**Regression Analysis to Predict the no. of Upcoming COVID-19 Cases in India**

1. **ABSTRACT**

The **COVID-19** (Coronavirus disease 2019), also known as the **coronavirus**, is an ongoing worldwide pandemic caused by a novel Coronavirus, it is also known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). It is the illness, which was first identified amid an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China, in December 2019. Now, this outbreak has been declared as **Public Health Emergency of International Concern**, by World Health Organization (WHO) on 30 January 2020, and a pandemic, on 11 March. Currently there is no known vaccine found(though there are more than 100 candidate vaccines which are in various stages of trials) or any specific antiviral treatment. Primary treatment is symptomatic and supportive therapy. With the proposed model, we ought to present the trajectory based on the rising number of COVID-19 cases in the coming months. Thus, the proposed model, described here considers several important parameters that are extremely useful in preventing the spread of COVID-19 and forecasts the upcoming number of cases.At the same time, various Regression models are analyzed based on their performance on this epidemic situation, and the relevant analysis has been developed.

INDEX TERMS: *Regression, COVID-19, Corona Virus, India, Prediction model, Polynomial Regression Model, Decision Tree, Random Forest Regression, Support Vector Regression, Machine Learning, Predictions, Forecasting*

1. **INTRODUCTION**

At the end of 2019, the new coronavirus (COVID-19) outbreak spread widely in China, now infecting a large number of populations all over the world. At present, all countries are trying hard to control this domestic outbreak, while the illness is spreading rapidly in other areas, thereby making it a “GLOBAL PANDEMIC”. Currently, as no vaccine is known for cure of the illness, all countries are trying to effectively reduce the spread of the pandemic by following the guidelines given by WHO and implementing travel restrictions, imposing lockdowns, workplace hazard controls, and facility closures etc, in the respective nations. The pandemic has globally caused issues like social and economic disruption, this also including the biggest global recession since the Great Depression. Due to this pandemic that spreads so fast, various cultural events or sports events are either cancelled or postponed worldwide. Due to uncertainness of the future policies and panic, millions of people around the world started storing food and this lead to panic buying , shortage of food in markets, this severely affected those people who cannot really afford to buy food in large quantity. Though on the greener side some good news came up like decreased emissions of pollutants and greenhouse gases. Schools, colleges, universities, and even the govt. offices have been closed temporarily  either on a local basis or nationwide in more than 172 countries, this has affected about 98.5 percent population of student around the world. Other than this misinformation also played a role in increasing the panic or rumors about the virus. In such mixed situation it is very much necessary that policy makers should have access to correct information.

Many countries and places have also worked to increase testing capacity and trace contacts of infected persons. The term Cases refer to the total number of people who have been tested for COVID-19 virus, and also test result has been confirmed positive, though there are asymptomatic patients also who haven’t shown any symptoms of COVID-19 and leaving a normal life so such cases are very hard to detect. Thus, the continuous check is being maintained across each nation in order to keep track of the spread of illness. The government at various nations, along with various authorities at global level, are inviting people to volunteer and come up with solutions, suggestions, technologies and even prepare medicines to help the mankind in this need of the hour. Thus, with this description, we ought to propose model which would predict the trajectory on the rising number of COVID-19 cases in the coming months, in India, while at the same time taking into consideration several important parameters and facts identified by health organizations to estimate an approximation about the number of cases upcoming in following months, and effectively control the wide spread of COVID-19.At the same time, various Regression models are analyzed based on their performance on data of this epidemic situation, and the relevant analysis has been developed.

1. **RELATED WORK**
   1. <https://www.researchgate.net/publication/260015447_A_Study_on_Multiple_Linear_Regression_Analysis> [1]

In this paper , paper discusses the regression models, regression can be linear or it can be multiple linear regression. In multiple linear regression , there are more than one independent linear variables. Multiple linear regression can be think of as the combination of more than 1 linear regression. This type of regression is also called multivariate regression. The paper also discusses the procedure that should be followed for example first the missing values must be dealt then normalization should be done then independent and dependent variables must be separated and so like this.

* 1. <https://www.tandfonline.com/doi/full/10.1080/16549716.2020.1760490> [2]

This paper studies and develops a correlation between the COVID-19 virus, number of cases and the ecological factors. This study has been done for places in China. The reason mentioned is that since the epicenter and origin of virus took place here. Here the paper mentions that virus has an incubation period of about 5-7 days. And based on this analysis the proposed system in this project also gives an incubation period of 7 days. This paper claims that there is an association between virus and the weather condition but also warns that this does not mean that virus will go away in summer season, this paper fails to predict or warn about how the cases will rise in different environment and does not show a clear picture. Also its worth to not that paper does not discuss about those countries where there is a warm environment but still virus has infected a lot of people. So this project does not take environment as an attribute in the dataset since virus can manipulate itself very quickly according to the environment and also keeping in mind that most of the transmission occurs due to lack of social distancing.

* 1. [https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230405 [3](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230405%20%5b3)]

This paper also predicts the number of cases three weeks ahead. The study is based on the data provided by Heibi provenance, China. This paper has used SIRD (Susceptible-Infectious-Recovered-Dead) mathematical model. This model is used to find the Reproduction-number (Ro) for this virus. The reproduction number can be described as the number of infected people that can generated by a single infected person in a scenario where all the people are susceptible to infection. According to the paper this reproduction number is greater than to 2 for this virus. To calculate this reproduction number the author assumes that total number of actual infected people will be higher than the reported cases and therefore takes total cases 20 times of reported cases and recovered cases 40 times of reported recovered cases and keeping the total deaths unchanged. Based on this the paper predicts about 1,80k (upper bound) and 45k (lower bound) number of cases by Feb 29,2020. Now this model is good to predict the range in a scenario where the cases are coming from a single region and therefore government can really press hard to limit the number of cases, and this where this model assumes that things will not change(majorly policies). So the Ro values will definitely change if there is policy shift.

* 1. <https://www.sciencedirect.com/science/article/pii/S0048969720323779> [4]

In this paper, the effect of weather on the outbreak and spread of COVID-19 has been analyzed and verified. The input data consist of the information statistics on COVID-19 in 50 states of US along with the corresponding weather information with includes parameters like Temperature (T) and Absolute Humidity (AH). Using this input data, trends of spread of COVID-19 along with the trends in weather conditions are analyzed during the training process. The resultant model is then used to predict the trends of the outbreak of COVID disease in a country which is more vulnerable to weather changes like India. As a result, some vulnerable ranges have been identified for the weather parameters and validated on different time intervals.

* 1. <https://www.researchgate.net/publication/340394777_SEIR_and_Regression_Model_based_COVID-19_outbreak_predictions_in_India>

In this paper, the outbreak of COVID-19 pandemic has been discussed and analyzed. SEIR and Regression models are developed based on the COVID-19 statics across a country like India. The obtained model is then used to predict the upcoming number of COVID-19 cases for the next two weeks. Along with the prediction, the performances of the SEIR and Regression model have been evaluated in detail based on the comparison of several parameters such as error rate, etc. this model evaluation is done using RMSLE. Though the model can be used to predict upcoming number of cases all over the world, the primary focus of the author is to analyze the data of India ( 17% population all over the world) currently and make the future predictions for this nation.

1. **PROPOSED SYSTEM**
   1. **Data Set Description**

Since this pandemic is very new to this world, it is very much necessary to collaborate and build some important datasets. Governments around the world are providing huge amount of data so that the programmers around the world could come up with some good models that will help them to look into the future and based on that make their policies now. Since this project aims to provide the forecast of number of COVID-19 cases in India so the project uses only data that is provided by the Indian Government. Its is worth noting that the dataset used in this project is made and maintained by the contributors of this project only and is not available anywhere. While choosing the parameters of the dataset , there are two approaches, first approach is that in general what are the factors that impact the rise of cases in communicable diseases that spread this fast ? and another approach is how government can effect this ? So in the first approach we came up with

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| dayafter31dec | TotalCases | Daily rise in cases | Daily Recovery | %increaseInInfection | Total deceased | Daily deceased |

**dayafter31dec :-** This column counts the number of days passed from the date 31 dec, 2019 when the China officially admitted the COVID-19 as a disease that can transmit human to human and also acknowledged its existence. It was taken as an attribute because transmission rate increases with every passing day.

**TotalCases : -** As the name suggest this attribute keeps track of the total number of cases in India. This is selected as a attribute because of two reasons first reason is that the project also forecast the total cases after 10 days from the present date and second reason is that these number of cases are responsible for further transmission of COVID-19 in India.

**DailyRiseInCases :** This gives an overview of how the cases are increasing on daily bases. These cases are the result of transmission that occurred some few days back and these cases will give rise to new cases. So it is an important feature.

**DailyRecovery :-** Now this may not impact the upcoming cases in India directly but they do impact the government policies. Because if the recovery is increasing then the government will also promptly trace and test.

**%increaseInInfection :-** This attribute tracks the percentage increase in the infection rate in India. It is known that if this factor is low then daily cases will decrease but if this keeps on increasing then daily cases will also increase.

**Total deceased :-** This attribute also drives the government policies as some government tries to hide the total deaths so that they seemed to be handling the disease very well.

**Daily deceased :-** This attribute tracks the daily deaths, also drives the government policies.

From the second approach we come with the following attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TotalTests | %increaseInTest | %NewInfection InDailyTests | LockDownUnlockPhases | PreviousPolicy BufferTime(7days) |

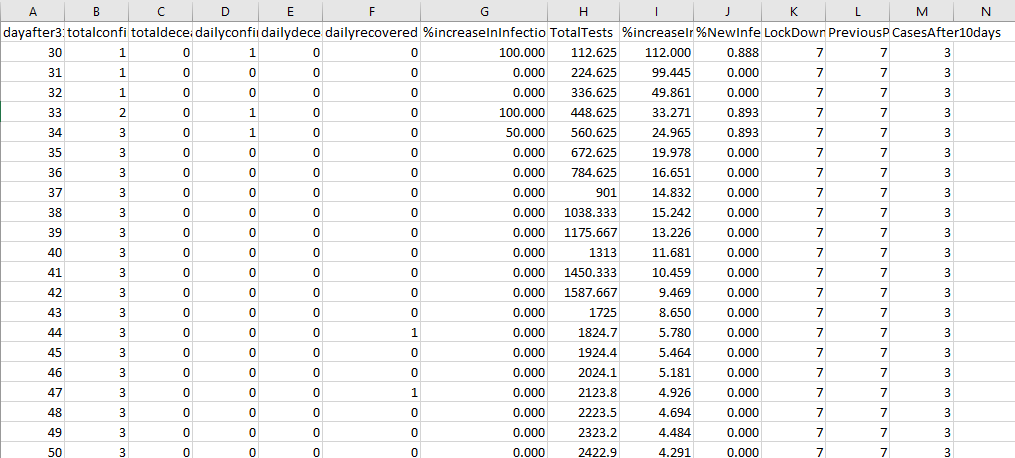
**TotalTests :-** This is government driven policy. Some government may decide to stop the testing or some government may decide to test more and more. In an infectious disease like when there is no medicine, the testing is the only way to tackle the situation. As testing increasing cases will rise but transmission will definitely decrease leading to less cases in the future.

**%increaseInTest :-** This tracks the percentage increase in overall testing, this helps in tracking the policy, whether the government is still pushing for increase in testing or they want to just limit the testing.

**LockDownUnlockPhases**:- This tracks the current government policy, we have a scale from 7 to 1. Where 7 denotes the full relaxation from all the curbs and 1 denotes the government strictest policy against infection, it include total lockdown, as we move up the scale relaxation increases.

**PreviousPolicyBufferTime(7days)** :- This attribute is added due to a general perspective that any policy change does not impact the upcoming cases suddenly. And also since this infection has some incubation period thus a previous policy may have the impact on daily rise in cases for some days. A period of 7 days is taken as buffer time for such impact of policy.

The last column in the dataset is what this project will predict I.e. **CasesAfter10days** . The project though is able to predict cases after a month or 15 days but keeping in mind the amount of data that this project has, such prediction may not show accurate result. In such a scenario this may not help anyone.

  
Fig : 1.1 ( a snippet of the dataset)

* 1. **Methodology**

In this project the proposed system deals with prediction of a number so the proposed system or project uses the Regression to predict the results. The purpose of this project is to get that model that will give the best result with minimum error. Since by only observing the data, the model cannot be selected and also keeping in mind that the real world data may change its type anytime based on the provided environment thus a system is required that can change dynamically according to the change in behavior of data provided and whatever the model suits at that point of time that can be applied to get the best result. So the proposed system deals with these type of problems and successfully can shift from one model to another based on the error rate.

Regression analysis have a set of ML methods that helps us to predict an outcome Y based on the values of a single or multiple predictor variables (X).

Generally regression is of two types :-

1. Linear Regression
2. Non Linear Regression

In **Linear regression** it can be Simple Linear regression of type Y= aX+c , where the Output Y only depends on one variable X, such a simple linear regression may not work in real life where there are countless factors that determine the outcome, therefore there is Multiple Linear Regression and it is of type Y = aX1+bX2+dX3+…. so on. Here the Outcome Y depends on more than one variable or factors. Here it is worth noting that the coefficients attached to independent variable X are the weights of that variable to which they are attached (Here the term independent variable means that the value of this variable does not depend on others. Since Y’s value is dependent on the values of X therefore it is not independent and is rather dependent) .Weights are assigned on the bases of training of the model though initially they can take arbitrary values. The weights are increased or decreased by the model by moving in that direction which decreases the error. Multiple linear regression will always work better than Simple linear regression because the model has various options to adjust the weights and not just simply always changing only one factor.

**Non Linear Regression** works for those set of data which does not fit into a linear relationship say Y=aX2 +c ; now this is a quadratic relationship between Y and X and this type of relationship cannot be represented correctly in Linear regression thus a non linear regression is used for such type of data. Non Linear Regression are majorly of 4 types in Machine Learning.

* Polynomial Regression : “A model is defined regression in which for a variable y (dependent) there is a variable x (independent) which has a relationship of the form y = x^n (note that n is the degree of polynomial) “.



This equation is of the polynomial form and it can be represented in matrix form also

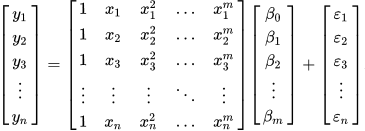


Fig : 1.2 (matrix representation of above equation, source : wikipedia.com )

The polynomial regression is also considered as a special case of multiple linear regression. An example to show the non-linear relationship between X and Y.

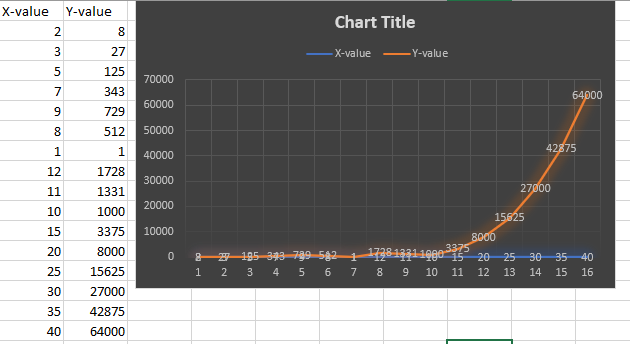


Fig : 1.3 showing the non-linear relationship

* **Support Vector Regression**: - Support Vector Regression uses the concept of SVM i.e. used in classification. In SVM we have decision boundary and a hyperplane. The SVM tries to find a hyperplane in the given plane or in the higher dimensions such that it separates the given points into classes. Decision boundary is that boundary on one side of which lie positive examples and on the other side lie the negative examples. In SVR we use hyperplane to predict the continuous output. And the decision boundary is used to limit the range of points to be considered. Points inside these boundaries are only considered while implementing the SVR. In SVR we have a hyperparameter called Kernel; the main purpose of the kernel is to find the hyperplane without increasing the computation cost as if we move to higher dimensions the cost increases. There are different kernel like ‘rbf’, ‘linear’, ‘gaussian’ etc. in this project we will be using ‘rbf’ and ‘linear’ only.
* **Decision Tree Regression:**- This type of regression builds a model in the form of a tree structure. What is does is that it breaks the given dataset information into smaller and smaller subsets. At every node we have some conditions that if full filled then prediction analysis moves in that side of the node. The algorithm for building the decision tree is called ID3 by J.R. Quinlan. This uses a top-down approach and greedy search through the branches of tree without backtracking. While implementation we have few hyperparameters to consider. The first one is Mean Square Error or similar metrics which is used to determine the splits. MSE measure the quality of an estimator. It is always non-negative and values which are closer to zero are considered better.
* **Random Forest Regression**:-The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one “forest.” The aim is not to heavily rely on a single learning model, but depend on group of decision models to improve the accuracy. Here also “mse”, “mae” are used to measure the quality of split. In the project we have used these hyperparameters like n\_estimator (this is number of trees in the forest) ,max\_features (number of features to consider when looking for the best split).
* **Architecture**

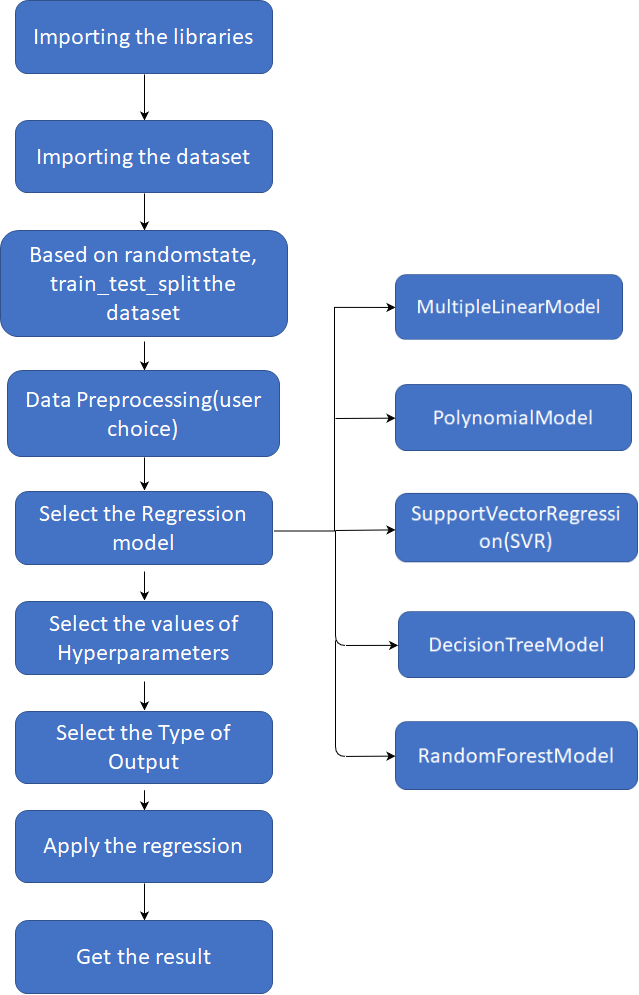


Fig : 1.4 ( Architecture of the Proposed system)

* 1. **Data Preprocessing**

Data preprocessing in Machine Learning is a crucial step that helps us to enhance the quality of data to promote the extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

When it comes to creating a Machine Learning model, data preprocessing is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data preprocessing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models.

There are several significant steps in data preprocessing in Machine Learning as shown below:

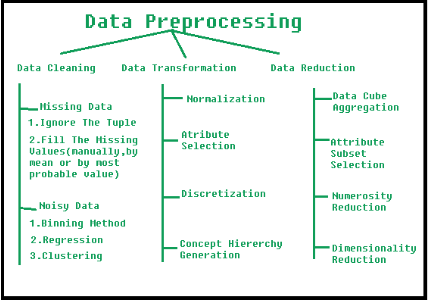


Fig-1.5 Steps in Data Preprocessing

1. **Data Cleaning:**In this step, data is cleansed. That means, it undergoes through processes such as filling in missing values, smoothing the noisy data, and even resolving the inconsistencies in the data.
2. Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

* + 1. Ignore the tuples:

We use this method only when our dataset is quite large and we have multiple values are missing within a tuple.

* + 1. Fill the Missing values:

There are various ways to do this task. The missing values can also be filled manually, by attribute mean or the most probable value.

* 1. Noisy Data:

Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

* + 1. Binning Method:

This method works on sorted data in order to smooth it. We divide the whole data into various segments of equal size and then apply some methods to complete the task. Each segmented is handled separately. Either all data in a segment can be replaced by its mean or one can only use boundary values to complete the task.

* + 1. Regression:  
       One method is to fit the data into a regression function , in order to make itsmooth.The regression used can be of two types:-linear regression (having one independent variable) or multiple regression (having multiple independent variables).
    2. Clustering:  
       This technique is used for grouping the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

NOTE: We are using various Regression methods in the proposed model.

1. **Data Integration:** Data with different representations are put together and conflicts within the data are resolved. This involves following ways:
   1. Normalization:  
      It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
   2. Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

* 1. Concept Hierarchy Generation:

Here attributes are converted from level to higher level in hierarchy.

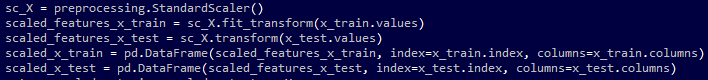
1. **Data Transformation:** Data is normalized, aggregated and generalized.
   1. Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

* 1. Scaling

We can have random values of every feature in a data point. So, it becomes important to scale them such that they may matches specified rules.

NOTE: We perform data Transformation by performing scaling on it.



Output: All the values will have been scaled between the given range.

* 1. Normalization

In normalization,we adjust the values in the feature vector in order to measure them on a common scale. Here, the values of a feature vector are adjusted so that they sum up to 1.Normalization is used to ensure that data points do not get boosted due to the nature of their features.

1. **Data Reduction:**Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. Thus we perform this step, so thatwe can present a reduced representation of the data in a data warehouse.
   1. Attribute Subset Selection:

Only the highly important and relevant attributes are to be used, all the other attributes can be discarded. Thus, to perform attribute selection, either level of significance or p- value of the attribute can be used. Usually, the attribute whose p-value is greater than significance level is considered to bebe discarded.

* 1. Numerosity Reduction:

This enables to store the model of data instead of whole data, for example: Regression Models.

* 1. Dimensionality Reduction:

This reduces the size of data by encoding mechanisms.It can be lossy or lossless. If we can retrieve the original data after reconstruction from compressed data, then such reductions are called lossless reduction, otherwise they are known as lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

1. **Data Discretization:** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.
   1. **Dynamic Model Selector:** As shown in the architecture the dynamic model selector is on of the module of the application. In this module we have the option of selecting what kind of model we want to use for this data prediction. The system supports 5 different types of models that are already described in the previous section , these models are "MultipleLinearModel","Polynomial","SupportVectorRegression(SVR)","DecisionTree","RandomForest". This feature allow this system to adjust dynamically according to the user’s choice. If the user finds that polynomial is working nicely then he has the freedom to choose another model without rewriting the whole code again for a different model.

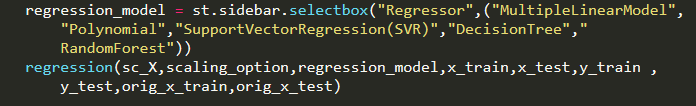


Fig : 1.5 (Code snippet for selecting the type of model)

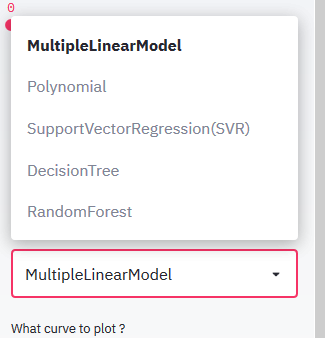
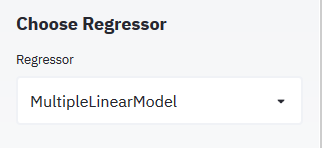


Fig 1.6 (The output of the snippet)

* 1. **Setting up the Hyper Parameters**

We use the term hyperparameter, in machine learning, so as  to distinguish between the predefined parameters and the standard model parameters. Basically, a machine learning model is the one which is build of a mathematical formula with a number of parameters that need to be learned from the data.

However, we come across another kind of parameters which cannot be directly learned from the regular training process. These parameters are used to express the “higher-level” properties of the model like its complexity or how fast it should learn etc. They are called **hyperparameters**. Hyperparameters are those parameters which are usually fixed before the actual training process begins.

Some examples of hyperparameters used in this proposed model are as follows:

* 1. **Degree in Polynomial Regressor**: - Polynomial regression is a form of regression analysis in which the relationship between the independent variable *x* and the dependent variable *y* is modeled as an (*n*th) degree polynomial in *x.* Thus, degree needs to be specified before training process begins.
  2. **Maximum No**.**Of Iterations in SVR : -**  In this , we need to specify the no. of iterations for the Support Vector Regression before proceeding to implement the training process.
  3. **Regularization Parameter in SVR:** - This parameter is required to maintain the trade-off between the test error and training error. Thus, varying the regularization parameter in an SVM can change the decision boundary for a non-separable dataset.
  4. **Criteria in Decision Tree and Random forest Regressor:** - It defines the quality of splitting up of nodes. It generally has 3 values – mse ,friedman\_mse, mae.
  5. **Max\_features in Decision Tree and Random forest Regressor: -** It specifies the no. of features which would be considered while looking for the best split. It can take 3 values – sqrt, auto and log 2, where auto is set as default value.
  6. **Splitter in Decision Tree: -** It explains that how the decision tree searches the given features for a split. The default value of splitter is set to “best”. But it can take two values – best and random.
  7. **Random\_state in multiple Linear model**: We set this value in order to get the same output and split corresponding to the same random number, if this parameter is not set, it might be possible that we het different data split each time we train the model.
  8. **N\_estimators in Random forest Regressor:** It represents the no. of trees that should be made in the forest. Thus, it is also set before executing the training process.
  9. **Training and Testing the Model**

For Training and testing the data in machine learning, we usually split our dataset into two sections – one is train data and other is test data.

* **Training Data**

The observations and the data provided in the training set form the experience that our model algorithm uses to learn. This forms the basis of supervised learning problems, in which each observation consists of an observed output variable and one or more observed input variables.

* **Test Data**

The test set is a set of observations or input data which is used to evaluate the performance of our model, using some performance metric, which it has learnt from train set. It is important to note that; here no observations from the training set are included in the test set. This is important because if the test set does contain examples from the training set, it will be very difficult to for us to know whether the algorithm has learned to generalize from the training set or if it has simply memorized it.

In our proposed model, we will see an app build using streamlit library. Here we will have numerous option available like

* User wants to scale data or not?
* Which regressor does the user selects to implement ?
* Which graph does the user wish to obtain as output?
* If the user wants to see the raw data or not?

Once the user selects all the required options for his/her regressor model, the menu will appear for selecting the hyperparameters as per the user’s choice. As soon as all selections are made, the user would select the regression button in order to run the implementation. Thus, the results would be displayed on the right panel of the screen as per the choices made by the user. The results have been described in the next section.

* 1. **Results :**
* **Using Multiple Linear Model**

Multiple Linear Model using the randomstate = 0

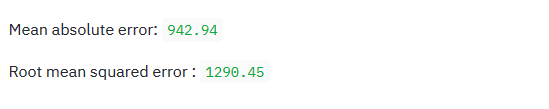


Fig 1.7 (showing the Error in prediction)

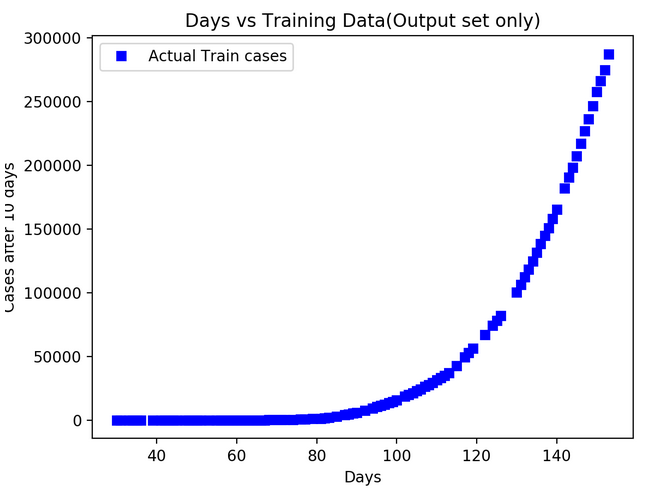
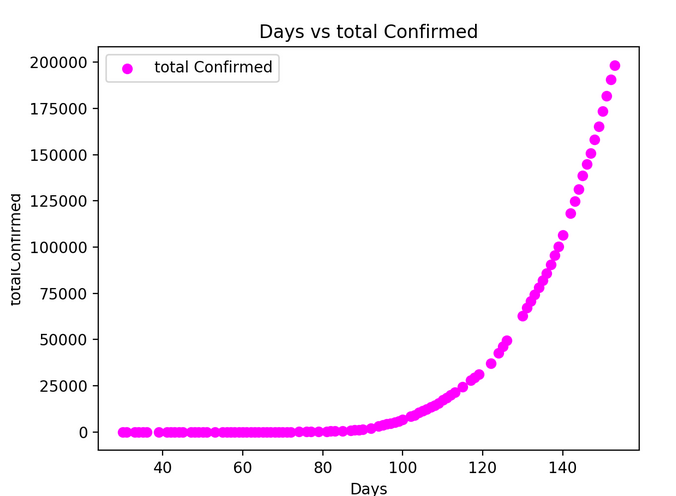


Fig 1.8 Days v/s TotalConfirmed Fig 1.9 Days vs Training data

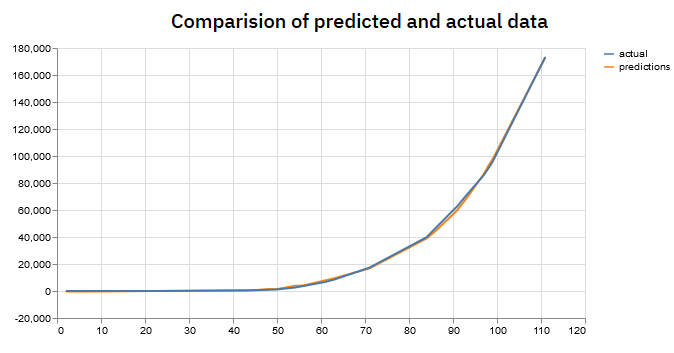


Fig 1.10 Comparsion of predicted and actual data

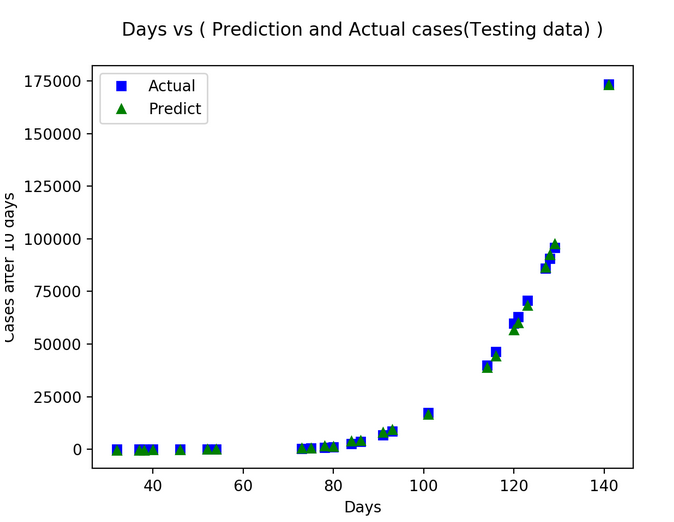


Fig 1.11 Days vs (Prediction and Actual cases(testing data)

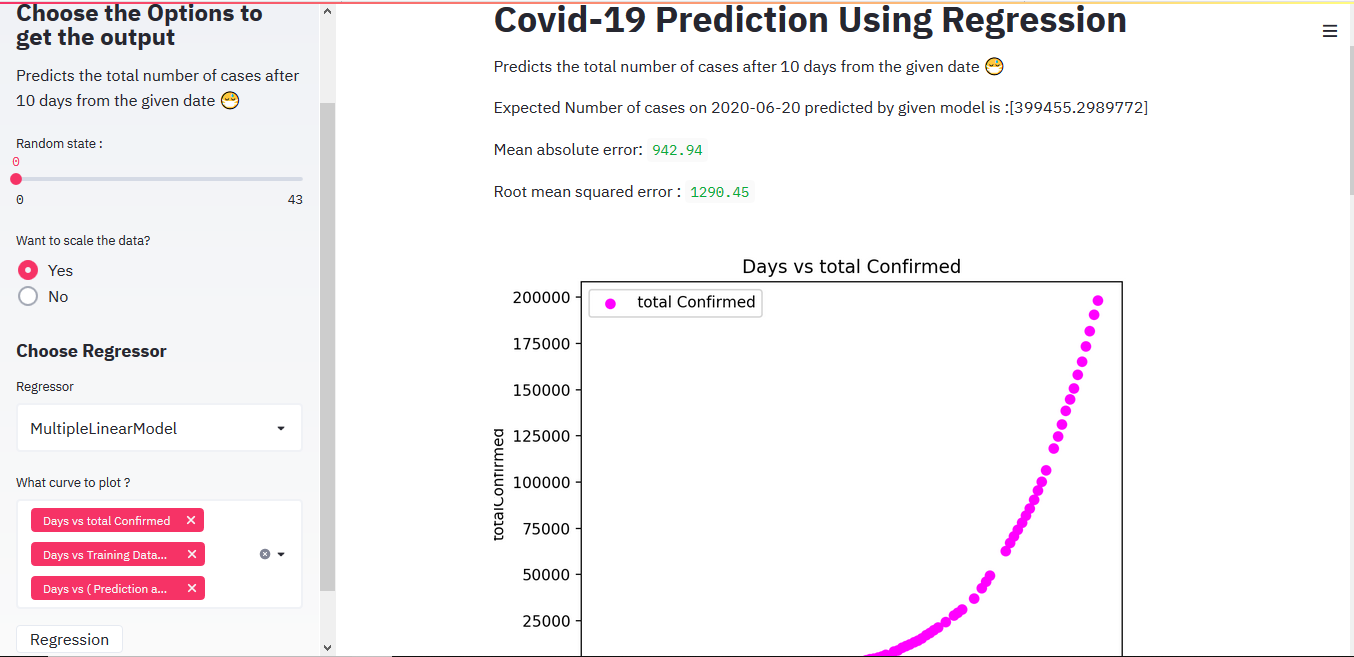


Fig 1.12 web view of application

Here it is observed that the mean absolute error and root mean squared error are around 1k cases. We can say that this is fairly a very good prediction model since model is able to forecast the cases which are in lakhs with an error of +-1k cases. We also observe that if random\_state is changed to 6 then the model predicts even better.



Fig 1.13 error with random\_state =6

* Using Polynomial with degree=2

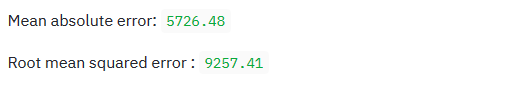


Fig 1.14 error with degree =2 and random\_state = 0

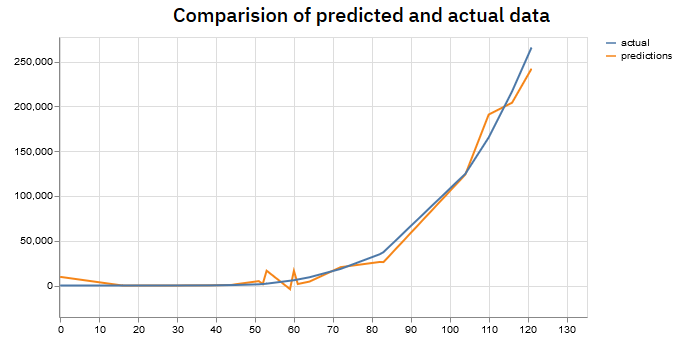


Fig 1.15 Comparison of predicted and actual data

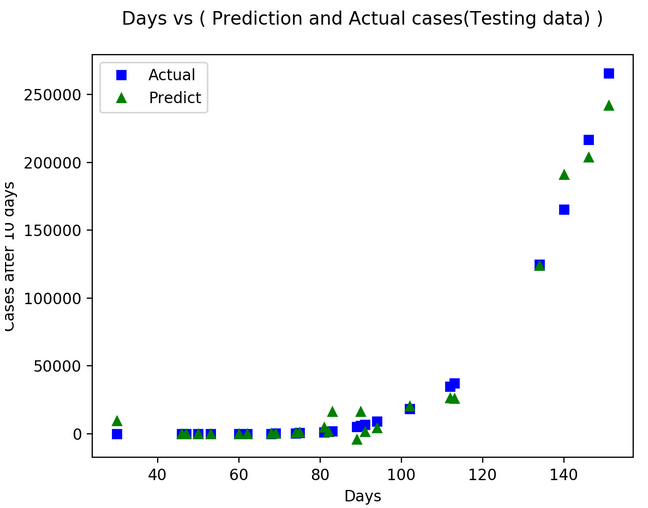


Fig 1.16 Days vs (Prediction and Actual cases(testing data)

We observe that as the degree increases the error also increases and also when degree is made 4 and higher the computation cost increases exponentially resulting in failure of the output. So in the system we will recommend it to keep it upto 3.

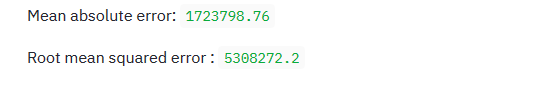


Fig 1.17 Error when degree = 3

* Using Support Vector Machine :

By keeping the following hyperparameters

* + 1. Maximum number of iteration : 300
    2. C (regularization parameter ) : 0.01

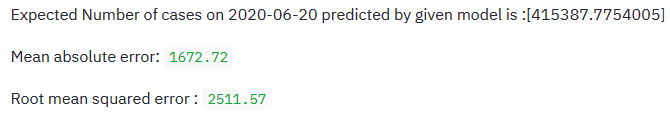


Fig 1.18 Expected Cases and MAE and R2error.

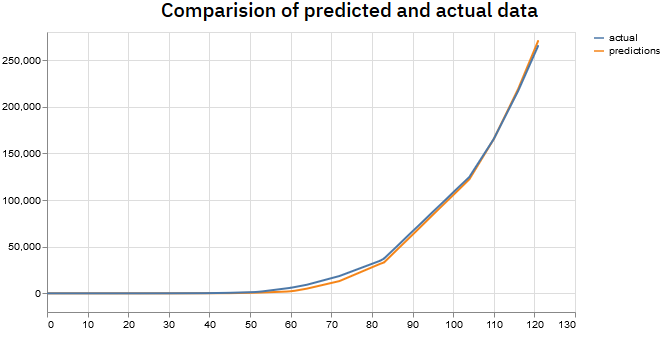


Fig 1.19 Comparison of predicted and actual data

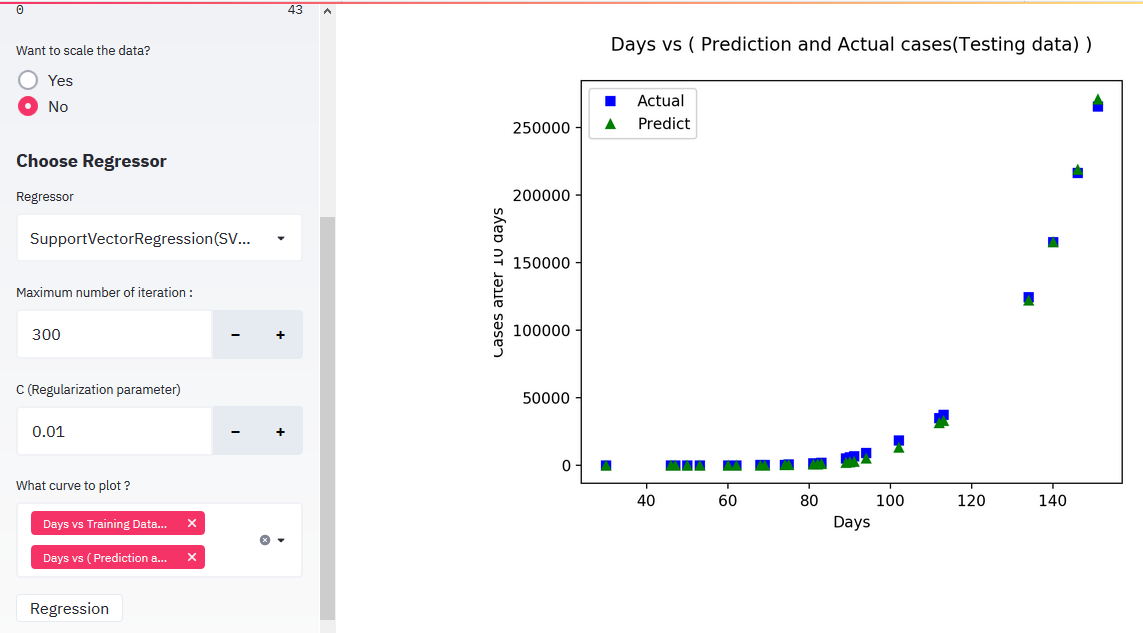


Fig 1.20 Days vs (prediction and Actual Cases )

We observe that even if we increase the Number of iteration the error does not decrease though by changing Regularization parameter accuracy can be increased.

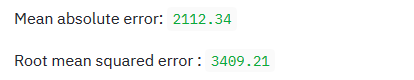


Fig1.21 Error when C=0.01 and iteration = 550

* Using Decision Tree :- We keep the following hyperparameters to minimize the error

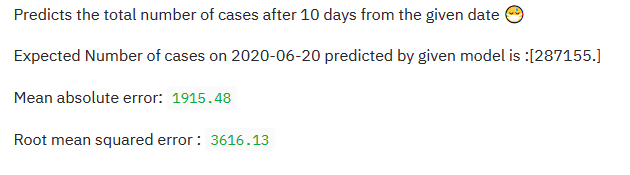
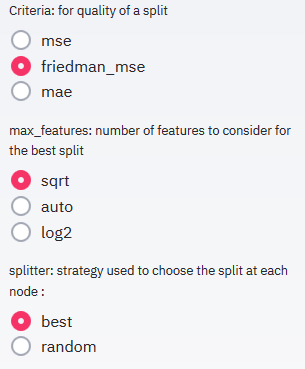


Fig 1.22 Hyperparameters Fig 1.23 Error MAE and RMSE

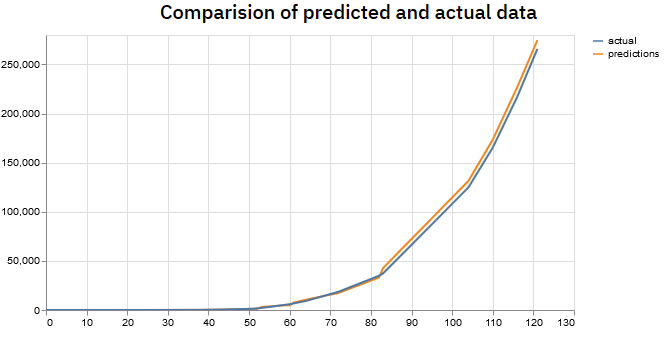


Fig 1.24 Comparison of predicted and actual data

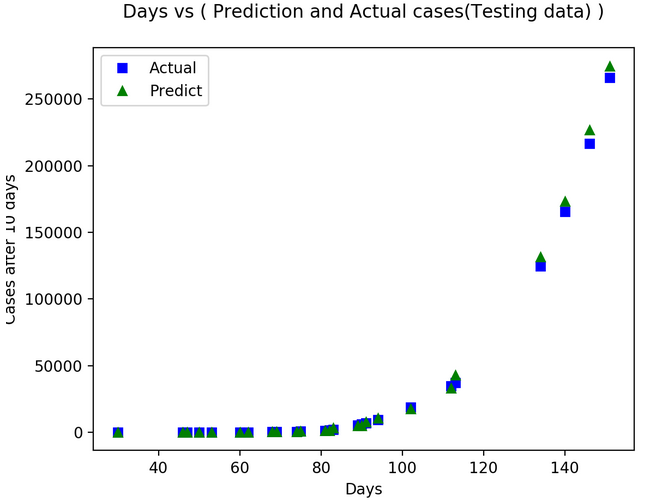


Fig 1.25 Days vs Prediction and Actual cases (Testing data))

We observe that though the error is less then 2k but the model is underfitting. This is due to the reason that SVR works better when the dataset is quite large i.e. about 10k samples. And also the predicted number of cases is less than the total number of cases till that day(10 Jun,2020). So it is obvious that cases on 20 Jun,2020 must be greater than 10 Jun,2020. So the model is not at all good.

* Using Random Forest : Following are the hyperparameters setting.

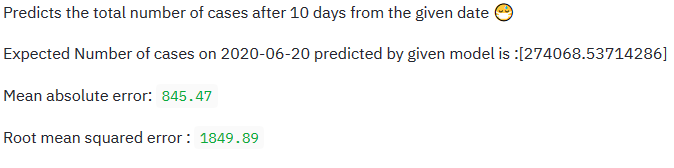
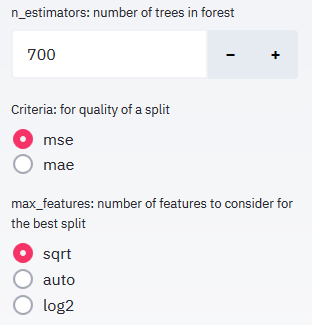


Fig : 1.26 Hyperparameters Fig: 1.27 Error and prediction

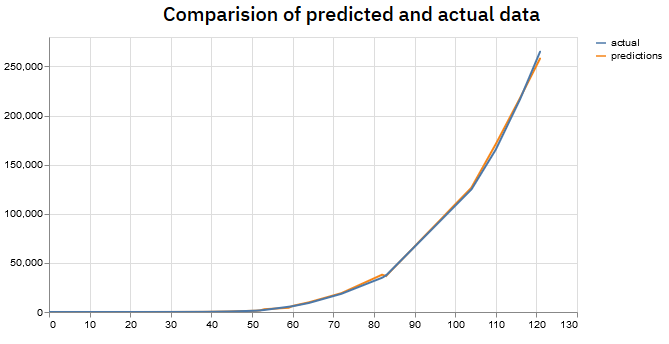


Fig 1.28: Comparison of predicted and actual data

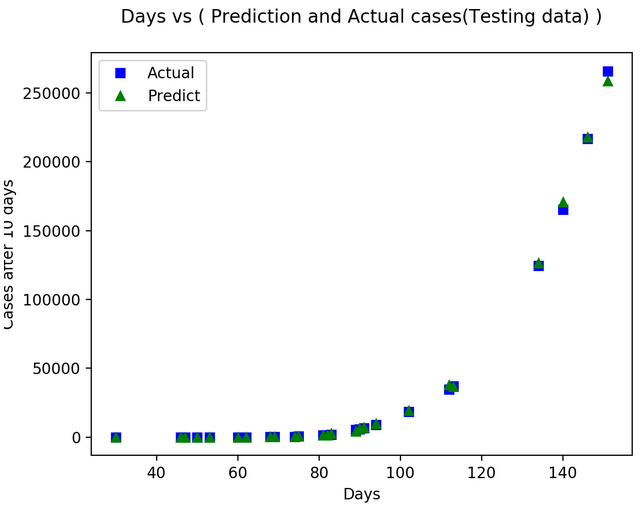


Fig 1.29 Days vs (Prediction and Actual cases)

We observe that here also the data is underfitting and therefore the predicted number of cases is less than total number of cases till that date (Jun 10,2020). This may happened because of less amount of data.

1. **CONCLUSION**

There is no doubt that COVID-19 has bring us those challenges which cannot be matched with any other challenge in the last century. This pandemic is faced by all the nations in all the continents of the world. Just within 6 months, there are more than 400k deaths around the world. To deal with such unprecedent challenge , it is very much necessary that we use all the available technologies and knowledge. Thus to help the government, a system or application is developed that can be used to forecast the number of COVID-19 cases in India. The system uses various regression models to check which model currently works the best. According to the results Multiple Linear regression and SVR works better than any other available model on the current dataset. Multiple Linear Regression has less error as compared to SVR. But the SVR predicted more accurate result as compared to Multiple Linear Regression. The rest of the models failed to predict any satisfying values as they underfit the data. Though as more data will come higher accuracy can be expected from them. This type of system has helped in getting and setting that environment which can predict the cases more correctly as it is very much flexible. The choice of setting the Hyperparameters dynamically will definitely decrease the amount of efforts and time required to get the results. Thus the system will definitely help the government in their effort of dealing with the virus.

1. **REFRENCES**
   1. <https://www.researchgate.net/publication/260015447_A_Study_on_Multiple_Linear_Regression_Analysis>
   2. <https://www.tandfonline.com/doi/full/10.1080/16549716.2020.1760490>
   3. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230405>
   4. <https://www.sciencedirect.com/science/article/pii/S0048969720323779>
   5. <https://www.researchgate.net/publication/340394777_SEIR_and_Regression_Model_based_COVID-19_outbreak_predictions_in_India>