COLLEGE OF ENGINEERING
AND COMPUTER SCIENCE
Spring 2020

Advanced Database Management Systems

Report for

Analysation and Prediction of the Spread of Corona

Report Prepared By:

CWID	**Name**
887480747	**Venkata Pranathi Immaneni**

TABLE OF CONTENTS

1. Abstract 4

2. Introduction 5

2.2 Problem Statement 5

2.3 Project Goal 5

3. Literature Review 6

3.1 Background of Literature Review: 6

4.Methodology 8

4.1 Approach 8

4.2 Data Collection Source 9

4.3 Data Preprocessing Steps 9

4.4 Machine Learning Algorithm - Suitable for our project: 12

4.4.1 Choosing Appropriate Regression Algorithms: 13

5. Implementation 14

5.1 Creating the Model: 14

6. Results 17

6.1 Predictions for Confirmed Coronavirus Cases 17

6.1.1 Predicting the test set Y values, by passing test set X values: 17

6.1.2 Accuracy Measurement: 18

6.1.3 Accuracy Results: 18

6.1.4 Line Graph Visualization for future prediction: 19

6.1.5 Bar Graph Visualization for future prediction: 20

6.2 Predictions for Death Cases: 21

6.2.1 Predicting the test set Y values, by passing test set X values: 21

6.2.2 Line Graph Visualization for future prediction: 22

6.2.3 Bar Graph Visualization for future prediction: 23

6.3 Predictions for Recovered Cases: 24

6.3.1 Predicting the test set Y values, by passing test set X values: 24

As we can see, the results are almost closer 24

6.3.2 Line Graph Visualization for future prediction: 25

6.4.2 Bar Graph Visualization for future prediction: 26

6.5 Visualisation of Actual Test Data, Predicted Test Data and Future prediction for USA 27

7. Implications 29

6.1 Benefits of the Project: 29

6.2 Lessons Learned: 29

8. Conclusion 30

9. Appendix 30

10. References 31

11. Source Code 31

1. Abstract

Through this project “Analyzation and Prediction of Spread of Corona”, we would like to present our Machine Learning Analysis of COVID 19.

This analysis can be useful for government officials and public health officials, to analyse the current situation and can get to know more insights about the impact of corona in the future. So with this project, we would like to predict the spread of corona for the future 30 days, for the whole world as well as for the United States of America.

Many were being infected by Corona from the past 4-5 months. There were many deaths due to Corona and many of them were fighting against it and getting recovered. The main concern of this Virus, the symptoms are seen only after 14-15 days, after the person is infected with this virus. So many people are falling prey for this virus. So, initially, when we submitted the proposal for this project, the cases were registered only in China and only a few cases in the USA. But, it was a hot topic at that time. And also at that time, we don’t have much data about this Virus, but our curiosity to analyse about this virus, made us land in this project.

So, Regression Machine Learning Algorithm is used for the prediction of the number of corona infected cases, deaths and recoveries for the whole and as the USA is the most affected country, we have chosen USA, to predict only the number of corona cases for the USA for the next 30 days. We are going to analyse the situation for future.

2. Introduction

2.1 Background

Currently, there are many people, who are being affected with CoronaVirus. It started in China and now it is spreading all over the world. Till now, there is no medicine for this virus and its killing millions of millions of people. So, it is a big question among all of us of how many people are going to be affected.

2.2 Problem Statement

Currently there is no application that can predict the spread of CoronaVirus for the future 30 days. So, with this project, we would like to create awareness among the people, by showing them how the corona rises for the future 30 days, so that they can take some preventive measures by staying indoor.

2.3 Project Goal

The main objective of this project is :

- Future prediction of the increase/decrease in the number of active Coronavirus Cases for the next 30 days - for the whole world as well as for the United States of America. We have chosen the USA among all the counties as it is the highly affected country due to corona. .
- Future prediction of the increase/decrease in the number of deaths due to Coronavirus for the next 30 days.
- Future prediction of the increase/decrease in the number of recovered cases due to Coronavirus for the next 30 days.

3. Literature Review

There is an outbreak of Corona in early december. This is caused due to severe acute respiratory syndrome coronavirus 2 , which is basically the family of SARS virus. The World Health Organisation did not announce this to the world till Jan 30 2020. So, on Jan 30 2020, this coronavirus outbreak was declared as Public Health Emergency by the world health organisation. Many governments all over the world are issuing their own preventive measures to control the spread of coronavirus. So, we have conducted the literature review regarding this virus, based on the information that is publicly available.

3.1 Background of Literature Review:

China alerted WHO on 31st december 2019 that many people are reported to be suffering from Pneumonia, in Wuhan City. They reported that it started on Dec 8th 2019, and there were an increasing number of patients who are working or living around the Huanan Seafood Wholesale Market.

When we started working on this project in the start of February, the Corona virus was majorly prevalent in China. Initially at the time of our project proposal, the mortality rate in China among all the confirmed cases is around 1.2% as of February 2020. And the mortality rate in all other countries, other than china was around only 0.2%. Among all the patients, who were admitted to the hospitals, the mortality rate, was around 11%. COVID-19 is increasing with a great speed, and now there is a relatively very high mortality rate.

3.2 A Way to Further Research :

So, we have performed this literature review, to analyse the spread of coronavirus. After analysing how increasingly its spreading all over the world, we thought of performing our own prediction regarding this virus, so as to make people aware of its spread, and with this they can take their own preventive measures, so that they do not fall prey to this dangerous virus.

We had very little amount of data when we started this project. It is a very trending topic all over the world. And millions of millions of people are losing their lives due to this virus. So, we are very curious to analyse this pandemic and so we have taken up this project.

We have found many datasets to collect the data regarding the corona cases. Some of them include Kaggle, John Hoppkins etc. So, we thought of choosing the dataset from John Hoppkins, as it's updating the dataset on a daily basis. So, we have collected the data from 22rd January 2020 - till date and performed our own future predictions.

4.Methodology

4.1 Approach

So, basically, we have followed the below approach to kick start with our project:

1. Firstly, we have started with research on choosing the datasets. On performing research on various datasets, we have finalised with John Hoppkins data set, as it gives us the live data on coronavirus.
2. Secondly, we have collected the data and performed our preprocessing operation, so as to make our data ready for future predictions.
3. Next, coming to choosing the machine learning algorithm. We have chosen appropriate machine learning(we will discuss below regarding this).
4. Finally, we have performed our predictions to analyse the active cases, deaths and recoveries for the next 30 days, based on the data available from the datasets and the chosen machine learning algorithm.

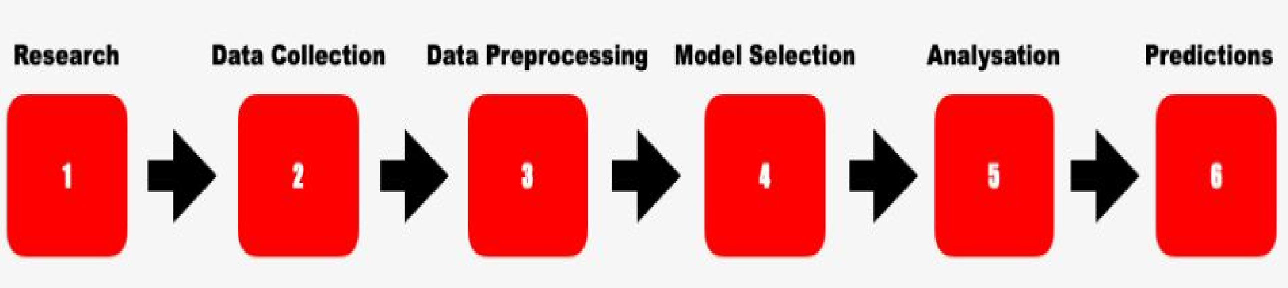


Figure 1: Approach

4.2 Data Collection Source

We have collected the data from the source of John Hopkins Github.

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

4.3 Data Preprocessing Steps

We performed data processing in 5 steps:

Step 1: Importing the data

We have imported the datasets of active cases, recovered cases and death cases from the below data sets:

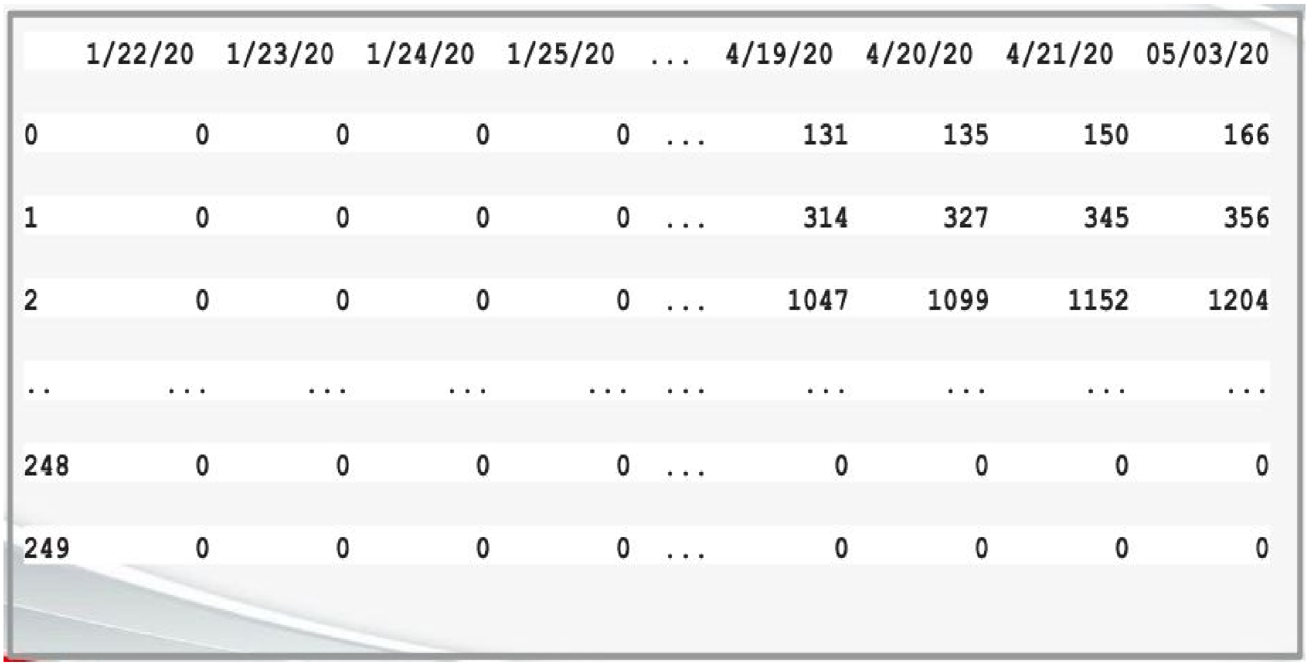
Data Source Links	
Confirmed cases	**https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv**
Deaths cases	**https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv**
Recovered cases	**https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv**

Step 2: Retrieving Required Columns from the datasets:

- As the next step of our data preprocessing, we have eliminated all the unwanted columns by retrieving all the required columns January 22nd 2020 to Till date.
- Below is the snippet of code that we used to retrieve the required columns.

Data Source Links
confirmedCases_ForGivenDates = corona_confirmed_cases.loc[:,Retrive_Required_Dates_Columns[4]:Retrive_Required_Dates_Columns[-1]]
deathsCases_ForGivenDates = corona_deaths_cases.loc[:,Retrive_Required_Dates_Columns[4]:Retrive_Required_Dates_Columns[-1]]
recoveredCases_ForGivenDates = corona_recovered_cases.loc[:,Retrive_Required_Dates_Columns[4]:Retrive_Required_Dates_Columns[-1]]

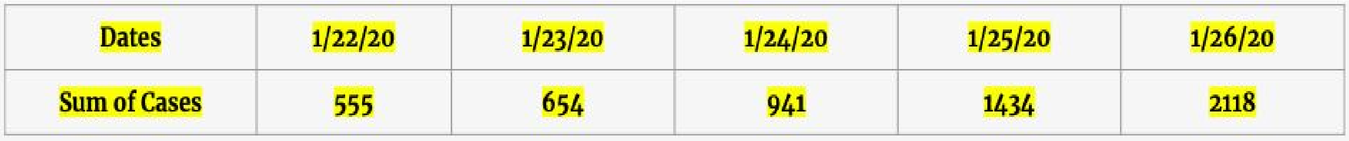
Output: For clear understanding of how the above step works, below is the output:



```
    1/22/20  1/23/20  1/24/20  1/25/20  ...  4/19/20  4/20/20  4/21/20  05/03/20

0         0        0        0        0  ...      131      135      150       166

1         0        0        0        0  ...      314      327      345       356

2         0        0        0        0  ...     1047     1099     1152      1204

..      ...      ...      ...      ...  ...      ...      ...      ...       ...

248       0        0        0        0  ...        0        0        0         0

249       0        0        0        0  ...        0        0        0         0
```

Step 3: Calculating the sum of the cases, for each day:

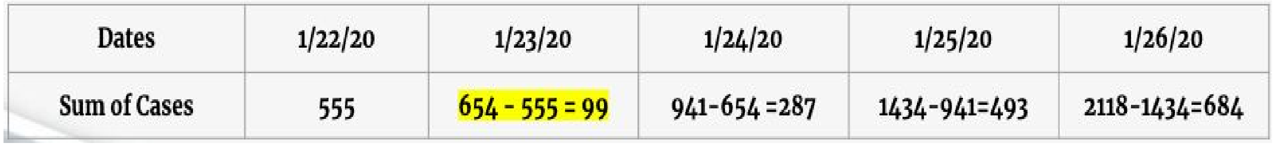
- So, as the next step in our data preprocessing, we are calculating the sum of active cases, death cases and recovered cases for every day.



Dates	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20
Sum of Cases	555	654	941	1434	2118

Step 4: Calculating the daily increase of the cases:

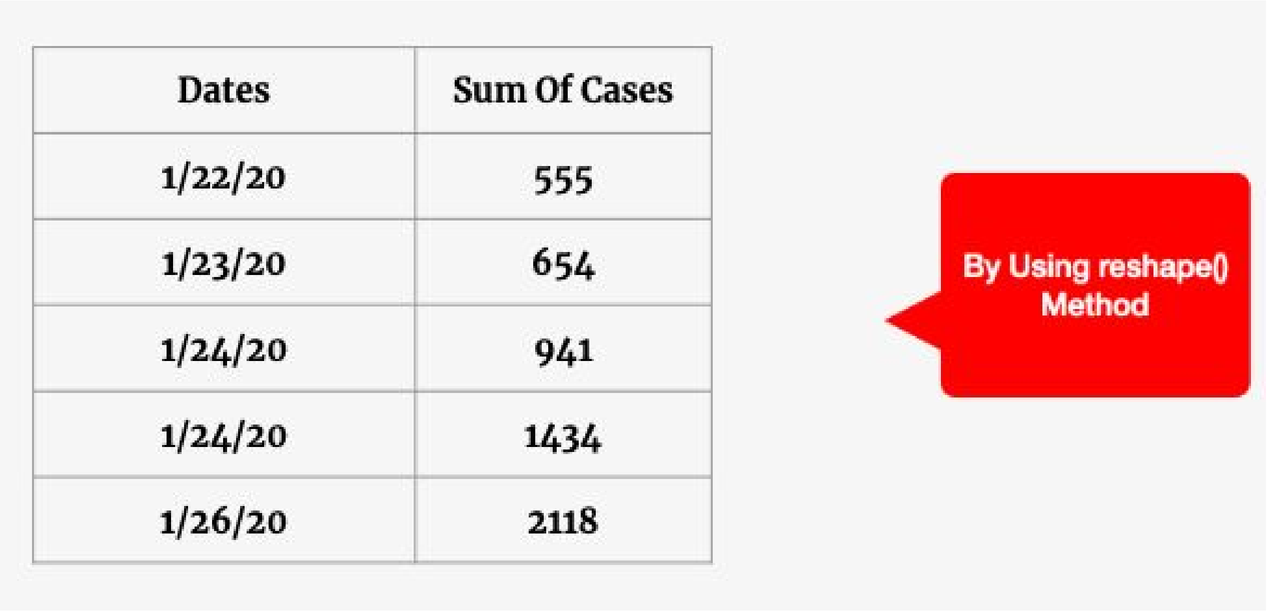
- The next step, in our data preprocessing is to calculate the daily increase of the confirmed, death and recovered cases for the whole world.
- It is basically the difference of the count from the immediate previous day.



Dates	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20
Sum of Cases	555	654 - 555 = 99	941-654 =287	1434-941=493	2118-1434=684

Step 5: Transposing the rows and columns:

- To continue with our prediction, we need to transpose the rows and columns for the data generated in the previous step.
- We have performed the transposing using reshape() method.

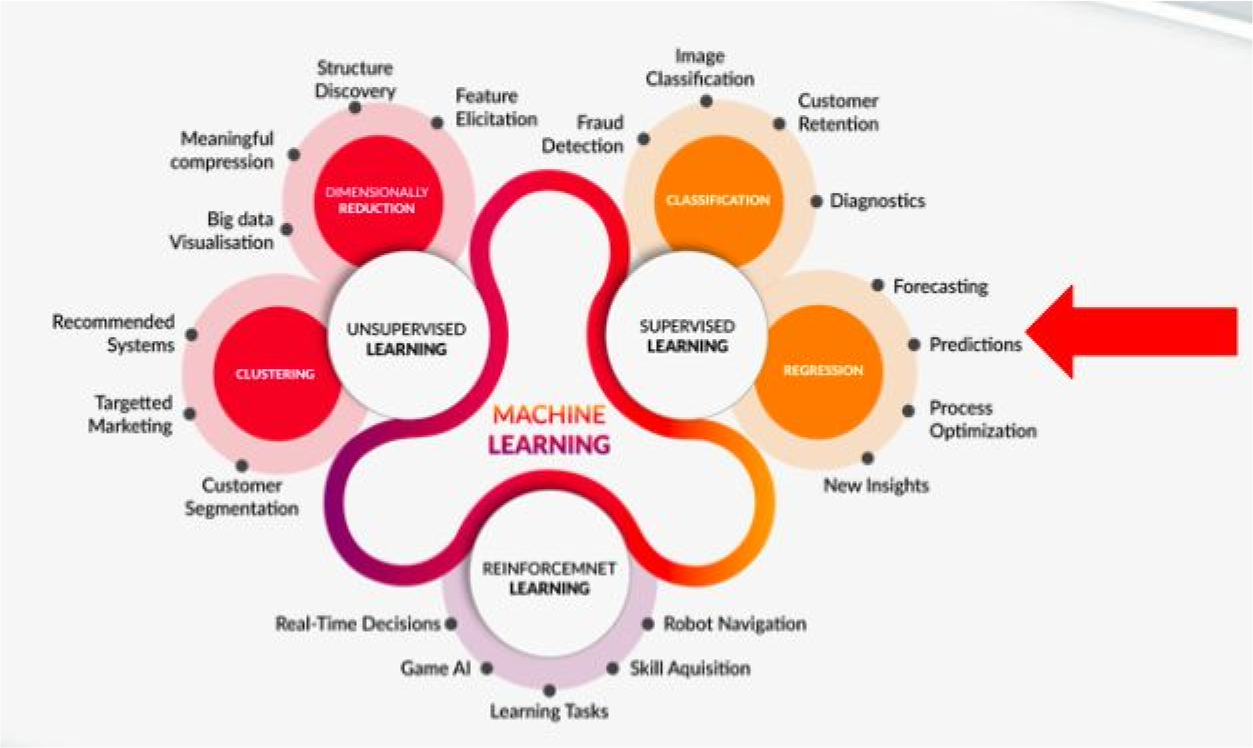


Dates	Sum Of Cases
1/22/20	555
1/23/20	654
1/24/20	941
1/24/20	1434
1/26/20	2118

By Using reshape() Method

4.4 Machine Learning Algorithm - Suitable for our project:

As our project is related to the future forecasting, we have chosen Regression Algorithm for our predictions. Regression algorithm comes under the category of Supervised learning of Machine Learning. With these regression algorithms, we can predict the output values based on the input values, with the help of the data sets that we feed into the algorithm.



4.4.1 Choosing Appropriate Regression Algorithms:

We have chosen two Regression algorithms to predict about the corona cases:

1. Support Vector Regression
2. Polynomial Regression

We have chosen a Support Vector Machine Regression(SVR) Algorithm to forecast the increase in corona cases, death and recovered cases for the World Case Scenario. But for predicting the US Cases, the SVR model seemed to produce inappropriate values, i.e there is a great difference in the test set actual values and the predicted values. So, we have chosen a Polynomial Regression algorithm to predict the number of corona cases in the USA.

5. Implementation

5.1 Creating the Model:

Step 1: Splitting of Dataset

Firstly we splitted the data into test set and Train set. Here, we have chosen 66% of the data to train the model and 34% of the data to test the model.

```
X_Train, X_Test, y_Train, y_Test = train_test_split(RecordedCases_StartDate,
TotalWorldCases_List, test_size=0.34, shuffle=False)
```

As our prediction is daywise, and as we want to predict for the next 30 days, we do not want to shuffle the data. So, we have chosen Shuffle as false. If we shuffle the data, it will produce the inaccurate results.

Step 2: Training the dataset using SVR Model

Next we are building the model using SVR. Below is the code snippet. Let us see the reasoning about why we have those parameters:

```
SVR_Model = SVR(shrinking=True, kernel='poly',gamma=0.01, epsilon=1,degree=4, C=0.1)


SVR_Model.fit(X_Train, y_Train)
```

'shrinking' Param: shrinking parameter is of boolean type. With this parameter, we can speed up the optimisation because it will shrink the dataset, when we set it to true.

‘kernel’ Param: We have different options that can be set to the kernel parameter. They include: linear, poly, rbf, sigmoid, precomputed, or callable. The default parameter for this rbf, if nothing is chosen. As poly suites best to our dataset, we have chosen poly.

‘gamma’ parameter: gamma parameter is of type ‘Float’. The lower the gamma parameter is, the farther points will be taken into the consideration.

‘epsilon’ parameter: It actually determines how much error can be allowed for our training data instance.

- It ranges from 0 to maximum allowable for our data instance.

‘degree parameter’: Here degree denotes the degree of our polynomial. If we have a higher value of this degree, there is a chance of overfitting. Here, we have chosen 4 as it best fits our dataset.

‘c’ parameter: This parameter is the tuning parameter. If we have the lower value of c, then there will be lesser chances to violate the boundaries.

Step 3: Training the dataset using SVR Model

Now, after building up the model, we will predict the test set Y values, by passing our test set X values, to our SVR model.

```
SVR_TestDataPredictionResult = SVR_Model.predict(X_Test)
```

Step 4 : Using the model for future forecasting

Now we predicted the increase in corona cases, deaths or corona recoveries for the future 30 days, by using the below snippet of code.

```
SVR_FutureDataPredictionResult =
SVR_Model.predict(Concat_PastNFutureDays[Current_Day:Next_30Day])
```

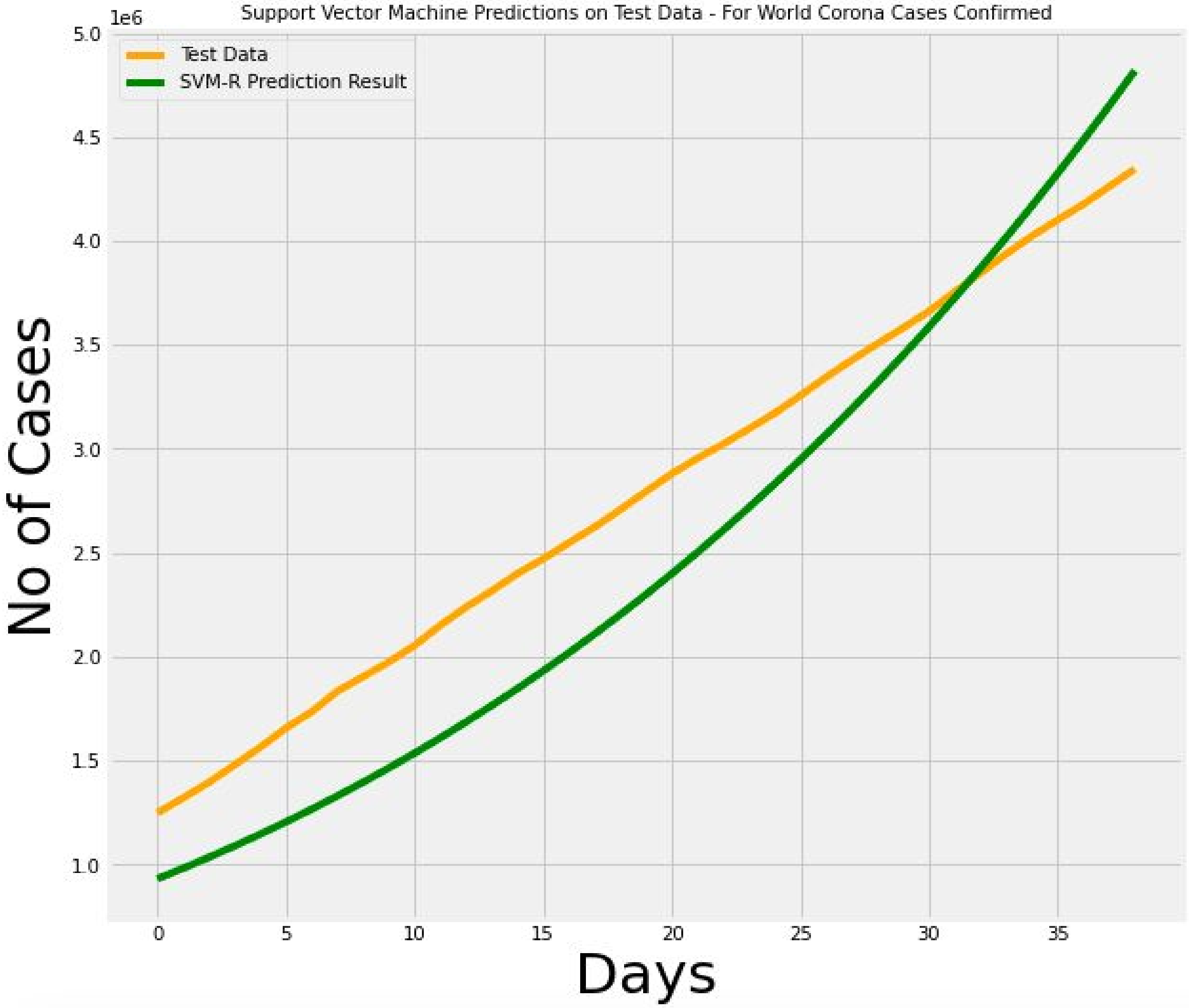
6. Results

6.1 Predictions for Confirmed Coronavirus Cases

6.1.1 Predicting the test set Y values, by passing test set X values:

Here the green line represents our SVR prediction result and Orange Line Represents the Test data result.

As we can see, the results are almost closer.

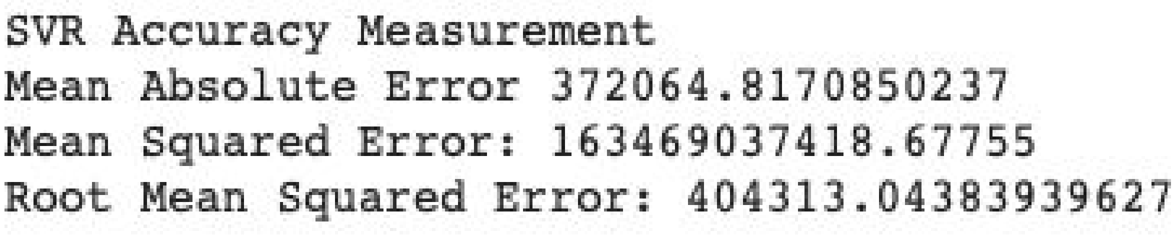


6.1.2 Accuracy Measurement:

For regression models, accuracy is measured by Mean Absolute error, Mean Squared Error and Root mean squared error. The following snippet of code is used to calculate the accuracy of our model:

```
print('SVR Accuracy Measurement')
print('Mean Absolute Error', mean_absolute_error(SVR_TestDataPredictionResult, y_Test))

print('Mean Squared Error:',mean_squared_error(SVR_TestDataPredictionResult, y_Test))

print('Root Mean Squared Error:',
np.sqrt(mean_squared_error(SVR_TestDataPredictionResult, y_Test)))
```

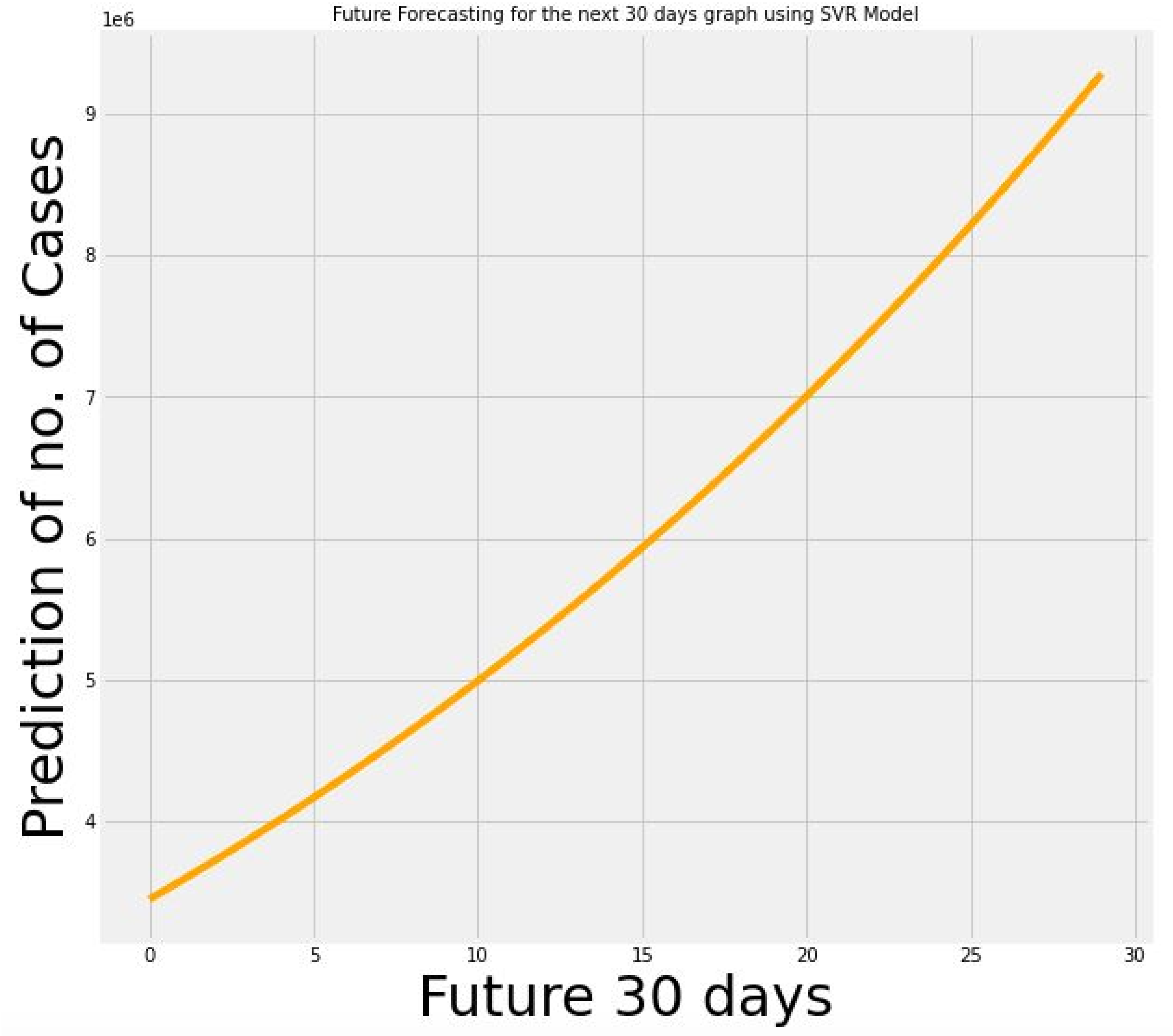
6.1.3 Accuracy Results:



```
SVR Accuracy Measurement
Mean Absolute Error 372064.8170850237
Mean Squared Error: 163469037418.67755
Root Mean Squared Error: 404313.04383939627
```

6.1.4 Line Graph Visualization for future prediction:

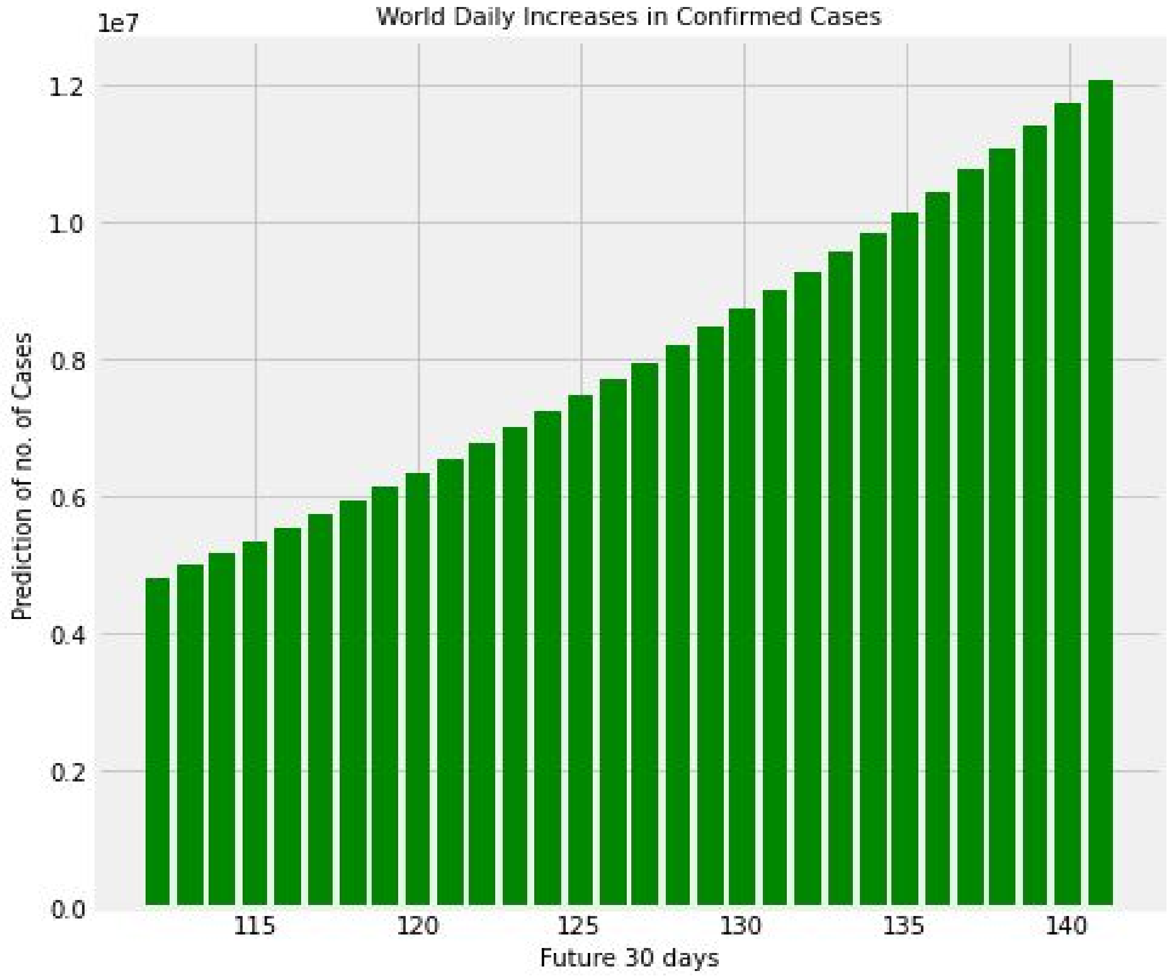
Now, let us visualise the future prediction of an increase in the number of coronavirus cases for the future 30 days.



Based on our prediction, On the 30th day, the number of active corona cases can be around 12078762.

6.1.5 Bar Graph Visualization for future prediction:

Below is the bar graph visualisation of the increase in the number of corona cases for the next 30 days.



Xaxis values starts from 113 - represents the May 13 (No of days from 22nd Jan to May 13th)

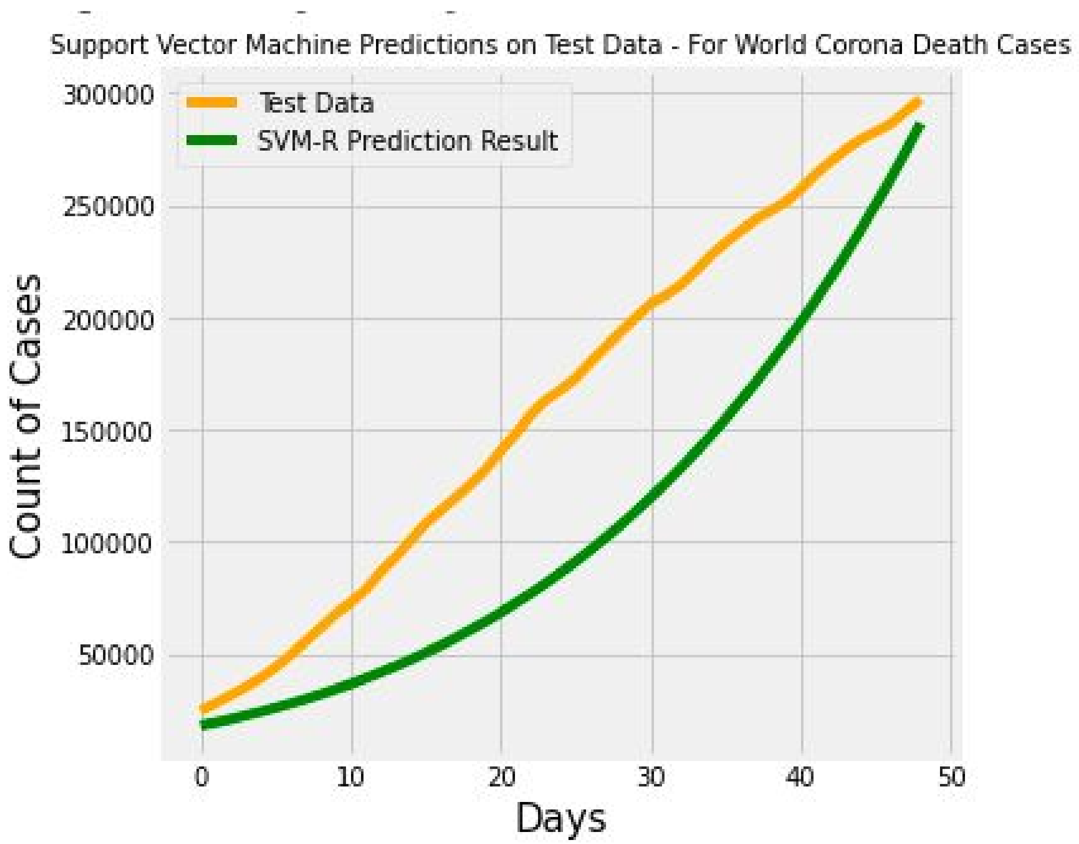
Xaxis values ends with 143 - represents future 30 days(113+30 = 143)

6.2 Predictions for Death Cases:

6.2.1 Predicting the test set Y values, by passing test set X values:

Here the green line represents our SVR prediction result and Orange Line Represents the Test data result.

As we can see, the results are almost closer

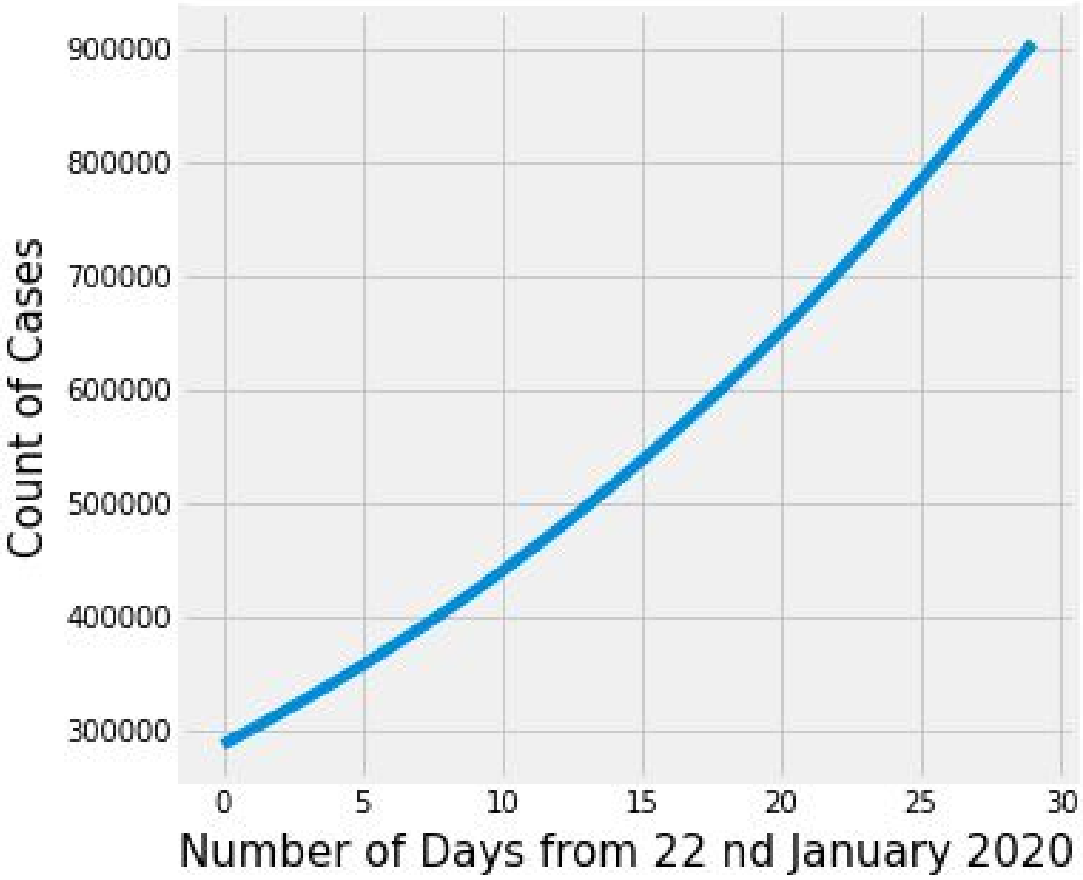


No of deaths (As of 12 th May 2020)

Actual Value: 286517

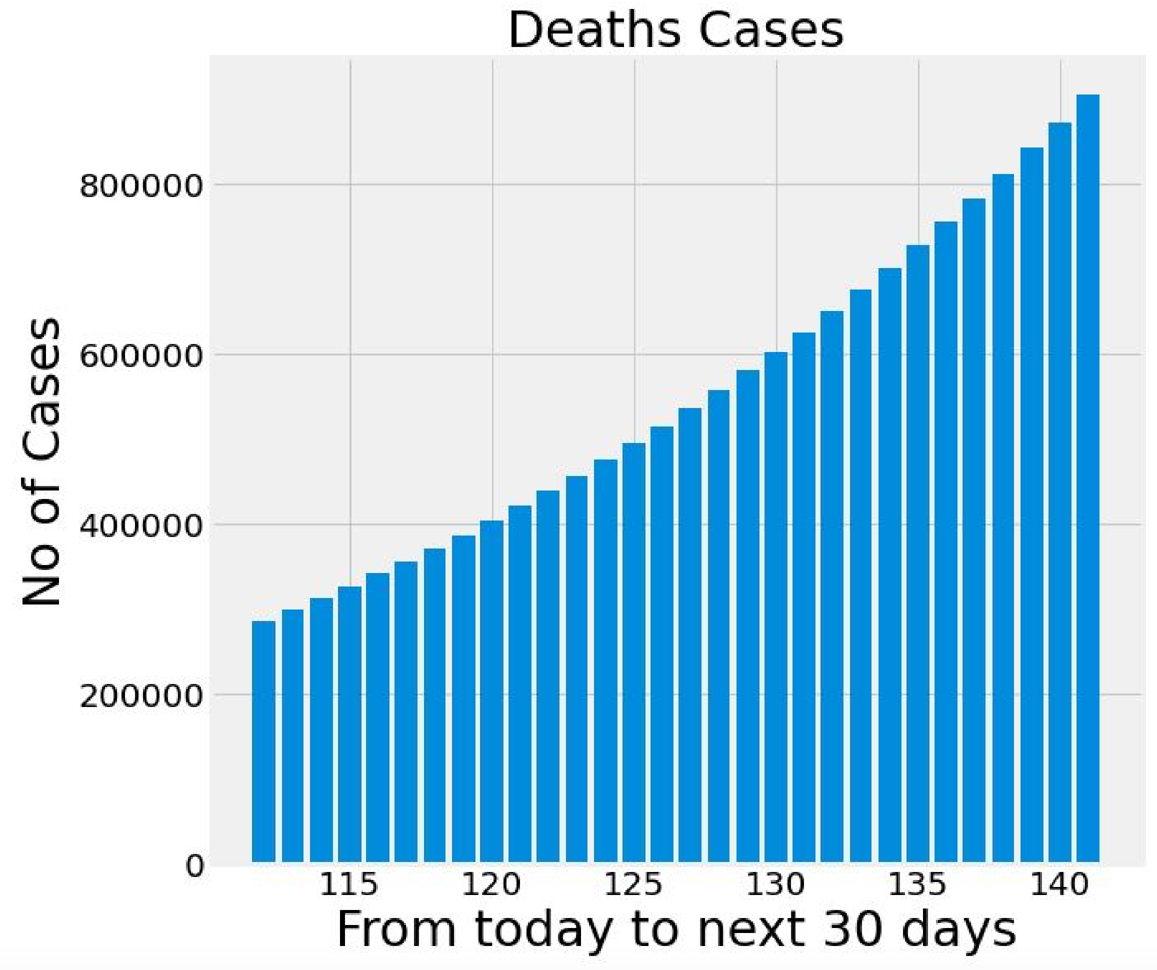
Predicted Value: 297197

6.2.2 Line Graph Visualization for future prediction:



Based on our prediction, On the 30th day, the number of death cases due to corona can be around 904364.

6.2.3 Bar Graph Visualization for future prediction:



Xaxis values starts from 113 - represents the May 13 (No of days from 22nd Jan to May 13th)

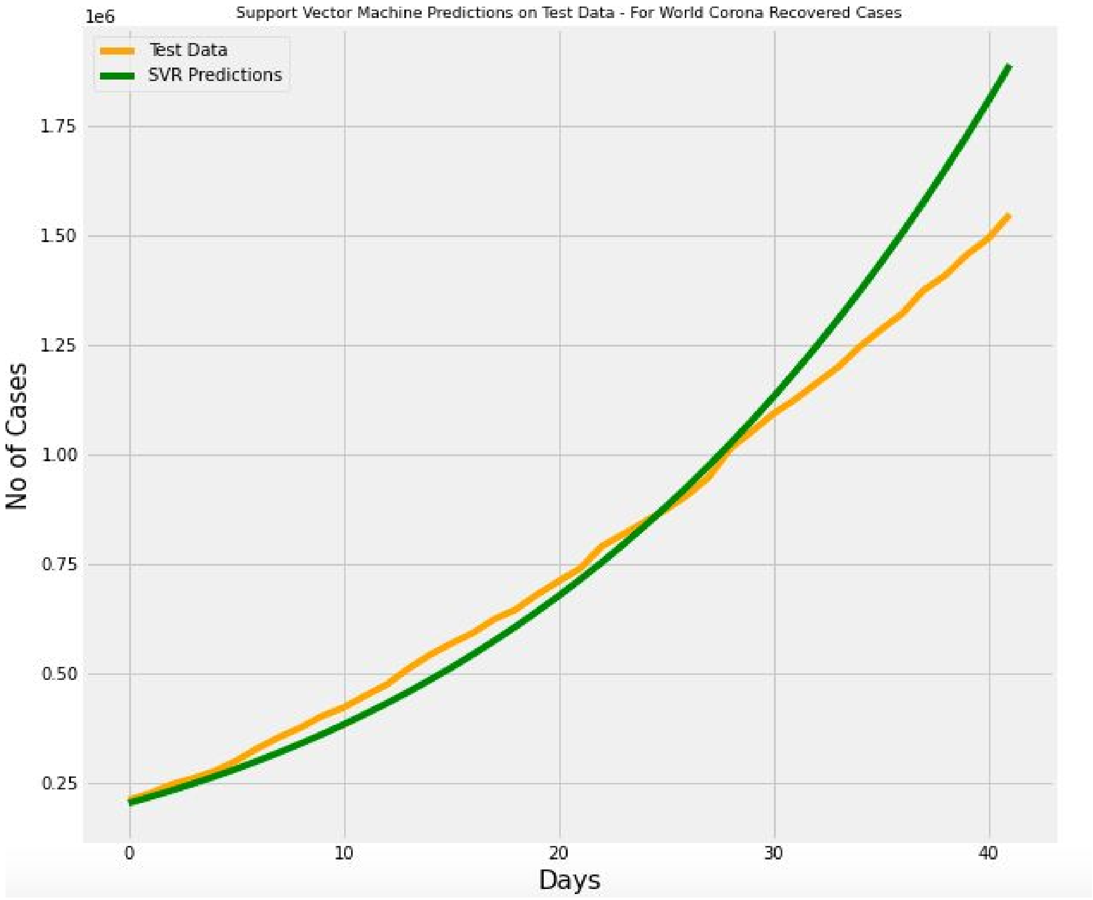
Xaxis values ends with 143 - represents future 30 days(113+30 = 143)

6.3 Predictions for Recovered Cases:

6.3.1 Predicting the test set Y values, by passing test set X values:

Here the green line represents our SVR prediction result and Orange Line Represents the Test data result.

As we can see, the results are almost closer

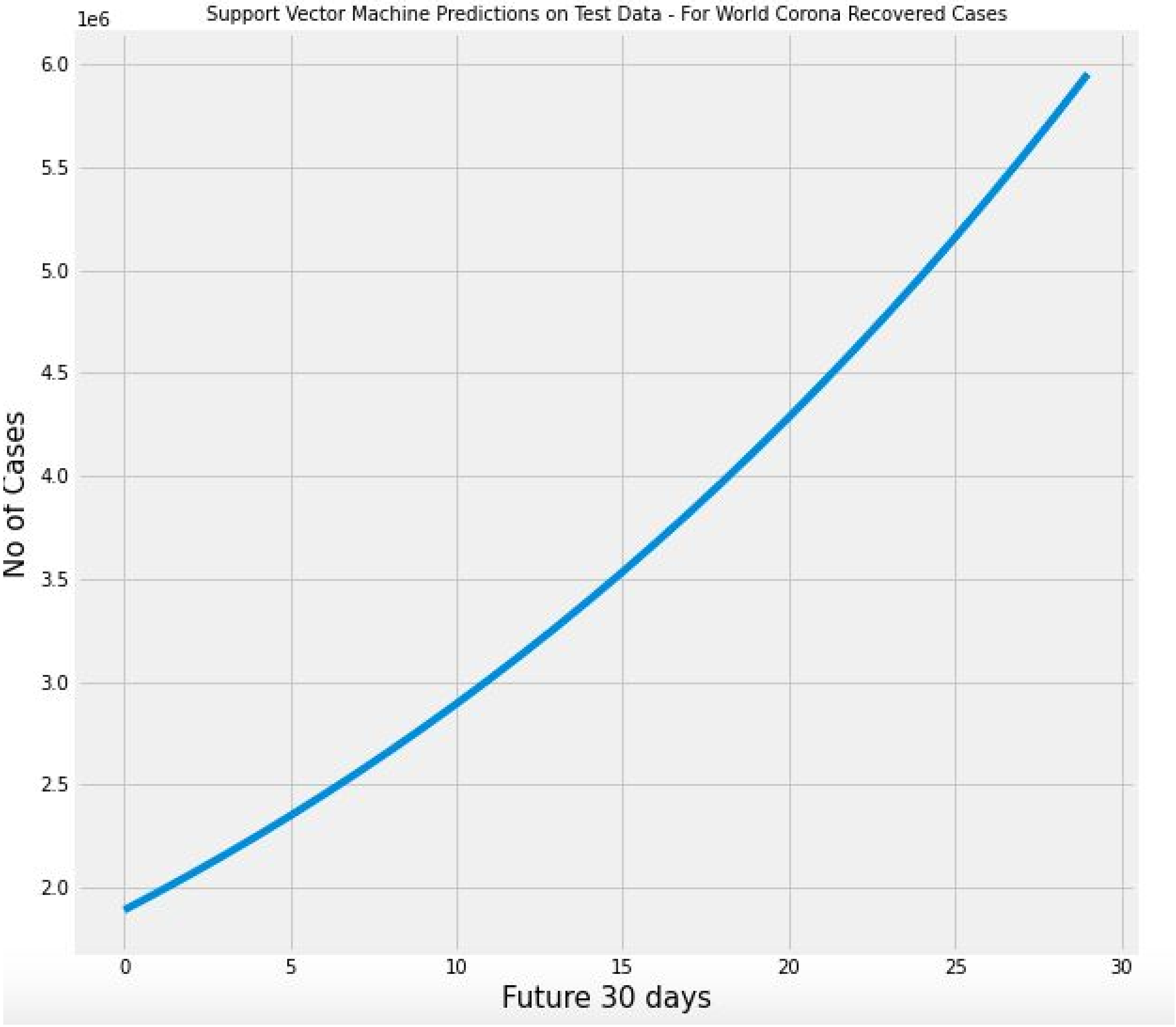


No of recoveries (As of 12 th May 2020)

Actual Value: 1548547

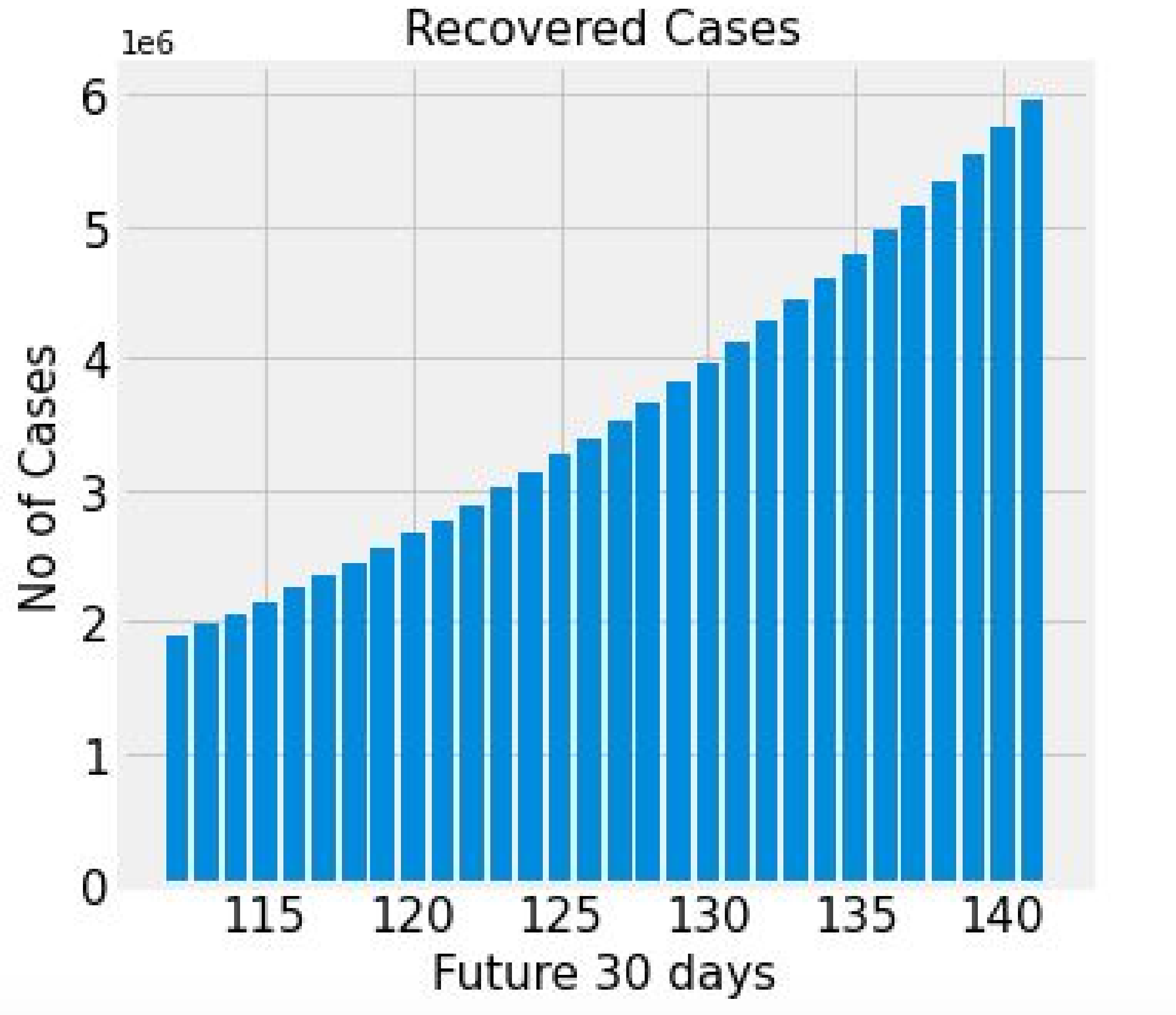
Predicted Value: 1889858

6.3.2 Line Graph Visualization for future prediction:



Based on our prediction, On the 30th day, the number of recovered cases from corona can be around 5951859

6.4.2 Bar Graph Visualization for future prediction:



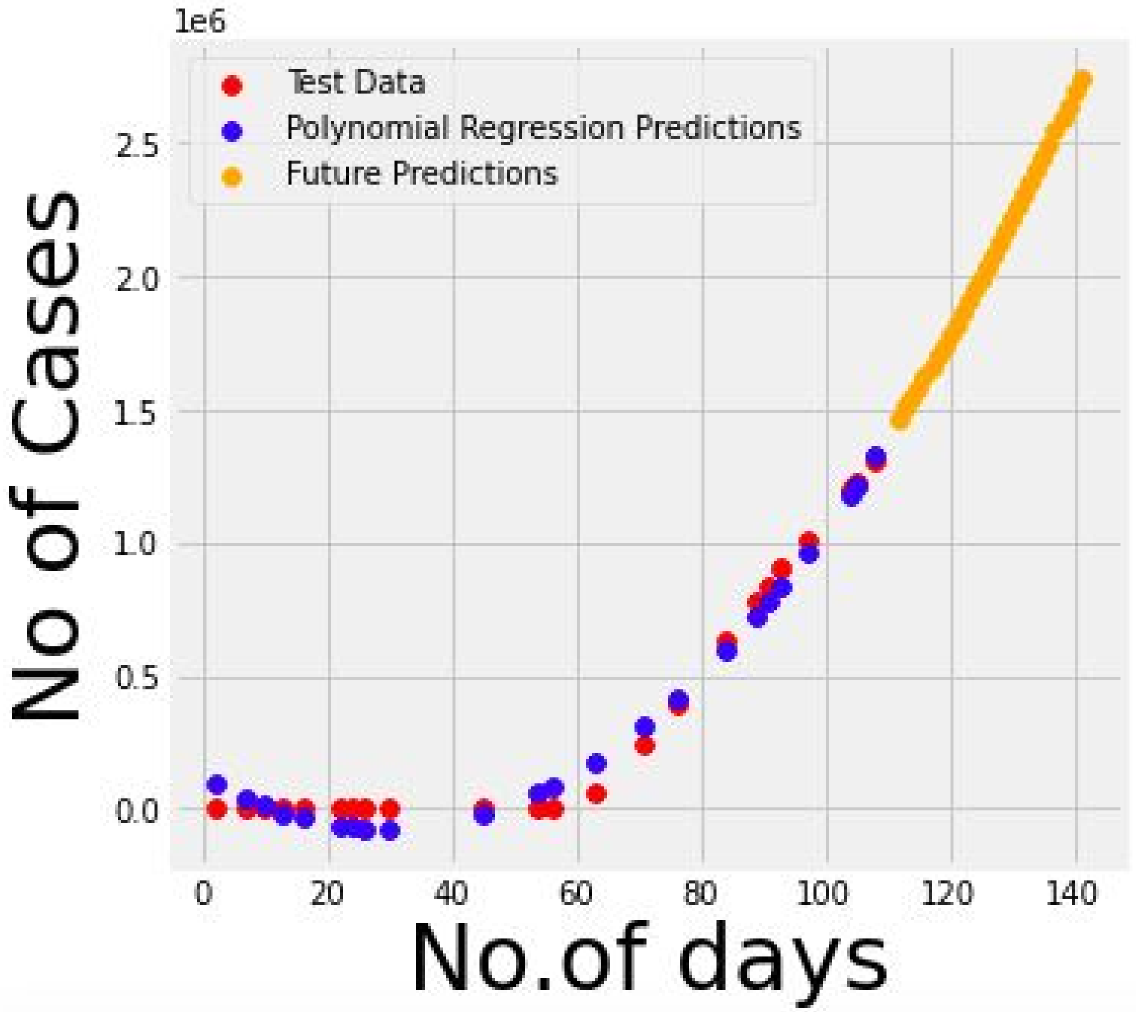
Xaxis values starts values from 113 - represents the May 13 (No of days from 22nd Jan to May 13th)

Xaxis values ends with 143 - represents future 30 days(113+30 = 143)

6.5 Visualisation of Actual Test Data, Predicted Test Data and Future prediction for USA

We have used polynomial regression with degree 2, to predict the increase in the number of corona cases for the USA. Using Support Vector Regression for the USA is giving us incorrect results. So, we tried using polynomial regression with a degree of 2. Below are the results:

For Simplicity, we have included all the results in the single graph for the better understanding of the results.



No of corona cases in USA (As of 12 th May 2020)

Actual Value:	**1390406**
Predicted Value:	**1322856**

Based on our prediction, On the 30th day, the number of recovered cases from corona can be around 2745544

7. Implications

6.1 Benefits of the Project:

- This project helps in the prediction of the coronavirus cases for the next 30 days, all over the world.
- With this, we can also predict the increase in corona cases in the USA.
- By this, we can know how fast the corona virus is spreading all over the world.
- We can create awareness among people.
- We can also create awareness in government, so that they can take preventive measures to stop the spread of corona.

6.2 Lessons Learned:

Initially, I had no idea of a Machine learning algorithm. I started learning about machines from scratch. I bought some Udemy tutorials and through that I learnt everything step by step. In the start of the project, I am not even aware of what machine learning algorithm to use.

It was really an exciting experience doing this project. I am inspired to take up a Machine Learning Course for my next semester to learn deeply about Machine Learning Algorithms.

I tried my level best, and contributed my 100% for this project.

Now, I came to know about machine Learning, different types of machine learning Algorithms, differences between classification and regression algorithms -when to use what, creating test and train sets, building up the model, by choosing the appropriate parameters, and performing future predictions. In the future, I would also love to take up the project related to Classification Algorithms.

8. Conclusion

- Finally to conclude, we have performed prediction using SVR and Polynomial Regression Algorithm.
- SVR predictions are mainly for predicting the world case scenario, which includes confirmed, death and recovered cases.
- Polynomial Regression is used for the prediction of US Cases.
- Based on the results, we believe that our predictions were almost accurate, with some little differences from the actual values.
- This project can be further scalable, to include the predictions for various individual countries.

9. Appendix

- We have used Google Collab for our project. As we are two members in the team, we have chosen this, because it enables us to simultaneously work on the project from different locations.
- No Installation Required.
- We just need to have a google account. And we can easily create a Google Collaboratory file in our google drive, just as Google docs.
- We will provide both .py file as well as .ipynb file along with this report, so as to run on google collab.
- .ipynb can be uploaded to google collab directly and the results of the projects can be easily checked.

10. References

1. **CSSEGISandData. (2020, May 15). CSSEGISandData/COVID-19. Retrieved from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series**
2. **Learn and Build Random Forest Algorithm Model in Python - Intellipaat. (2020, April 6). Retrieved from https://intellipaat.com/blog/what-is-random-forest-algorithm-in-python/**
3. **Brownlee, Jason. "Your First Machine Learning Project in Python Step-By-Step." Machine Learning Mastery, 17 Jan. 2020, machinelearningmastery.com/machine-learning-in-python-step-by-step/.**
4. **Brownlee, Jason. "How to Perform Data Cleaning for Machine Learning with Python." Machine Learning Mastery, 7 May 2020, machinelearningmastery.com/basic-data-cleaning-for-machine-learning/.**

11. Source Code

SourceCode.zip