REGRESSION MODELS FOR PREDICTION OF COVID-19 CASES

Atanu Das

M.Tech in Information Technology (Data Science)

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

West Bengal, India
dasa90552@gmail.com

Abstract— In 2019, a very serious pandemic virus COVID -19 started spreading, which has globally infected over 31 million people. Many people were infected by the virus few were recovered but the death rate is on peek. In such a case, a complete understanding of the disease and its rate of spreading is vital. This infectious growth is needed to be analyzed to take preventive measures for the people against this pandemic. If the rate of spreading is calculated in an accurate manner, then the protection measures can be proposed easily by the government. To analyse the rate of spreading, Machine learning plays a vital role. Machine learning proved itself a prominent field to solve many complex real-world problems. This project illustrates how the machine learning algorithm predicts the number of upcoming cases with the help of historical data of Coronavirus. These algorithms will predict the upcoming number of newly infected cases, recoveries and death. The performance of these algorithms will be analyzed based on their prediction accuracy and an efficient approach will be proposed.

Keywords— ***Prediction, COVID-19, Regression, Linear, Support Vector Machine.***

I. INTRODUCTION

Coronavirus disease 2019 is a contagious infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China, and has since spread globally, evolving into an ongoing pandemic. Common symptoms include cough, fever, fatigue, breathlessness and loss of smell and taste. The World Health Organization (WHO) announced the outbreak of a Public Health Emergency of International Concern in January and a pandemic on March [2]. The COVID-19 has largely impacted on all the sectors like economy, education, healthcare, logistics and mental health of people.

The pandemic has caused severe global economic disruption and has led to the postponement or cancellation of many events. According to the World Trade Organization, the trade has been plunged due to the pandemic and is expected to fall between 15% and 30%. Many economic experts state that it might take 10 years to improve the economy to its normal state [3].

The WHO states that COVID-19 has impacted significantly in the health sector for non-communicable diseases such as cancer, Alzheimer's etc. Since there are no vaccines for this disease, it has become a humongous task and utmost priority

for the healthcare department to prevent the widespread of the disease [3].

With the help of predictive analysis and supervised learning, we can predict future cases which might be helpful for taking much better preventive measures and precautions. The proposed model is shown in Fig. 1. Here we have used 2 supervised machine learning models for the regression of the data. The data set after a series of visualization seems to be linear and hence we have used 2 basic regression models.

We have used Linear Regression as it is simple to implement and easier to interpret. Linear Regression tends to perform better when the current data is also linear.

We have also used SVR as it is one of the basic and simplest algorithms available for regression. One of the major advantages of SVR is the complexity of the model does not depend on the dimensions of the data.

II. LITERATURE SURVEY

After extensive research and survey, we have found a paper with similar kinds of work but more extensively. "COVID-19 Future Forecasting Using Supervised Machine Learning Models" a journal written by honourable professors [8], have used similar techniques and models, but with more research and experimentation. We will not be comparing our work with theirs as they have used fewer data compared to ours.
They have made use of many models such as LASSO Regression, Support Vector Machine, Linear Regression and Exponential Smoothing.

III. PROPOSED MODEL

In the proposed model there are two stages: training and evaluation. Before training, the dataset is pre-processed by removing null values and fields which are non-significant for this study. In the training stage, the model is trained and tested for prediction. The results are evaluated using 3 measures like MSE, R^2 and MAE [5].

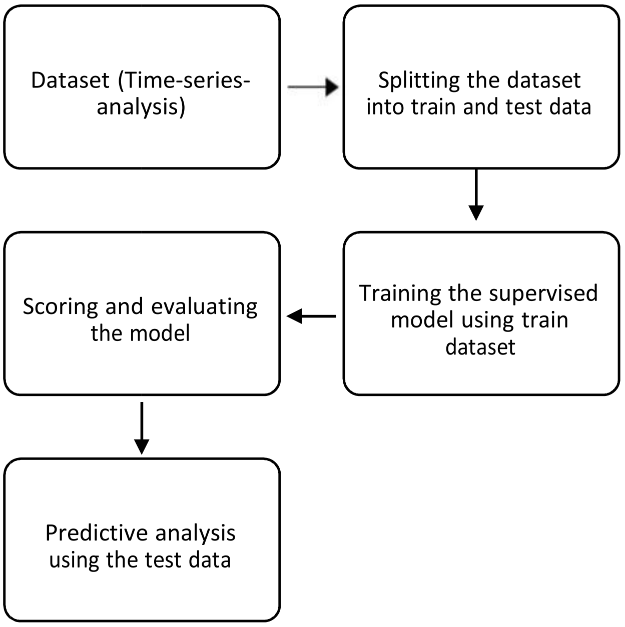

Dataset (Time-series-analysis)
Splitting the dataset into train and test data
Training the supervised model using train dataset
Scoring and evaluating the model
Predictive analysis using the test data

Figure 1: Proposed model

A. Data Acquisition and Selection

The Centre for Systems Science and Engineering (CSSE) is a research collection centre housed within the Department of Civil and Systems Engineering (CaSE) of John Hopkins University, has collected the data. They have released multiple forms of the dataset, and in this case, we have selected Time Series Dataset, which is updated every day. We have used the dataset collected from 01/22/20 to 06/22/20, which is available from their official GitHub website [4].

Table 1: Time Series data set of confirmed cases

Province/ State	Country/ Region	Lat	Long	1/22/20	1/23/20
NULL	Afghanistan	34	65	0	0
NULL	Albania	42.1533	20.1683	0	0
NULL	Algeria	28.0339	1.6596	0	0
NULL	Andorra	40.5063	1.5218	0	0
NULL	Angola	12.2027	17.8739	0	0

B. Data pre-processing

Pandas package is used for converting the CSV file into Data frame. Filtering of data is a manipulation and transforming method to fit into the requirements. Here we make separate data frames such as future_forecast_dates, unique_countries, etc.

Lat, Long, Province/State columns from the data frame are removed as they are less useful in prediction. We add the number of future dates required to the original dates in the data frame for future prediction. This is done by using the date/time package of python.

We used the train_test_split function in order to split the data into training and testing data sets using the sklearn.models package.

C. Selecting the model

Regression is a widely used Machine learning technique for prediction purposes. In this work, we used supervised learning for future prediction. Under supervised learning, two regression models, Support Vector Regression and Linear Regression are used for the conduction of experiments. The supervised learning algorithm is used because of two input factors (X) and an output factor (Y) which are utilized by the algorithm to learn the mapping from input to output. The objective is to train the model by mapping in a good manner that a new input data given to the model can predict the output factor (Y) [1].

Support Vector Regression:

It is a common application form of Support Vector Machine that supports linear and non-linear regressions. SVR requires training data X and Y which covers the domain of interest and is accompanied by solutions on that domain [10]. The SVR is a supervised learning machine derived directly from the Support Vector algorithm to estimate the function we use to generate the training set to reinforce some data. It is available in the Scikitlearn package of Python [11].

Linear Regression:

It is a commonly used type of predictive analysis and is used to foresee a numeric result given an arrangement of autonomous factors.[14] The overall idea is to find a relation between two variables by fitting a linear equation to the observed data. The linear regression model can be used by importing the sklearn.linear_models package in Python [6].

IV. TRAINING

Steps in training each model: Google Collaboratory is used to train the model as it is a free cloud service Jupyter Notebook and supports free GPU. It is research-oriented and does not require an environment setup. It supports many machine learning libraries which can be loaded easily without any dependencies on hardware [7].

The dataset is employed to validate the model.

A. *Training SVR*

Here in this paper by using the GridSearchCV present in the sklearn we can tune the hyperparameters. The best_search function is used to find the best hyperparameters. Then we train the model using the train data. After training the data we send the test data with future dates for the prediction.

B. *Training Linear Regression*

The linear model is trained with parameters normalize and fit_intercept set to TRUE. The data is trained with train data and the test data with future dates are used for prediction.

Comparison of both algorithms is studied based on the accuracy for the effective prediction.

V. EXPERIMENTS AND DISCUSSIONS

A. Dataset

The basic dataset is having 266 rows and 157 columns. The dataset has all the recorded positive cases dated from 01/22/20 to 06/22/20 based on countries. From the dataset, all the dates and the number of cases are extracted and made into separate data frames. The dataset is then split into train and test data by dividing 30:70, 50:50, 60:40,80:20 percentage.

For the actual prediction, the dataset is split between the ratio of 85:15 train and test dataset. The dataset is split in such a manner to experiment and observe how the model learns with different quantities of training set using the evaluation parameters.

B. Evaluation Parameters

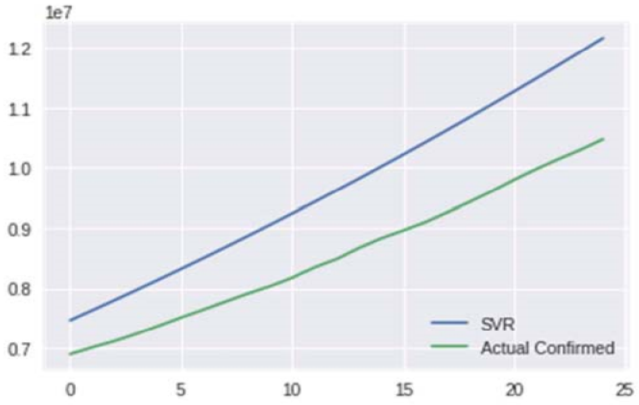
In this paper, the evaluation of the model's performance is measured in the terms of R-Squared(R^2), mean squared error(MSE) and mean absolute error(MAE) [8]. Here in this paper, we have mainly concentrated on the R^2 score. When the performance of the model is evaluated against the R^2 if the value is negative, it indicates that the model's performance is arbitrarily worse. And if the value is nearing to or is 1.0, the model is evaluated to be having the best performance [12].

C. Comparison of Models

SVR Prediction: The SVR model is trained and tested with different ratios of train and test datasets. GridSearchCV class and the estimator function is used to find the best hyperparameter for the model.

As mentioned above, the model is trained with different ratios of train and test datasets and performance is evaluated for each. The performance of the model can be visualized by plotting the predicted values and the actual values.

Figure 2: Cases confirmed by SVR vs Actual Confirmed



As shown in Fig. 2. the lines do not coincide and is predicting a higher value than supposed to, hence visualizing that the prediction done by SVR is not accurate. This can be proved by evaluating the performance of the model.

Table 2: Performance measure of SVR

Data Set Split	R2	MAE	MSE
30:70	-7.801137	2315545.839	8.32699E+12
50:50	-1.010918	2019287.999	4.54883E+12
60:40	0.8061677	885613.1922	1.39289E+12
80:20	0.2934247	1270539.112	1.85497E+12
85:15	0.3061489	1128135.156	1.3784E+12

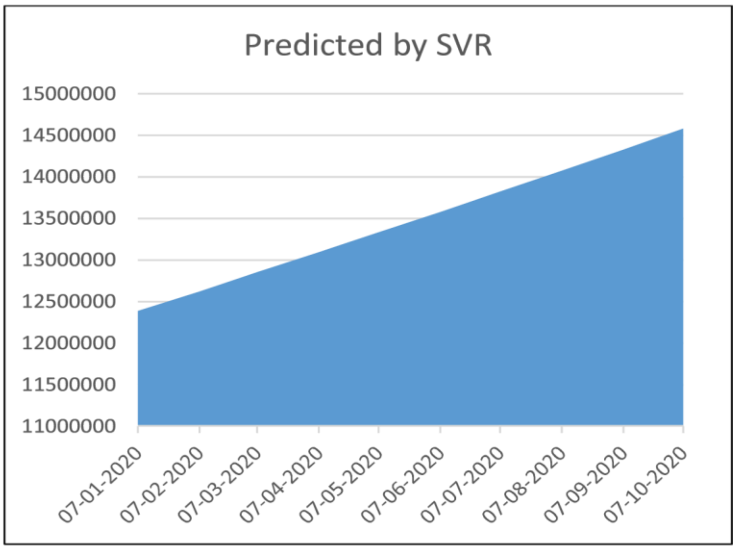
As highlighted in Table 2, the model performs with only 30% accuracy. It is also observed that as the ratio of the training set is increased the model slowly improves its performance.

The performance accuracy of the model can further be confirmed by comparing the predicted cases by SVR and the actual cases confirmed by the WHO [9].

Table 3: Time Series prediction made by SVR and Actual cases confirmed on particular dates

Date	Predicted by SVR	Actual cases confirmed by WHO
07-01-2020	12389159.28	10795162
07-02-2020	12621307.64	10974342
07-03-2020	12856385.35	11188120
07-04-2020	13094365.21	11383908
07-05-2020	13335265.04	11562295
07-06-2020	13579102.63	11734031
07-07-2020	13825895.85	11942118
07-08-2020	14075662.34	12156020
07-09-2020	14328420.06	12379660
07-10-2020	14584186.77	12616578

Figure 3: Cases predicted by SVR   Table 4: Difference between cases



predicted by SVR and Actual cases confirmed

Date	Actual Cases – SVR Predicted	Difference
07-01-2020	10795162-12389114.28	-15,93,952.28180
07-02-2020	10974342-12621307.64	-16,46,965.64
07-03-2020	11188120-12856385.35	-16,68,265.35
07-04-2020	11383908-13094365.21	-17,10,457.21
07-05-2020	11562295-13335265.04	-17,72,970.04
07-06-2020	11734031-13579102.63	-18,45,071.63
07-07-2020	11942118-13825895.8	-18,83,777.8
07-08-2020	12156020-14075662.34	-19,19,642.34
07-09-2020	12379660-14328420.06	-19,48,760.06
07-10-2020	12616578-14584186.77	-19,67,608.77

In Table 4, there is a large difference between the predicted values of SVR and the actual confirmed cases. Hence it can be declared that the SVR's performance in predicting is not very reliable.

Linear Regression: The Linear Regression model is trained and tested with parameters normalize and fit_intercept set to TRUE.

As mentioned above, the model is trained with different ratios of train and test datasets and performance is evaluated for each. The performance of the model can be visualized by plotting the predicted values and the actual values.

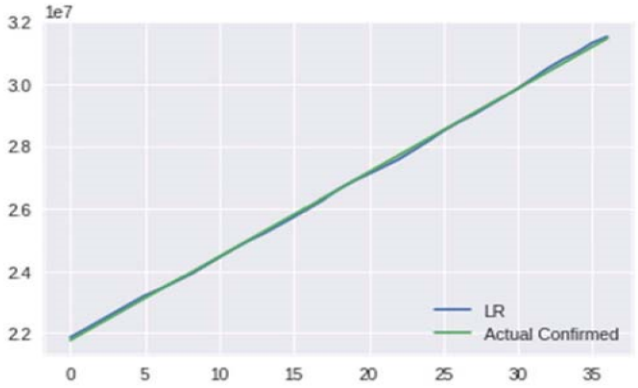


Figure 4: Cases confirmed by LR vs Actual Confirmed

As shown in Fig. 4., both the lines almost coincide with each other and is predicting values almost the same as it is supposed to, hence visualizing that the prediction done by the Linear Regression model is quite accurate. This can be proved by evaluating the performance of the model.

Table 5: Performance measure of LR

Data Set Split	R2	MAE	MSE
30:70	0.97755317	327024.633	1.58435E+11
50:50	0.98899298	192823.384	51664683589
60:40	0.98946368	161481.827	36408939271
80:20	0.99637816	57369.3079	4608462954
85:15	0.99632163	56956.1964	4317204498

As highlighted in Table 5, the model performs with almost 99% accuracy.

The performance accuracy of the model can be further confirmed by comparing the predicted cases and the actual cases confirmed by the WHO [9].

Table 6: Time Series prediction made by LR and Actual cases confirmed on particular dates.

Date	Predicted by LR	Actual cases confirmed by WHO
07-01-2020	10518411.13	10795162
07-02-2020	10668646.39	10974342
07-03-2020	10818881.65	11188120
07-04-2020	10969116.91	11383908
07-05-2020	11119352.17	11562295
07-06-2020	11269587.43	11734031
07-07-2020	11419822.69	11942118
07-08-2020	11570057.94	12156020
07-09-2020	11720293.2	12379660
07-10-2020	11870528.46	12616578

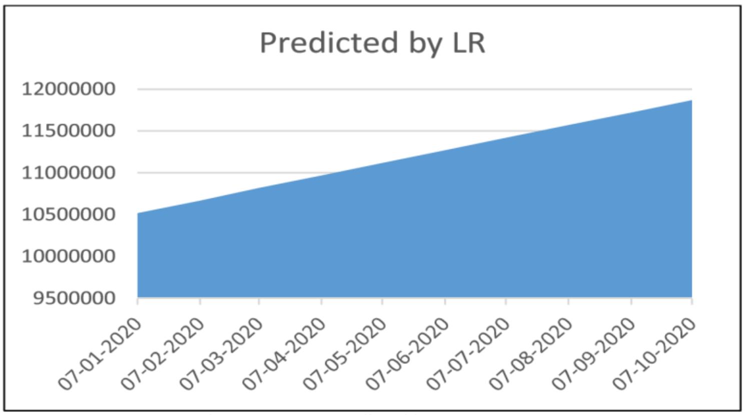


Figure 5: Cases predicted by LR

Table 7: Difference between cases predicted by LR and Actual cases confirmed.

Date	Actual cases – LR Prediction	Difference
07-01-2020	10795162 -10518411.13	2,76,750.87
07-02-2020	10974342-10668646.39	3,05,695.61
07-03-2020	11188120-10818881.65	3,69,238.35
07-04-2020	11383908-10969116.91	4,14,791.09
07-05-2020	11562295-11119352.17	4,42,942.83
07-06-2020	11734031-11269587.43	4,64,443.57

07-07-2020	11942118-11419822.69	5,22,295.31
07-08-2020	12156020-11570057.94	5,85,962.06
07-09-2020	12379660-11720293.2	6,59,366.8
07-10-2020	12616578-11870528.46	7,46,049.54

In Table 7, the difference between the predicted values of LR and the actual confirmed cases does not have a large difference when compared to SVR. It is evident that the LR's performance is quite reliable.

Both SVR and LR are used for prediction analysis using a time series data set [10] [12]. But the major disadvantage of SVR is that it cannot handle large data sets and hence its performance is less compared to LR which can be seen in the above tables, Table 4 and Table 7 [11].

VI. CONCLUSION

The machine learning algorithms such as linear regression, Gaussian processes, support vector regression gives accuracy in prediction, quickly and easy prediction of spreading virus so that the protective measure can be taken earlier. among these algorithms support vector regression results in better accuracy.

These models acquired remarkable accuracy in COVID-19 recognition. Bearing in mind these projected active results, the current estimate for COVID-19 containment needs to be reinforced or updated. Our framework could assist and protect healthcare professionals, government officials in making plans appropriate to cope with the influx of future COVID-19 patients.

REFERENCES

[1] Geddam Jaishankar Harshit, Rajkumar S, "A Review Paper on Cricket Predictions Using Various Machine Learning Algorithms and Comparision Among Them", 2018

[2] Covid-19 Pandemic,https://en.wikipedia.org/wiki/COVID-19_pandemic
Covid-19 Pandemic effects on world economy, https://www.wto.org/english/news_e/pres20_e/pr855_e.htm

[3] Covid-19 Pandemic effects on health sector, https://www.who.int/newsroom/detail/01-06-2020-covid-19-significantly-impacts-health-servicesfor-noncommunicable-diseases

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

[5] Documentation of R2 score by Scikit Learn,https://scikitlearn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[6] Documentation of Linear Regression by Scikit Learn, https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegressi on.html

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

[8] 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

[9] https://www.worldometers.info/coronavirus/worldwide-graphs/#totalcases, World confirmed cases counts by WHO

[10] 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.