 2019. India reported its first case on January 30 2020 and now has become a global pandemic. This situation should be handled wisely so as to take proper precautions before the count goes out of control.

Many scientists are taking major efforts to save mankind from this disaster. In this era of technology, Artificial Intelligence and Machine Learning are playing vital roles by giving a stand for innovating the technology. Many of the data scientists worldwide are engaged in getting the right datasets and building the strong models in order to fight against this pandemic.

AI has been contributing in various fields like making early warnings and alerts, tracking and prediction, data dashboards, social control, treatments and cures, etc.

This paper focuses on building the prediction model to predict the case count for next 20 days. Polynomial Regression and Support Vector Regression are the two techniques of Machine Learning used for obtaining the results. Python language makes it simpler to obtain the desired outputs. It acts as a necessary tool to uncover hidden insights and predict future trends.

### 1.1 Collecting data for predictions

Data collection is often a very challenging path while developing any machine learning model and it is essential to make hands dirty at this stage. The perfect dataset probably doesn't exist as the tremendous growth of data day by day or rather at every next second. In order to collect data for creating a primary dataset, following steps are to be followed.

The dataset is provided and updated periodically by the John Hopkins University and is used for the study factors such as confirmed cases, number of deaths, number of recoveries, locations, active cases, dates and so on. The dataset extracted for this study is from January 1 2020 for the prediction. The latest dataset is 25 May 2020.

### 1.2 Data Visualization

The theoretical meaning of Data Visualization is the graphical or pictorial representation of data or information in the form of graphs, charts and maps, etc. In the world of data science, data visualization is used for analysing large amounts of data and information and producing data-driven decisions. Human eyes are subtle to colours and data presented in different colours gives easy grasp to the user and hence can quickly analyse it.

The libraries such as pandas, matplotlib, seaborn in order to plot graphs with a python program.

- **Pandas Visualization**
  Pandas is used to create plots from pandas data frames and libraries. It is an open source data structure and data analytics tool used in Python. Pandas can import various formats of files while .csv is the most popular.
- **Matplotlib**
  Matplotlib is a library used in python to create static, animated and interactive visualizations. It comes up with a wide variety of applications using GUI toolkits like Tkinter, , wxPython, etc. This paper uses the pyplot API to plot areas, plot lines and decorate the plot with labels.

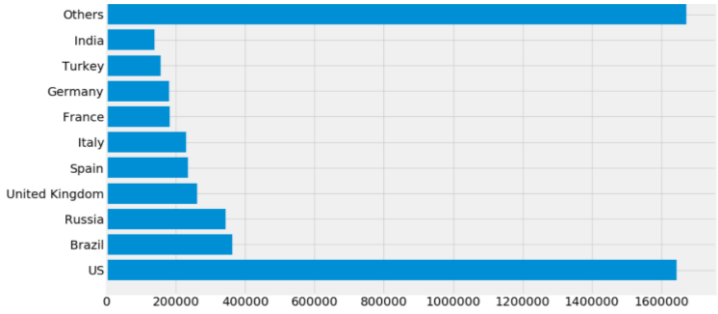


**Fig-1:** Top 10 highly affected countries by COVID-19

The example above shows the bar graph of the data representing the top 10 highly affected countries by this pandemic.

As it is said, data is only good when it's well presented and thus this makes it easier to get a clear idea of data.

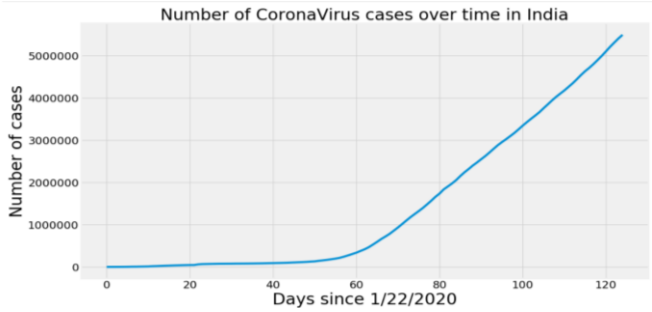


**Fig -2:** Number of COVID-19 cases over time in India

## 2. PREDICTION

"Prediction" refers to the future outcome of a situation by an algorithm. This is prior trained with a historical dataset and is applied to the new data for prediction. The prediction models provide the insights that result in the tangible outcomes.

### 2.1 Regression

Regression involves predicting the outcomes based on the inputs. Regression techniques in Machine Learning varies from Linear Regression, Multiple Regression, Polynomial Regression, SVR, etc. These are the supervised learning techniques. This paper includes prediction using polynomial regression and support vector regression techniques.

#### 2.1.1 Polynomial Regression

Polynomial Regression is simply a form of Linear Regression whereas Linear Regression is always a straight line and Polynomial Regression outputs a curve. Polynomial Regression shows the relationship between the variables x and y and finds a best way to draw a line through the data points. Given below is an equation for Polynomial Regression:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_2x_1^3 + \ldots\ldots b_nx_1^n$$

where,

y = dependent variable

$x_1 \ldots x_1^n$ = independent variables

$b_1 \ldots b_n$ = coefficients

$b_0$ = constant

To start working, the data is split into train and test sets respectively. The data is then transformed for Polynomial Regression by considering the data values and the degree. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used to calculate the prediction error rates and specify the performance of the model.

```
]: X_train_confirmed, X_test_confirmed, y_train_confirmed, y_test_confirmed =
   train_test_split(days_since_1_22, india_cases, test_size=0.2, random_state =

]: #Transform the data for polynomial regression
   poly = PolynomialFeatures(degree=3)
   poly_X_train_confirmed = poly.fit_transform(X_train_confirmed)
   poly_X_test_confirmed = poly.fit_transform(X_test_confirmed)
   poly_future_forecast = poly.fit_transform(future_forecast)

]: #Polynomial Regression
   linear_model = LinearRegression(normalize = True, fit_intercept = False)
   linear_model.fit(poly_X_train_confirmed, y_train_confirmed)
   test_linear_pred = linear_model.predict(poly_X_test_confirmed)
   linear_pred = linear_model.predict(poly_future_forecast)
   print('MAE:',mean_absolute_error(test_linear_pred, y_test_confirmed))
   print('MSE:',mean_squared_error(test_linear_pred, y_test_confirmed))

   MAE: 2185.7626758879287
   MSE: 6612932.5084639145
```

**Fig-3:** Implementation of Polynomial Regression

The predictions are made since the day 22 January 2020 and are visualized with the help of the graph below. The graph shows the prediction for the future 20 days in India.

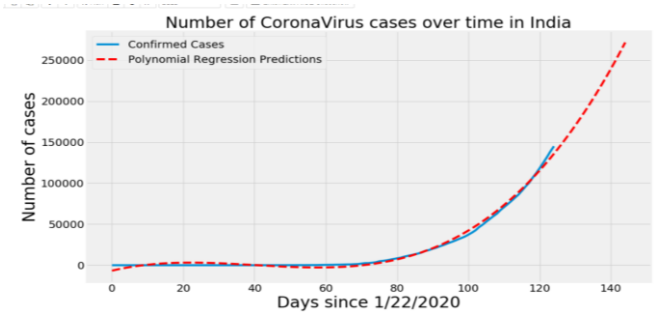


**Fig-4** Graph of Prediction using Polynomial Regression

In order to get the date wise prediction of count for the future 20 days, the data frame is created using pandas library and is shown in the form of a table. The table shows the prediction count for India in the upcoming 20 days according to the dates.

| | Date | Predicted number of cases in India by Polynomial Regression |
|---|---|---|
| 0 | 05/26/2020 | 141226.0 |
| 1 | 05/27/2020 | 146770.0 |
| 2 | 05/28/2020 | 152451.0 |
| 3 | 05/29/2020 | 158269.0 |
| 4 | 05/30/2020 | 164226.0 |
| 5 | 05/31/2020 | 170325.0 |
| 6 | 06/01/2020 | 176566.0 |
| 7 | 06/02/2020 | 182951.0 |
| 8 | 06/03/2020 | 189482.0 |
| 9 | 06/04/2020 | 196160.0 |
| 10 | 06/05/2020 | 202988.0 |
| 11 | 06/06/2020 | 209966.0 |
| 12 | 06/07/2020 | 217096.0 |
| 13 | 06/08/2020 | 224379.0 |
| 14 | 06/09/2020 | 231818.0 |
| 15 | 06/10/2020 | 239414.0 |
| 16 | 06/11/2020 | 247168.0 |
| 17 | 06/12/2020 | 255083.0 |
| 18 | 06/13/2020 | 263159.0 |
| 19 | 06/14/2020 | 271398.0 |

**Fig-5** Count for future 20 days

### 2.1.2 Support Vector Regression (SVR)

SVR has proven to be an effective tool in real value function estimation. It is used to predict a continuous variable. Other regression models often try to minimize the errors occurred while SVR tries to fit best in any threshold value. The SVM is trained with the SVM train function.

```
#SVM
svm_confirmed = SVR(shrinking=True, kernel='poly', gamma=0.01, epsilon=1,degree=5)
svm_confirmed.fit(X_train_confirmed, y_train_confirmed)
svm_pred = svm_confirmed.predict(future_forecast)
```

```
svm_test_pred = svm_confirmed.predict(X_test_confirmed)
plt.plot(y_test_confirmed)
plt.plot(svm_test_pred)
plt.legend('Test Data','SVM Predictions')
print('MAE:',mean_absolute_error(svm_test_pred, y_test_confirmed))
print('MSE:',mean_squared_error(svm_test_pred, y_test_confirmed))
```

MAE: 2816.121760878107
MSE: 14276221.447054649

**Fig-6** Implementation of SVR model

The important keys points in SVR are:

- Kernel is a function used in SVR to map the lower dimensional data points to higher dimensional data points. It is a crucial part of SVR.
- Hyper Plane is a line that is used to predict a continuous value.
- Boundary Lines are the two parallel lines drawn to the two sides of Support Vector with threshold error.
- Support Vector is the line from which distance is minimum from two boundary data points.

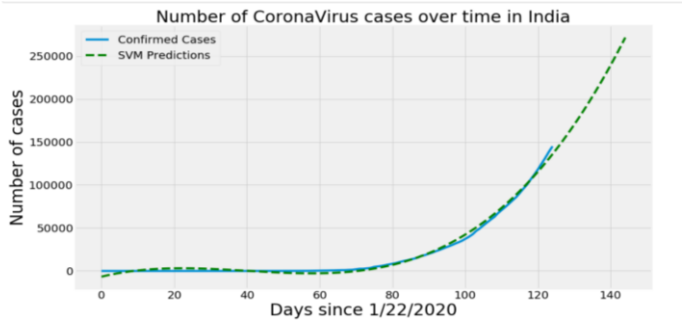


**Fig-7** Graph of prediction using SVR

| | Date | Predicted number of cases India by SVM |
|---|---|---|
| 0 | 05/26/2020 | 136912.0 |
| 1 | 05/27/2020 | 142514.0 |
| 2 | 05/28/2020 | 148297.0 |
| 3 | 05/29/2020 | 154265.0 |
| 4 | 05/30/2020 | 160423.0 |
| 5 | 05/31/2020 | 166774.0 |
| 6 | 06/01/2020 | 173324.0 |
| 7 | 06/02/2020 | 180077.0 |
| 8 | 06/03/2020 | 187038.0 |
| 9 | 06/04/2020 | 194211.0 |
| 10 | 06/05/2020 | 201602.0 |
| 11 | 06/06/2020 | 209215.0 |
| 12 | 06/07/2020 | 217055.0 |
| 13 | 06/08/2020 | 225127.0 |
| 14 | 06/09/2020 | 233437.0 |
| 15 | 06/10/2020 | 241989.0 |
| 16 | 06/11/2020 | 250790.0 |
| 17 | 06/12/2020 | 259843.0 |
| 18 | 06/13/2020 | 269155.0 |
| 19 | 06/14/2020 | 278732.0 |

**Fig-8** Count for future 20 days

## 3. CONCLUSIONS

COVID-19 was first encountered in Wuhan region, China and is still playing its game all over the world affecting human life, world trade, businesses and economy. The rate of the spread of the virus is much higher and so it becomes difficult to get the situation under control. Under such a crisis, Artificial Intelligence and Machine Learning techniques are playing a vital role across the world so as to analyze, predict, protect, track, diagnose, cure and create social control to fight against this pandemic. The prediction models built using Polynomial Regression and SVR would help predict the future count and to get the situation under control before it hits hard.

Machine Learning has proven to be a very promising and important tool to deal with this crisis. Data scientists across the globe are doing their best to build such strong models to help mankind and to give an end to this tragic attack.

## REFERENCES

[1] Upendra Kumar Trivedi and Rizwan Khan "Role of machine learning to predict outbreak of COVID-19 in India," Research Gate, April. 2020.

[2] Mohammad Mehran, Austin George, Umesh Yadav, R. Logeshwari, "Epidemic outbreak prediction using AI", Vol 7, Issue 4, April 2020.

[3] Wim Naude,"Artificial Intelligence against COVID-19: An early review", IZA Institute of Labor economics, April 2020.

[4] Rohan Taneja, Vaibhav, "Stock market prediction using regression", Vol 5, Issue 5, May 2018.

[5] Sakshi Deshmukh, Tashfin Ansari, Dr. Almas M.N Siddiqui, Aniket Kotgire, Dr. Gaikwad A. T., "An overview of detection of COVID-19 in medical imaging using machine learning, Vol 7, issue 4, April 2020.

[6] Joseph George and Ranjeesh R Chandran, "Comparison of regression models on covid19 cases", Vol 7, Issue 5, May 2020

[7] https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[8] https://www.datarobot.com/wiki/prediction-explanations/

## BIOGRAPHIES

**Name:** Richa Dinesh Tamhane, Student at Computer Science and Engineering, SVERI's College of Engineering, Pandharpur

**Name:** Sumeet Shivraj Mulge, Student at Computer Science and Engineering, SVERI's College of Engineering, Pandharpur