**Final Project report:** **CSE 5095: Machine Learning for Time Series**

**Data Analysis**

Covid-19 Prediction Using Machine Learning Methods

*Soumyashree Sahoo (* [*Soumyashree.sahoo@uconn.edu*](mailto:Soumyashree.sahoo@uconn.edu)*) – PhD Student, First Year – Computer Science & Engineering, University of Connecticut*

**Abstract**

Coronavirus was detected in December 2019 in a bulk seafood shop in Wuhan, China. COVID-19 is a pandemic that has affected over 170 countries around the world. The number of infected and deceased patients has been increasing at an alarming rate in almost all the affected nations. The ongoing coronavirus disease (COVID-19) outbreak has infected over 38 million individuals and caused over 2 million deaths worldwide. This paper analyses COVID-19 data at worldwide basis by gathering data from several authentic sources. Various data visualization techniques are used to study world-wide confirmed cases, total deaths, total recovery numbers. Similar visualizations are performed for country wise confirmed cases, reported total deaths, and reported total recoveries. Time series forecasting techniques such as Polynomial Regression and Support Vector Regression are deployed to perform prediction using relevant confirmed cases, deaths, and recoveries in terms of global data. These predictions might help prepare against possible threats and consequences. This paper aims to comprehensively review role of Machine Learning models as one significant method in the arena of predicting and forecasting of COVID-19 pandemic. With the common data about confirmed, death and recovered cases across various countries for over the time length helps in anticipating and estimating the not so distant future. For extra assessment or future perspective, case definition and data combination must be kept up persistently.

**Keywords** – COVID-19, Forecasting, Machine learning, Polynomial Regression Technique(LR), Support Vector Machine technique(SVR), Long Short-Term Memories(LSTMs)

**1 Introduction**

Coronaviruses are a family of viruses that can cause illnesses such as the common cold, severe acute respiratory syndrome(SARS) and Middle East respiratory syndrome(MERS). In 2019, a new coronavirus was identified as the cause of a disease outbreak that originated in China. The virus is now known as the severe respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease it causes is called coronavirus disease 2019 (COVID-19). In March, the World Health Organization(WHO) declared the COVID-19 outbreak a pandemic, while the virus continues to spread. As on 4 May 2020, a total of 3,581,884 confirmed positive cases have been reported leading to 248,558 deaths. Moreover, various studies reported that the disease caused by CoV-2 is riskier for people with weak immune system. The elderly people and patients with life threatening diseases like cancer, diabetes, neurological conditions, coronary heart disease and HIV/AIDS are more vulnerable to severe effects of COVID-19. In the absence of any curative drug, the only solution is to slow down the spread by practicing social distancing to block the chain of spread of the virus.

There are lots of studies performed for the prediction of different diseases using machine learning techniques. In particular, the study [1] is focused on forecasting of COVID-19 confirmed cases using various forecasting models and study [4] is also focused on the forecast of COVID-19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively.

This study intends to apply the machine learning models simultaneously with the forecast of expected reachability of the COVID-19 over the nations using the real-time data from the John Hopkins dashboard. The dataset is retrieved from the official repository of John Hopkins University. The data consists of daily case reports and daily time series summary tables. In this study, I have selected time-series data in CSV format having three data frames for confirmed, death, and recovered cases of COVID-19 with six properties. For example, province/state, country/region, last update,confirmed,death, and recovered cases. The CSV data are available in GitHub repositories.

Figure 1 and 2 are the snapshots of the visual representation of confirmed cases across world and United States. This clearly presents how the COVID-19 pandemic has been spreading over countries throughout time. The blue dots show the location data corresponding daily confirmed cases in global map and United States map.

Fig. 1 Confirmed Cases global

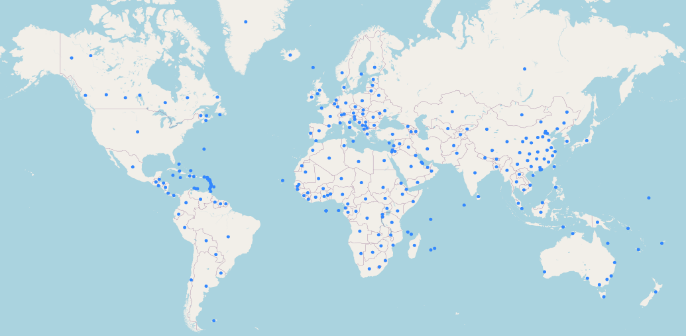
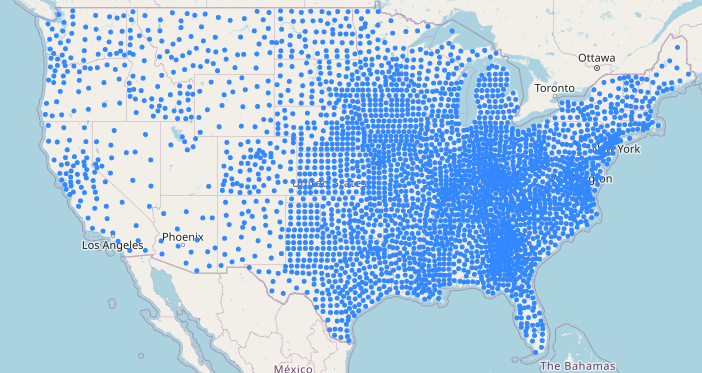


Fig. 2 Confirmed Cases United States



According to Jiang et al. [5] the fatality rate for this virus has been estimated to be 4.5% but for the age group 70-79 this has gone up to 8.0% while for those >80 it has been noted to be 14.8%. This has led to elderly persons above the age of 50 with underlying diseases like diabetes, Parkinson’s disease, and cardiovascular disease to be considered at the highest risk. Symptoms for this disease can take 2-14 days to appear and can range from fever, cough, shortness of breath to pneumonia, kidney failure and even death [6].

There have been developments of various forecasting techniques, inculcated assisting various design strategies. Recent study has been done for forecasting of COVID-19 based on various parameters such as environmental factors, incubation period etc. on datasets collected from big data accessed from WHO/National databases and data from social media[1]. Nowadays machine learning techniques are used worldwide for predictions due to its accuracy. Researchers have done predictions based on datasets that are available and used the best machine learning model as per the dataset. Kumar and Hembram [7] presented a model based on the Logistic equation, Weibull equation, and the Hill equation to find infection rates in China and Italy. In my study LR,SVR and LSTMs machine learning models are applied on John Hopkins University dataset having separate instances of daily confirmed cases, death cases and recovered cases to check the forecasted cases for future 10 days.

**2 Related Work**

SOHINI , Sareeta and Garima[2] have analyzed COVID19 data initially at a global level and then drilled down to the scenario obtained in India. They have used Time Series Forecasting techniques including Machine Learning models like LR,SVR, Polynomial Regression and Deep Learning Forecasting Model like LSTMs are deployed to study the probable hike in cases and in near future. In [3], an analysis is done on COVID-19 datasets to understand which age group is mostly affected in this pandemic. Different prediction algorithms such as random forest regressor and random forest classifier has outperformed the other machine learning models like Support Vector Machine, KNN classifier etc. As per study [8], outbreak of COVID-19 was analyzed for India till 30th March 2020 and predictions have been made for the number of cases for the next 15 days. SEIR model and Regression model have been used for predictions based on the data collected from John Hopkins University repository in the time period of 30th January 2020 to 30th March 2020. The performance of these models was evaluated by RMSLE and achieved 1.52 for SEIR model and 1.75 for the regression model achieved 2.01.The value of R0 score which is the speared of the disease was calculated to be 2.02. As per this study the expected cases were predicted to rise between 5000-6000 in the next 15 days. They predicted this study may help the government and doctors in preparing their plans for the next two weeks.

**3 Methods and Materials**

Linear Regression(LR) is a machine learning model that assumes a linear relationship between input variables and single output variable[2]. More specifically, an output variable can be calculated from a linear combination of the input variables. If the goal is prediction, forecasting or error reduction, the LR model can be used to fit a predictive model to observed dataset of values of the response and explanatory variables. LR shows the connection between two variables by fitting a straight condition to based information. One variable is viewed as an independent and the other is viewed as a dependent. An LR1 line has a condition of the structure:

(1)

Here x is the independent and Y is the dependent variable. is the bias and is the weight in the regression equation.

Polynomial regression is a special case of linear regression where a polynomial equation is fitted on the data with a curvilinear relationship between the target variable and the input variables. In a curvilinear relationship, the value of the target variable changes in a non-uniform manner with respect to the predictor.

The linear equation can be used to represent linear relationship, but in polynomial regression, we have a polynomial equation of degree n represented as:

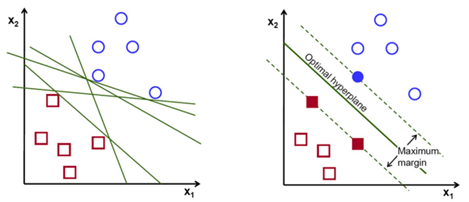
(2)

Here x is the independent and Y is the dependent variable. is the bias. ,,…., are the weights in the polynomial regression equation. The number of higher order terms increases with increasing values of n, by making equation more complicated.

Support vector machines (SVM)  is a supervised learning algorithm, which is used for classification and regression problems. SVR is based on the same principles as SVM for classification ,to find a hyperplane in a d-dimensional space that uniquely classifies the data points. SVR uses a non-parametric technique, which means, the output from the SVR model does not depend on distributions of the dependent and independent variables. SVR technique is basically dependent on kernel functions, which allows for the construction of a non-linear model without changing the explanatory variables, which helps in better interpretation of the resultant model. In these algorithms, a hyperplane is found that separates the different features. The produced model by SVM does not depend on the training points that lie outside the margin but instead depends on a subset of the training data as the cost function. Similarly, in SVR, support vectors find the closest data points and the actual function represented by them. We get closest to the actual curve if the distance between the support vectors to the regressed curve is maximum.

A hyperplane is a function that classifies the points in a higher dimension or other words hyperplanes are the boundaries that help in the classification of the data points. If the margin for any hyperplane is maximum, then that hyperplane is the optimal hyperplane. The points which are closest to hyperplane are called support vector points and the distance of the vectors from the hyperplane are called the margins, as shown in Figure SVR.

**Fig SVR**



The study in [9] states how LSTMs have attracted a lot of interest recently. Long Short-Term Memory networks are a special kind of artificial neural networks (specifically recurrent neural networks) capable of learning long-term dependencies from long sequences of observations. The key idea behind LSTM is we want some information to persist over time. Because of its complex architecture consisting of several gates, LSTMs can remember values over time and can effectively regulate the flow of information. LSTM exploits long term dependencies using cell states , sometimes referred to as a conveyor belt, that runs through the different cells, enabling the persistence of information from one memory block to the next. The flow of information is also controlled by gates, which determine whether values are added to the cell state or discarded. There are 3 types of gates: input , output , and forget gate . These gates consist of nonlinear sigmoid function, which given the previous hidden state and input x, it returns a value from 0 (discard value) to 1 (keep value). Given input ,…., and ,…. , the following equations determine unit activations:

=

=

=

=

=

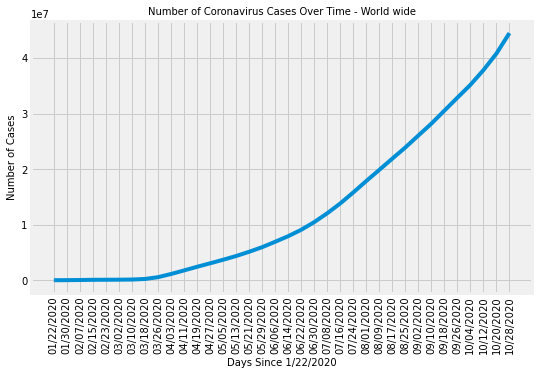
In recent studies[10], researchers have compared the performance of SVR and LSTM on time series forecasting using ecommerce data, concluding that the multivariate LSTM has higher accuracy than multivariate SVR.

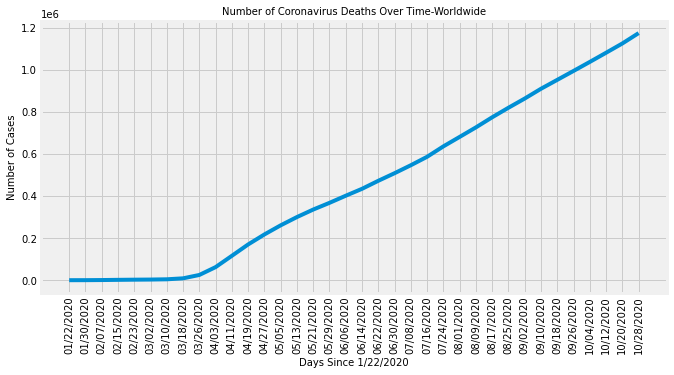
**4 Experimental Results**

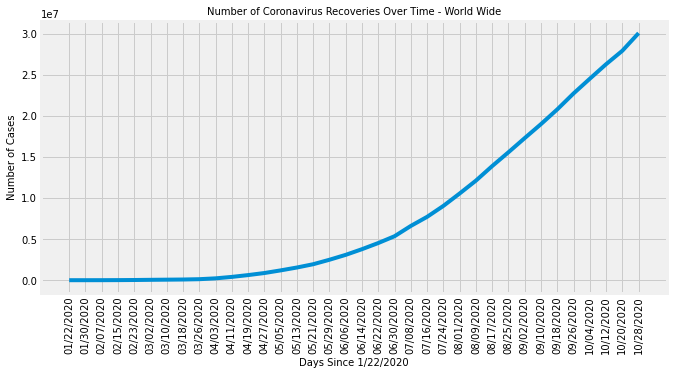
The structure of dataset is based on study dates starting January 22nd, confirmed, deaths, recovered cases extracted from the global databases of John Hopkins Coronavirus Resource Center. The current worldwide cases and country wise datasets presents how the case counts has been changing since the origination of this virus. As mentioned by WHO, right now many countries are in the second and even in third phase indicating very few or more cases.

In this paper, dataset has been analyzed to determine a summary of world cases, total deaths, mortality rate, recovery rate, total recovered cases and total active cases counts. Country wise confirmed case count, death counts and recovered case counts are revealed and are user for analyzing status of COVID-19 worldwide and in USA. Figure-4 provided the visualization of dataset in terms of total number of cases, total number of deaths and total number of recoveries over time.

**Fig 4 Number of COVID cases, deaths, and recoveries over time**





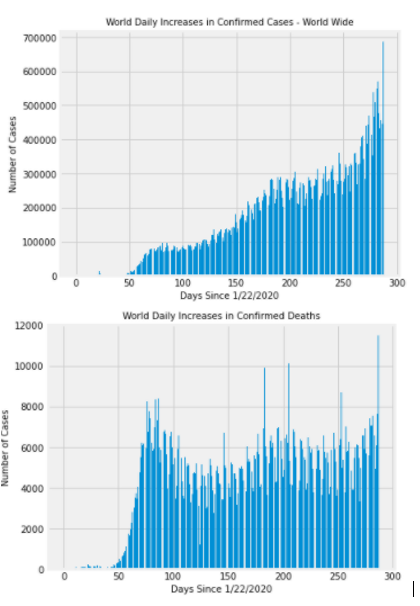


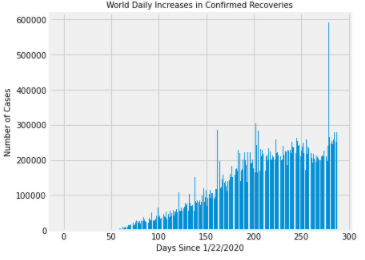
In illustration it can be understood that there has been significant rise of infection count along with time. Number of death cases was diminished a little after May but gradually increasing afterwards. Along with these recoveries also have steadily growing. The dataset revolves around the confirmed, recovered, and death of cases because of the COVID-19 outbreak over the time frame of around 10 months. According to Figure 5, there has been considerable increase of daily increase of confirmed cases and deaths worldwide. The infections have been increasing since September while slowing down the recoveries. The percentage of deaths due to pneumonia, influenza, and COVID-19 (PIC) remained at approximately equal levels from mid-September through mid-October. Both COVID-19 related hospitalizations and mortality for the most recent weeks may increase as more data are received. USA has been tremendously impacted in terms of number of confirmed cases and number of death counts. Recovery analysis shows that India have a great recovery rate in comparison to Brazil and USA. Appreciably the date attribute is holding a higher level of importance and that is reason globally the measures have been taken for social distancing. Across the globe, leaders of the nation are carrying out various trial and error methods to combat the seriousness of the disease.

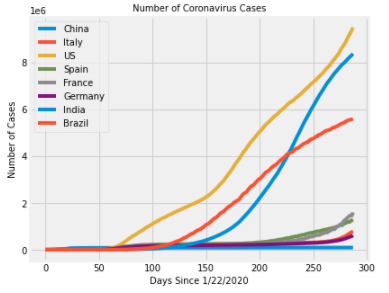
Prediction gives pertinent and consistent input about the past, present, and future happenings with certain statistical and scientific approaches. Helps in string decision making in all perspectives. Broadly classified into qualitative and quantitative approaches. Steps involved in forecasting is the deciding factor of the task. Initial understanding of the problem with complete analysis, making a strong foundation, collecting data based on the previous two steps followed by future estimation. Comparison between actual and estimated with follow up actions. Various applications like economic and sales predictions, budget, census and stock market analysis, yield projections and many more fields.

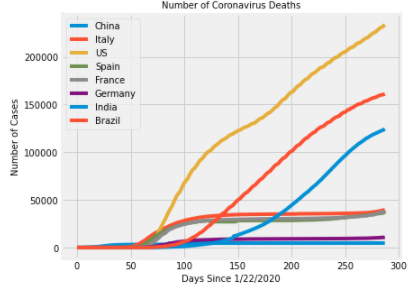
Machine learning models play an essential role in the pandemic investigation and prediction. It also helps for exposing epidemic patterns. Eventually, an immediate response could be prepared to prevent the spear of the virus. Few machine learning models such as Linear Regression, Support vector regression and Long short-term memory are utilized to recognize collective behavior together with the prediction of expected spread of the COVID-19 across the society by analyzing real time data from John Hopkins Dashboard.

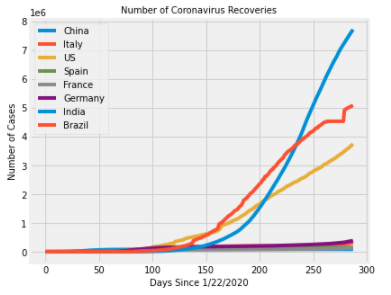
**Fig 5 Daily increase of COVID cases, deaths, and recoveries over time**







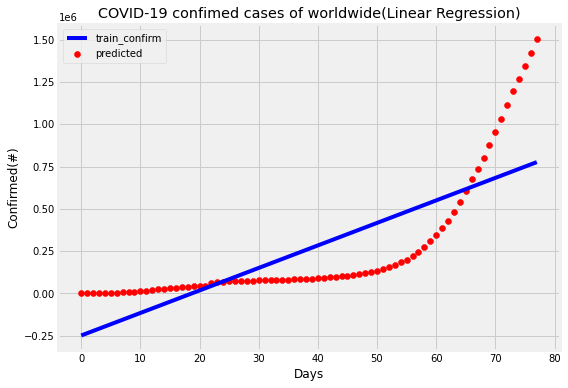




**5 Dataset Preparation and Evaluation**

Preparation of dataset begins by splitting the data for training, validation, and testing purpose. Pandas dataframe is used for storing confirmed, deaths and recovered csv files in form of dataframe. One index column is separately created for keeping track of dates starting Jan 22nd. Days since Jan 22nd , which is the start date is used for creating a set of values in form of NumPy array and then reshaped for prediction purpose. World confirmed ,deaths and recovered cumulative case counts for each day are collected and reshaped. Future days for prediction are determined based on incrementing order of days in terms of study dates. Based on analysis using above data it can be derived that cumulative case counts worldwide is less in comparison to number of confirmed cases and recovery count. To fit the linear regression model two parameters are selected considering dates and number of confirmed cases from the training, validation, and testing sets. In the verge of applying Linear regression model by thinking the continuous structure of data, a visualization is shown in Figure 6. After applying the linear regression, it can be seen that

**Fig 6 Confirmed cases prediction using Linear regression**



the predicted line does not fit with the independent variables. Which means that the line is unable to capture the patterns in the dataset and it may lead to under fitting situation.

Evaluating the model’s accuracy is an essential part of the process in creating models to describe how well the model is performing in its predictions. As the r2 score is closely related to the mean square error(MSE) and it clearly presents variance in the prediction that is predictable from the given dataset, hence it is considered as the major evaluation metrics for prediction. Other metrics such as MAE(Mean absolute error) and RMSE(root mean square errors) are also found out. For confirmation, the r2 score of the model is predicted as 0.61. R2 score is 61%. It means that 61% of variance is covered by the model. Now, to overcome the under-fitting problem, I must increase the complexity of the model. Here, I have used Polynomial linear regression technique because it best fit for the prediction of how diseases spread across the territory or the world. Polynomial linear regression is utilized to predict future confirmed cases. Polynomial regression is a form of linear regression in which the relationship between predictive values and study data is modelled as an nth degree polynomial. In our problem, a degree of 10 is considered. I tried to find out relationships to hypothesize the result to be curvilinear. Study data is transformed in terms of polynomial value and then model is fitted using transformed independent values and to be predicted values. In the same process I try to predict cases using polynomial converted future dates.

Fig 7 **Confirmed cases prediction using Polynomial Regression**

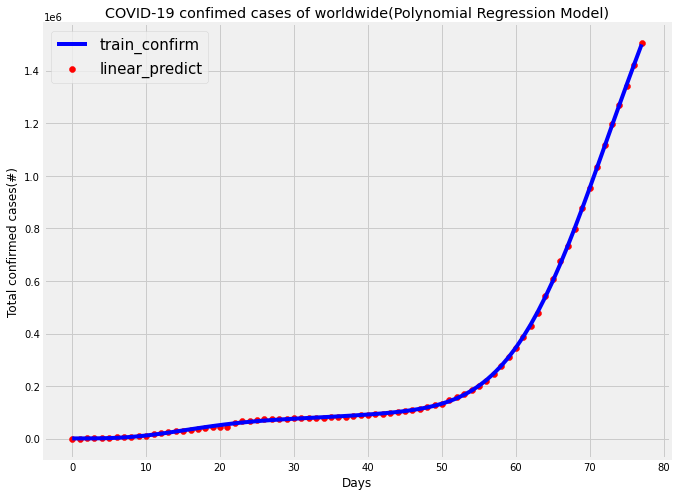
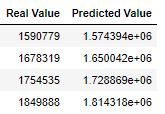


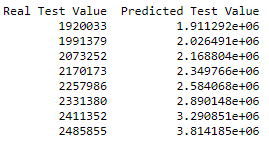
Figure 7 shows how the polynomial line connects all points and predicts all values almost covering all data. While trying to validate the model with few data points, the confirmed cases of the data is predicted as in Table 1.

**Table 1 Predicted Confirmed cases values using validation data**



For confirmation, the r2 score of the model is predicted as 0.91. R2 score is 91%. It means that 91% of variance is covered by the model and can be considered to have achieved considerable accuracy and more correlated in comparison to linear regression model. Now to apply the accurate model for the test data, the predicted data is found out as in Table 2.

**Table 2 Predicted Confirmed cases values using test data**

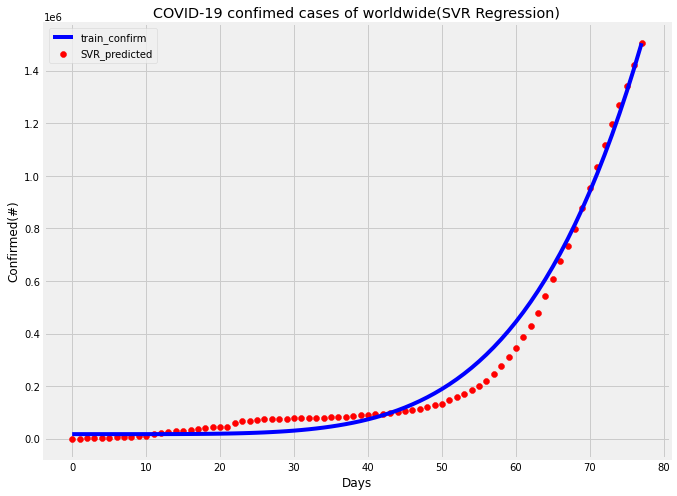


The support vector regression model is linear model for prediction problems and can solve both linear and nonlinear problems. The algorithm creates a line or hyperplane which separates data into classes. Perfecting this model to our problem, the original data is to be applied to the model to predict the future cases. The generalized equation for hyperplane could be represented as , where w is the weights and b is the intercept at X=0. The margin of tolerance is represented by epsilon . SVR regression model is imported from SVM class of sklearn library. The regressor is fit on the training dataset and model parameters chosen for analysis is as shown below.

SVR(shrinking=True, kernel='poly', gamma=0.01, epsilon=1,degree=5, C=0.1)

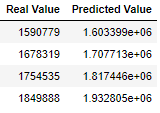
SVR model is fitted using dates index values and confirmed cases training dataset for prediction. The scatter plot function and line plot are used from matplotlib to show prediction using actual data and predicted data. In Figure 8, it’s clear that most of the prediction using SVR regression model aren’t correctly correlated with actual dataset.

**Fig 8** **Confirmed cases prediction using Support Vector Regression**

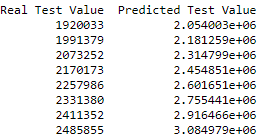


After applying the SVR regression, the predicted line does not fit with the independent variables. It means that the line is unable to capture the patterns in the data. This may be an example of under-fitting. For confirmation, the r2 score of the model is predicted as 0.98. It means that 98% of variance is covered by the model and can be considered for predicting future forecast values.

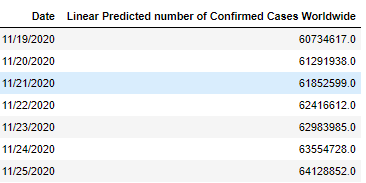
**Table 3 Predicted Confirmed cases values using validation data**

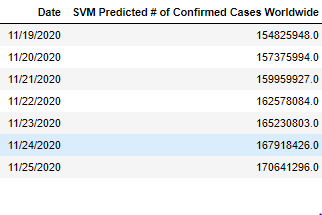


**Table 4 Predicted Confirmed cases values using test data**



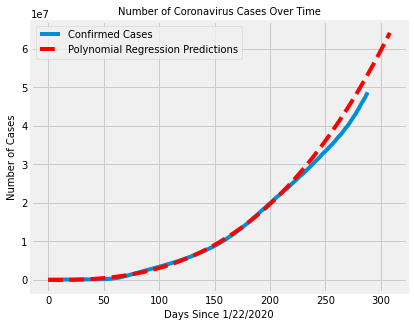
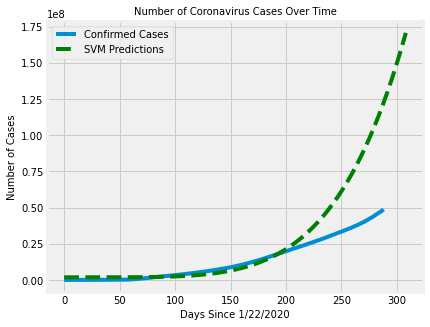
To understand cumulative effect of prediction on number of confirmed cases, the accurate model is applied on validation and test data and values are shown in Table 3 and 4. Prediction using validation and test data presents increasing trend of cases throughout time. To find out current prediction with date starting November 11th, future confirmed cases are predicted for next 7 days using models created so far. This helps to visualize the cumulative cases increasing trend in future 7 days. Table 5 shows predicted case counts using LR and SVR.

**Table 5 Future Predicted Confirmed Cases using Linear Regression and SVR**



To better understand behavior of linear and support vector models that is being created, future forecasting of COIVD-19 cases is done. In Figure 9 the increasing trend of confirmed cases can be visualized. Results show that model performed better in determining future cases by Polynomial regression model than Support Vector Regression. In Polynomial Regression model, increasing trend is more consistent with prediction line using actual non future data.

**Fig 9 Number of COVID-19 cases future prediction using LR and SVR**

**6 Conclusion**

Machine Learning models help in the decision-making process based on the past data with the data analytics and data mining perspectives. The Coronavirus data is huge in size and gathering information and retrieving interesting pattern out of the cumulative data is a challenging task. With the current data about confirmed , recovered and death across worldwide for over time durations helps in predicting and forecasting in near future. The accuracy of model could be increased by introducing other factors such as age of person, environmental aspects, gender, and underlying health conditions to prevent the propagation of COVID-19. According to latest reports, it is very clear that to handle this situation is by social distancing or quarantine and isolation. By noticing the predicted values and matching with cases from John Hopkins University data we can conclude that the Polynomial Regression method is giving good prediction results than that of the LR and SVR method. In future we can work with some deep learning methods for forecasting time series data for getting better predictions.

**References**

1. Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. [SN COMPUT SCI](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289234/). 2020; 1(4): 197. Published online 2020 Jun 11. doi: [10.1007/s42979-020-00209-9](https://dx.doi.org/10.1007%2Fs42979-020-00209-9)
2. Covid-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms. SOHINI Sengupta, Sareeta Mugde, Garima Sharma doi: <https://doi.org/10.1101/2020.06.25.20140004>
3. Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms, Volume 8. No. 5, May 2020 International Journal of Emerging Trends in Engineering Research Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter117852020.pdf <https://doi.org/10.30534/ijeter/2020/117852020>
4. G. Grasselli, A. Pesenti, and M. Cecconi, ‘‘Critical care utilization for the COVID-19 outbreak in lombardy, italy: Early experience and forecast during an emergency response,’’ JAMA, vol. 323, no. 16, p. 1545, Apr. 2020.
5. Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C.W. and Xia, Z., 2020. Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). Journal of General Internal Medicine, pp.1-5.
6. World Health Organization, 2020. Coronavirus disease 2019 ( COVID19): situation report, 67.
7. Kumar J, Hembram KPSS. Epidemiological study of novel coronavirus (COVID-19). 2020 arXiv preprint [https://arXiv:2003.11376](http://arxiv.org/abs/2003.11376).
8. SEIR and Regression Model based COVID-19 outbreak predictions in India, Gaurav Pandey , Poonam Chaudharya , Rajan Gupta , Saibal Pal.
9. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. [Ralf C. Staudemeyer](https://arxiv.org/search/cs?searchtype=author&query=Staudemeyer%2C+R+C), [Eric Rothstein Morris](https://arxiv.org/search/cs?searchtype=author&query=Morris%2C+E+R).
10. E-commerce Time Series Forecasting using LSTM Neural Network and Support Vector Regression. Ghassen Chinti,Houda Bakir,Hedi Zaher. [BDIOT2017: Proceedings of the International Conference on Big Data and Internet of Thing](https://dl.acm.org/doi/proceedings/10.1145/3175684) December 2017 Pages 80–84<https://doi.org/10.1145/3175684.3175695>