INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL

MINOR PROJECT

ON

PREDICTION AND FORECASTING OF CORONAVIRUS (COVID-19) OUTBREAK

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF TECHNOLOGY

SUBMITTED BY:-SUBMITTED TO:-

Deepak kumar
(17U03016)
Harsh Soni
(17U03032)
Tanuj Kumar
(17U03040)

UNDER THE GUIDANCE OF:-

Dr.Priyank Jain

Dr.Nitesh KBhardhwaj

Dr.Yadunath Pathak

SESSION 2019-2020

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that **Harsh Soni, Deepak Kumar , Tanuj Kumar** students of B.Tech 3rd year(INFORMATION TECHNOLOGY),have successfully completed their project "**PREDICTION AND FORECASTING OF CORONAVIRUS (COVID-19) OUTBREAK**" in partial fulfillment of their minor project in Information Technology.

Dr.Priyank Jain
Dr.Nitesh K BhardhwajDr.Priyank Jain
Dr.Yadunath Pathak
(Project Mentor) (Project Coordinator)

SESSION 2018-19

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
BHOPAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We,hereby,declare that the following report which is being presented in the Minor Project "***Prediction and Forecasting of Coronavirus (COVID-19 Outbreak***"is the partial fulfilment of the requirements of the third year(sixth semester) Minor Project in the field of Information and Technology.It is an authentic documentation of our original work carried out under the able guidance of **Dr.Nitesh K Bhardhwaj, Dr.Yadunath Pathak &Dr.Priyank Jain** and the dedicated co-ordination of **Dr.Priyank Jain** .The work has been carried out entirely at Indian Institute of Information Technology, Bhopal. The following project and its report,in part or whole,has not been presented or submitted by us for any purpose in any other institute or organization.

We,hereby,declare that the facts mentioned above are true to best of our knowledge.In case of any unlikely discrepancy that may possibly occur,we will be the ones to take responsibility.

Harsh Soni (17U03032)
Deepak Kumar (17U03016)
Tanuj Kumar (17U03040)

AREA OF WORK

Our project will mainly focus on DATA SCIENCE, DATA ANALYSIS and MACHINE LEARNING using a set of tools and technologies that are required for development of large-scale architectures.
While one may argue that the same thing could have been done using big data, but that's where the real catch is.

As it is already mentioned that the tools and architecture we'll be using mainly focus on large scale deployments.

It's been a fairly long period since humans started computing and then extracting useful insights from the data, but in past decade the immense growth in data has posed a serious challenge for computing that's where big data came and with time we also acquired the capacity to process real time massive data sets and ever since it's just keeps growing bigger and bigger.

In our project we'll be looking at architecture of real time data pipeline and We'll also look at analysis of big data using spark mlib.

We will be using different machine learning algorithms like SVM, Polynomial regression and Bayesain Ridge Regression in order to gain best results and predictions of this global pandemic.

Here we assume that the data provided is accurate and won't change with time under this covid 19 outbreak to hamper the prediction.

ACKNOWLEDGEMENT

With due respect,we express our deep sense of gratitude to our respected guide Dr.Nitesh K Bhardhwaj, Dr.Priyank Jain &Dr.Yadunath Pathak for their invaluable support and guidance. We are thankful for the encouragement that they have given us in completing this project successfully. Their rigorous evaluation and constructive criticism wasof great assistance.

It is imperative for us to mention the fact that this minor project could not have been without the periodic suggestion and advice of our project co-ordinator Dr.Priyank Jain.

We are also grateful to our respected Director Dr.NarendraSingh Raghuvanshi for permitting us to utilize all the necessary facilities of the college.Needless to mention is the additional help and support extended by our respectedNodalOfficer, Dr.Meenu Chawla, in allowing us to use the department laboratories and other services.

We are also thankful to professor in chargeexaminationDr. DheerajK. Agrawal who continuously supported and encouraged us by his advices and also helped us in our project presentation that has improved our presentation skills.

We are also thankful to Student Welfare in charge Dr.JaytrilokChoudhary for constantly motivating us to work harder for the completion of the project. We extend our sincere thanks to Dr. Amit Bhagat who co-operated with us nicely for smooth development of this project. We would also like to thanks all the other faculty, staff members and laboratory attendants of our department for their kind co-operation and help. Last but certainly not the least, we would like to express our deep appreciation towards our family members and batch mates for providing the muchneeded support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES AND ILLUSTRATIONS

ABSTRACT

Data science has attracted a lot of attention, promising to turn vast amounts of data into useful predictions and insights. Our perspective is that data science is the child of statistics and computer science. While it has inherited some of their methods and thinking, it also seeks to blend them, refocus them, and develop them to address the context and needs of modern scientific data analysis.

A **Machine Learning model** is a set of assumptions about the underlying nature the data to be trained for. The **model** is used as the basis for determining what a **Machine Learning** algorithm should learn. A good **model**, which makes accurate assumptions about the data, is necessary for the **machine** to give good results.Machine Learning is the process of predicting things, usually based on what they've done in the past.Machine Learning tries to find relationships in your data that can help you predict what will happen next [5].The process of making a model starts from Business understanding which means intension of project is outlined. Data is understood after that and collected from different or same source(s). Data preparation is most one of the most difficult step as the raw data get cleaned using different methods and checked for questionable, missing, or ambiguous cases. Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model.

What will be the global impact of the novel corona virus (COVID-19)? Answering this question requires accurate forecasting the spread of confirmedcasesaswellasanalysisofthe number of deaths and recoveries. Forecasting, however, requires ample historical data. At an equivalent time, the prediction is uncertain as the future rarely repeats itself in the same way as that in the history. Moreover, forecasts are influenced by the reliability of the data, vested interests, and what variables are being predicted. Also, psychological factors play a significant role in how people perceive and react to the danger from the disease and the fear that it may affect them personally. This paper introduces an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so [1]. Assuming that the data used is reliable and accurate and that the future will continue to follow the same past pattern of the disease, our forecasts suggest an unbroken

increase within the confirmed COVID-19 cases with size able associate dun certainty.The risks are far from symmetric as underestimating its spread like epidemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. This paper describes the timeline of a live forecasting exercise with massive potential implications for planning and decisionmaking and provides objective forecasts for the confirmed cases , active cases , deaths occurred and recovered cases of COVID19 worldwide on the continental map Data Science and Machine Learning Techniques.

INTRODUCTION

World is moving through a very distressing stage by the spread of novel coronavirus (SARS-CoV-2). It is a highly contagious disease and the World Health Organization (WHO) has declared it as a global public health emergency. It is originated in Wuhan, Hubei Province, People's Republic of China (PRC) in late December 2019, when a case of unidentified pneumonia was reported. PRC Centers for Disease Control (CDC) experts declared that pneumonia as novel coronavirus pneumonia (NCP) as caused by a novel coronavirus and WHO officially named the disease COVID-19. However, the International Committee on Taxonomy of Viruses (ICTV) named the virus as severe acute respiratory syndrome coronavirus 2 [1]. This is a class of -coronavirus and has many potential natural hosts, intermediate hosts and final hosts. Due to these characteristics, there is a great challenge for prevention and treatment of the virus infection. Despite of the large number of cases worldwide and low mortality rate compared to SARS and the middle east respiratory syndrome, this virus has high infectivity and transmissibility. Preventive measures for COVID-19 include maintaining social distancing, washing hands frequently, avoiding touching the mouth, nose, and face.

Medical professionals and others must get correct and up-to-date information about how the coronavirus situation changes day by day. Several organizations, including Johns Hopkins University, IBM and Tableau, have released interactive databases that offer real-time views of what's happening with the virus.
Many of these sources pull from data provided by trusted bodies such as the U.S Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO). They also include direct links to those places so that people have quick, easy access to reliable information. We are buiding a model that can inform people the number of confirmed cases, fatalities and recoveries. Then, whether a person is on the front lines of the coronavirus fight or a concerned citizen trying to stay informed, they can get all or most of the information they need in one place.

In response to the sudden explosion of nCoV-2019 , the Research and Development wing of WHO is actively trying to find the appropriate diagnostics and vaccination . Day by day, medical experts are working rigorously and trying to explore its severe consequence such as human body respiratory symptoms get affected, particularly in the elderly people imparting mortality rate.

In this paper, we provide statistical forecasts for the confirmed cases of COVID-19 using robust time series models , machine learning techniques , and we analyze the trajectory of confirmed cases, active cases, deaths , recoveries ,mortality rate and recovery rate of COVID-19 patients worldwide via hovering over the world map.

LITERATURE REVIEW

The accuracy of traditional forecasting largely depends on the availability of data to base its predictions and estimates of uncertainty. In outbreaks of epidemics there is no data at all in the beginning and then limited as time passes, making predictions widely uncertain. On February 18, 2020, a New York Times article cautioned against excessive optimism about the crisis peaking, even though there were close to 50 days since the virus had been identified. Besides, there are concerns that the data may not be reliable, as was the case of bird flu and SARS when the number of affected people and deaths were misreported to hide the extent of the epidemic. Similarly, in the case of COVID-19, the reporting did not reflect the correct numbers as well when on the February 13 a new category of “clinically diagnosed” was added to “lab-confirmed” ones. Such problems decrease forecasting accuracy and increase uncertainty, making the drawing of definite conclusions more difficult. Related to forecasting accuracy and uncertainty, there is a more severe problem that has to do the perception of epidemics and pandemics. Politicians are concerned with regards to the measures to be taken while the general population fears about the impact on the epidemic on their health/lives. Furthermore, the pharmaceutical firms are working on vaccinations for the new virus with considerable commercial interest [2]. This was the case with SARS when governments persuaded on the severity of the virus bought large numbers of vaccines that were never used as its spread stopped without the need to vaccinate people.

Of course, the big problem is the asymmetry of risks and the irrational fear of a pandemic with its possible catastrophic consequences, as happened with the 1918 Spanish flu that killed an estimated 50 million worldwide. In contrast, the SARS killed a total of 774 in 2003 and the bird flu around 100 in 1997. COVID-19 has resulted in an estimated 5.8 thousand deaths until now (15/03/2020). At the same time, there is much less concern over the seasonal flu that kills about 646,000 people worldwide each year [4]. Medical predictions are often not accurate while their uncertainty is seriously underestimated . Predicting the future of epidemics and pandemics is much more difficult as the number of cases to be studied can be measured in one hand. At one end of the scale is the case of SARS where the fear of becoming a pandemic was overblown, resulting in overspending and the application of restrictive measures to be contained that it turned out to be unnecessary [3]. At the other end is the Spanish flu that turned out to be a serious pandemic with catastrophic consequences, arguably in a different era when communication and the ability to raise public awareness (and possibly exaggerated fear) were limited. Despite the inaccuracies associated with medical predictions, still forecasting is invaluable in allowing us to better understand the current situation and plan for the future.

METHODOLOGY AND WORK DESCRIPTION

For developing this data science project we used the methodology which is widely used by data science community which follow 6 stages follows [5]:-

1. **Business Understanding/ Problem Understanding:-**This stage is the most important because this is where the intention of the project is outlined. In this step we predefined our objectives i.e. to help in the this pandemic to analyse, predict and visualize the covid-19 fatalities,recovery, deaths and confirmed cases to tackle the global crisis better.

2. **Data Understanding:-**Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods.The data were collected from various sources like WHO , JHU and worldmeter.info .

Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/5/20	5/6/20	5/7/20	5/8/20	5/9/20	5/10/20	5/11/20	5/12/20
Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	3224	3382	3563	3778	4033	4402	4687	4963
Albania	41.1533	20.1683	0	0	0	0	0	0	...	820	832	842	850	856	868	872	876
Algeria	28.0339	1.6596	0	0	0	0	0	0	...	4838	4997	5182	5369	5558	5723	5891	6067
Andorra	42.5063	1.5218	0	0	0	0	0	0	...	751	751	752	752	754	755	755	758
Angola	-11.2027	17.8739	0	0	0	0	0	0	...	36	36	36	43	43	45	45	45

Figure 1 : Understanding and Analyzing Data-Set

3. **Data Preparation** :- Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases.

4. **Modeling:-**Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary.

```
total_deaths.append(death_sum)
total_recovered.append(recovered_sum)
total_active.append(confirmed_sum-death_sum-recovered_sum)

# calculate rates
mortality_rate.append(death_sum/confirmed_sum)
recovery_rate.append(recovered_sum/confirmed_sum)

# case studies
india_cases.append(confirmed_df[confirmed_df['Country/Region']=='India'][i].sum())
china_cases.append(confirmed_df[confirmed_df['Country/Region']=='China'][i].sum())
italy_cases.append(confirmed_df[confirmed_df['Country/Region']=='Italy'][i].sum())
us_cases.append(confirmed_df[confirmed_df['Country/Region']=='US'][i].sum())
spain_cases.append(confirmed_df[confirmed_df['Country/Region']=='Spain'][i].sum())
france_cases.append(confirmed_df[confirmed_df['Country/Region']=='France'][i].sum())
germany_cases.append(confirmed_df[confirmed_df['Country/Region']=='Germany'][i].sum())
uk_cases.append(confirmed_df[confirmed_df['Country/Region']=='United Kingdom'][i].sum())
russia_cases.append(confirmed_df[confirmed_df['Country/Region']=='Russia'][i].sum())

india_deaths.append(deaths_df[deaths_df['Country/Region']=='India'][i].sum())
china_deaths.append(deaths_df[deaths_df['Country/Region']=='China'][i].sum())
italy_deaths.append(deaths_df[deaths_df['Country/Region']=='Italy'][i].sum())
us_deaths.append(deaths_df[deaths_df['Country/Region']=='US'][i].sum())
spain_deaths.append(deaths_df[deaths_df['Country/Region']=='Spain'][i].sum())
france_deaths.append(deaths_df[deaths_df['Country/Region']=='France'][i].sum())
germany_deaths.append(deaths_df[deaths_df['Country/Region']=='Germany'][i].sum())
uk_deaths.append(deaths_df[deaths_df['Country/Region']=='United Kingdom'][i].sum())
russia_deaths.append(deaths_df[deaths_df['Country/Region']=='Russia'][i].sum())

india_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='India'][i].sum())
china_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='China'][i].sum())
italy_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Italy'][i].sum())
us_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='US'][i].sum())
spain_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Spain'][i].sum())
france_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='France'][i].sum())
germany_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Germany'][i].sum())
uk_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='United Kingdom'][i].sum())
russia_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Russia'][i].sum())
```

Figure 2 : Preparation of the Data for Machine Use

5. **Evaluation:-** The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

6. **Deployment:-** In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

Flexibility is required at each step along with communication to keep the project on track. At any of the six stages, it may be necessary to revisit an earlier stage and make changes. The key point of this process is that it's cyclical; therefore, even at the finish you are having another business understanding encounter to discuss the viability after deployment. The journey continues.

TOOLS AND TECHNOLOGY USED

- **Jupyter Notebook:-**The **Jupyter Notebook** is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

- **Python: It** is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language **used** by data scientist for various **data science** projects/application. **Python** provide great functionality to deal with mathematics, statistics and **scientific** function.

- **Matplotlib:-**Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

- **Scikit-learn** : It is a free machine **learning** library for **Python**. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports **Python** numerical and scientific libraries like NumPy and SciPy .

- **Machine Learning:-**Machine Learning is the process of predicting things, usually based on what they've done in the past. Machine Learning tries to find relationships in your data that can help you predict what will happen next.

IMPLEMENTATION AND CODING

Analytic Approach:

The data that has been collected from the sources contain various attributes , the confirmed dataframe consist of country coulumn with respect to it a time series of covid-19 cases are labelled. Using this data we predicted the covid-19 cases worldwide.

SVM MODEL CODE:-

In Python, scikit-learn is a widely used library for implementing machine learning algorithms. SVM is also available in the scikit-learn library, and we follow the same structure for using it(Import library, object creation, fitting model and prediction.

```
In [16]: # svm_confirmed = svm_search.best_estimator_
svm_confirmed = SVR(shrinking=True, kernel='poly',gamma=0.01, epsilon=1,degree=4, C=0.1)
svm_confirmed.fit(X_train_confirmed, y_train_confirmed)
svm_pred = svm_confirmed.predict(future_forcast)
```

```
C:\Users\DELL\Anaconda3\lib\site-packages\sklearn\utils\validation.py:578: DataConversionWarning: A column
-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for exam
ple using ravel().
  y = column_or_1d(y, warn=True)
```

```
In [17]: # check against testing data
svm_test_pred = svm_confirmed.predict(X_test_confirmed)
plt.plot(y_test_confirmed)
plt.plot(svm_test_pred)
plt.legend(['Test Data', 'SVM Predictions'])
print('MAE:', mean_absolute_error(svm_test_pred, y_test_confirmed))
print('MSE:',mean_squared_error(svm_test_pred, y_test_confirmed))
```

```
MAE: 745500.934578296
MSE: 596128381134.7928
```

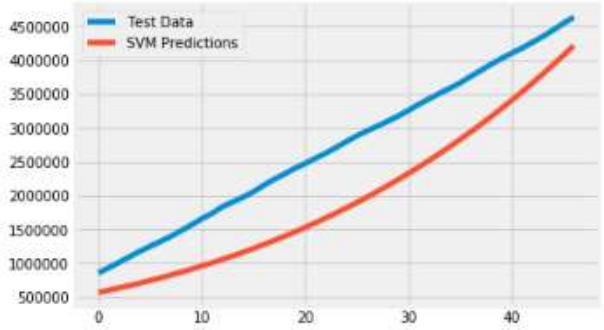


Figure 3 : Code for Support Vector Model Algorithm of Machine Learning

POLYNOMIAL REGRESSION PREDICTIONS CODE:-

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y —x) [5].

```
In [18]: # transform our data for polynomial regression
poly = PolynomialFeatures(degree=3)
poly_X_train_confirmed = poly.fit_transform(X_train_confirmed)
poly_X_test_confirmed = poly.fit_transform(X_test_confirmed)
poly_future_forcast = poly.fit_transform(future_forcast)

bayesian_poly = PolynomialFeatures(degree=4)
bayesian_poly_X_train_confirmed = bayesian_poly.fit_transform(X_train_confirmed)
bayesian_poly_X_test_confirmed = bayesian_poly.fit_transform(X_test_confirmed)
bayesian_poly_future_forcast = bayesian_poly.fit_transform(future_forcast)

In [19]: # polynomial regression
linear_model = LinearRegression(normalize=True, fit_intercept=False)
linear_model.fit(poly_X_train_confirmed, y_train_confirmed)
test_linear_pred = linear_model.predict(poly_X_test_confirmed)
linear_pred = linear_model.predict(poly_future_forcast)
print('MAE:', mean_absolute_error(test_linear_pred, y_test_confirmed))
print('MSE:',mean_squared_error(test_linear_pred, y_test_confirmed))

MAE: 539003.0790611312
MSE: 613121423470.118

In [20]: print(linear_model.coef_)

[[-6.55757132e+04  1.77039969e+04 -7.48960036e+02  9.62601905e+00]]

In [21]: plt.plot(y_test_confirmed)
plt.plot(test_linear_pred)
plt.legend(['Test Data', 'Polynomial Regression Predictions'])

Out[21]: <matplotlib.legend.Legend at 0x154fe23aa58>
```

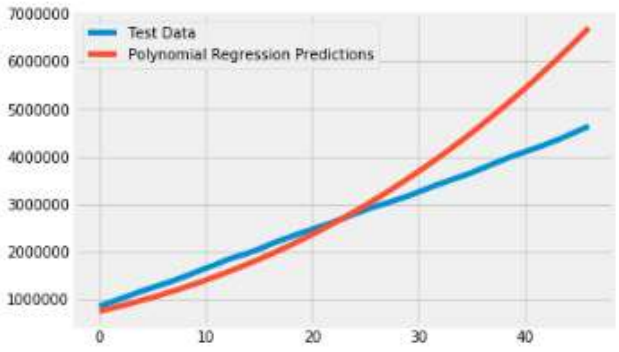


Figure 4 : Code for Polynomial Regression Predictions in Machine Learning

3) Bayesian Ridge Regression : Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response ‘y’ is assumed to drawn from a probability distribution rather than estimated as a single value. We have used it as it provides less errors in cases where prediction is uncertain [5].

Now the fitting of the data is done and then it is trained for getting the output.Now, Graphing the number of Confirmed Cases, Active Cases, Deaths , Recoveries ,Mortality Rate and Recovery Rate due to COVID-19 is done.

Graphing the number of confirmed cases, active cases, deaths, recoveries, mortality rate, and recovery rate

In [26]:
```
adjusted_dates = adjusted_dates.reshape(1, -1)[0]
plt.figure(figsize=(16, 9))
plt.plot(adjusted_dates, world_cases)
plt.title('# of Coronavirus Cases Over Time', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()

plt.figure(figsize=(16, 9))
plt.plot(adjusted_dates, total_deaths)
plt.title('# of Coronavirus Deaths Over Time', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()

plt.figure(figsize=(16, 9))
plt.plot(adjusted_dates, total_recovered)
plt.title('# of Coronavirus Recoveries Over Time', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()

plt.figure(figsize=(16, 9))
plt.plot(adjusted_dates, total_active)
plt.title('# of Coronavirus Active Cases Over Time', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Active Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```

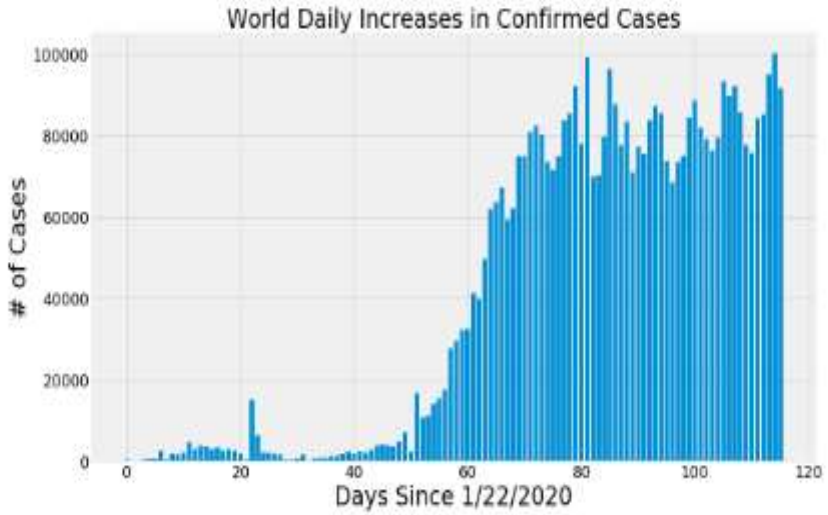
Figure 5 : Code for Graphing and Analyzing Data

/blob/master/coronavirus-covid-19-visualization-prediction.ipynb

In [27]:
```
plt.figure(figsize=(16, 9))
plt.bar(adjusted_dates, world_daily_increase)
plt.title('World Daily Increases in Confirmed Cases', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()

plt.figure(figsize=(16, 9))
plt.bar(adjusted_dates, world_daily_death)
plt.title('World Daily Increases in Confirmed Deaths', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()

plt.figure(figsize=(16, 9))
plt.bar(adjusted_dates, world_daily_recovery)
plt.title('World Daily Increases in Confirmed Recoveries', size=30)
plt.xlabel('Days Since 1/22/2020', size=30)
plt.ylabel('# of Cases', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



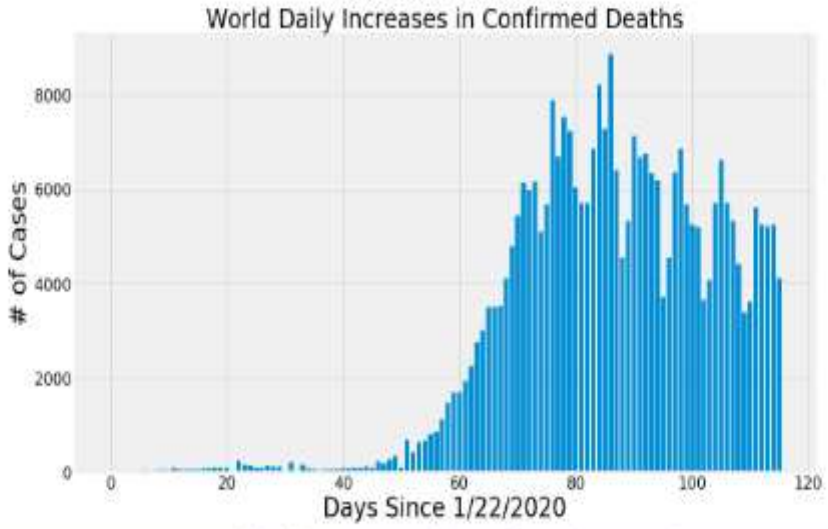


Figure 6 : Visualization of Data Set

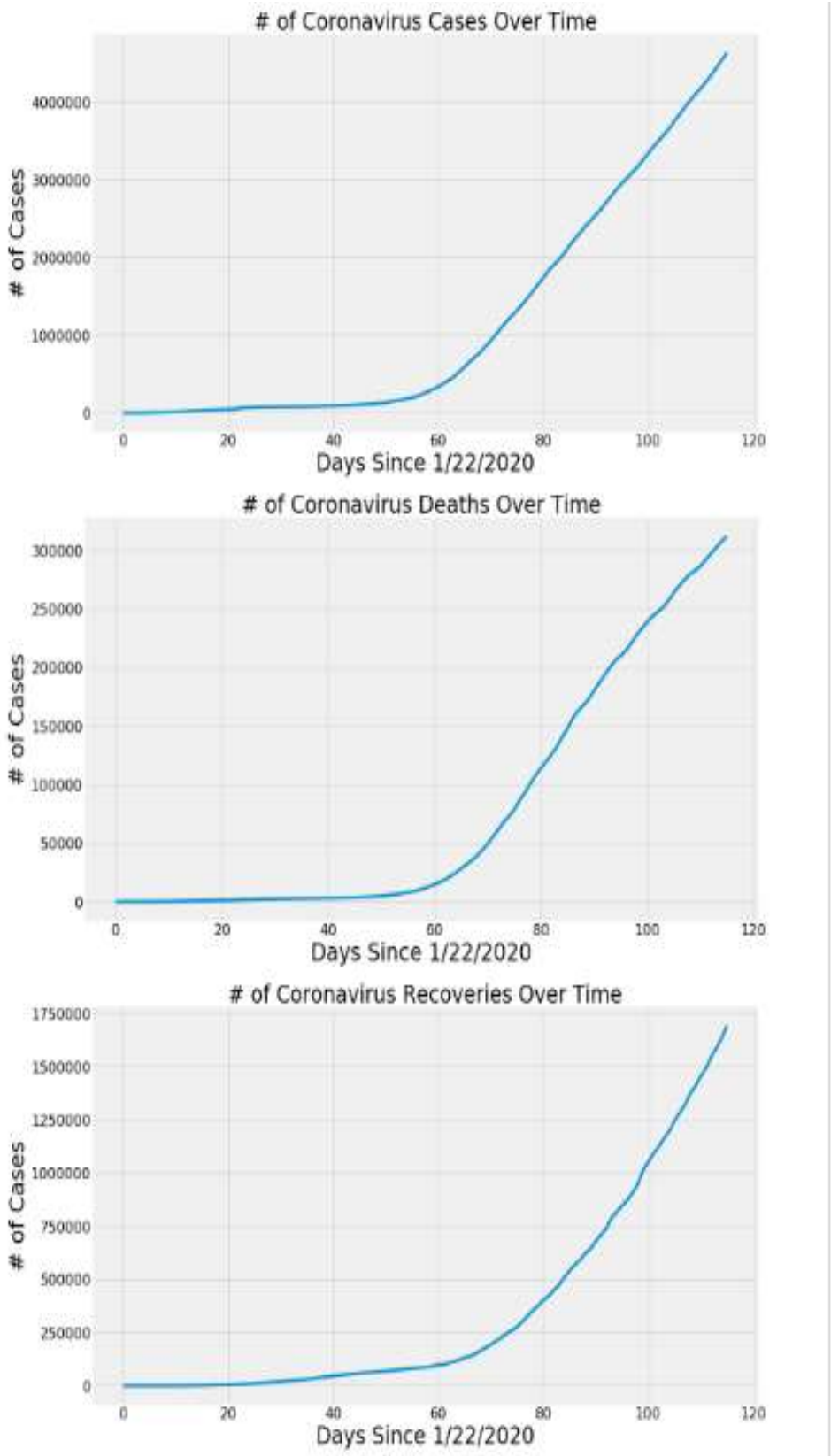


Figure 7 : Visualization of Data Set

Pie Chart Visualizations for COVID-19

In [71]:
```
def plot_pie_charts(x, y, title):
    c = random.choices(list(mcolors.CSS4_COLORS.values()),k = len(unique_countries))
    plt.figure(figsize=(20,15))
    plt.title(title, size=20)
    plt.pie(y, colors=c)
    plt.legend(x, loc='best', fontsize=15)
    plt.show()
```

In [72]:
```
plot_pie_charts(visual_unique_countries, visual_confirmed_cases, 'Covid-19 Confirmed Cases per Country')
```

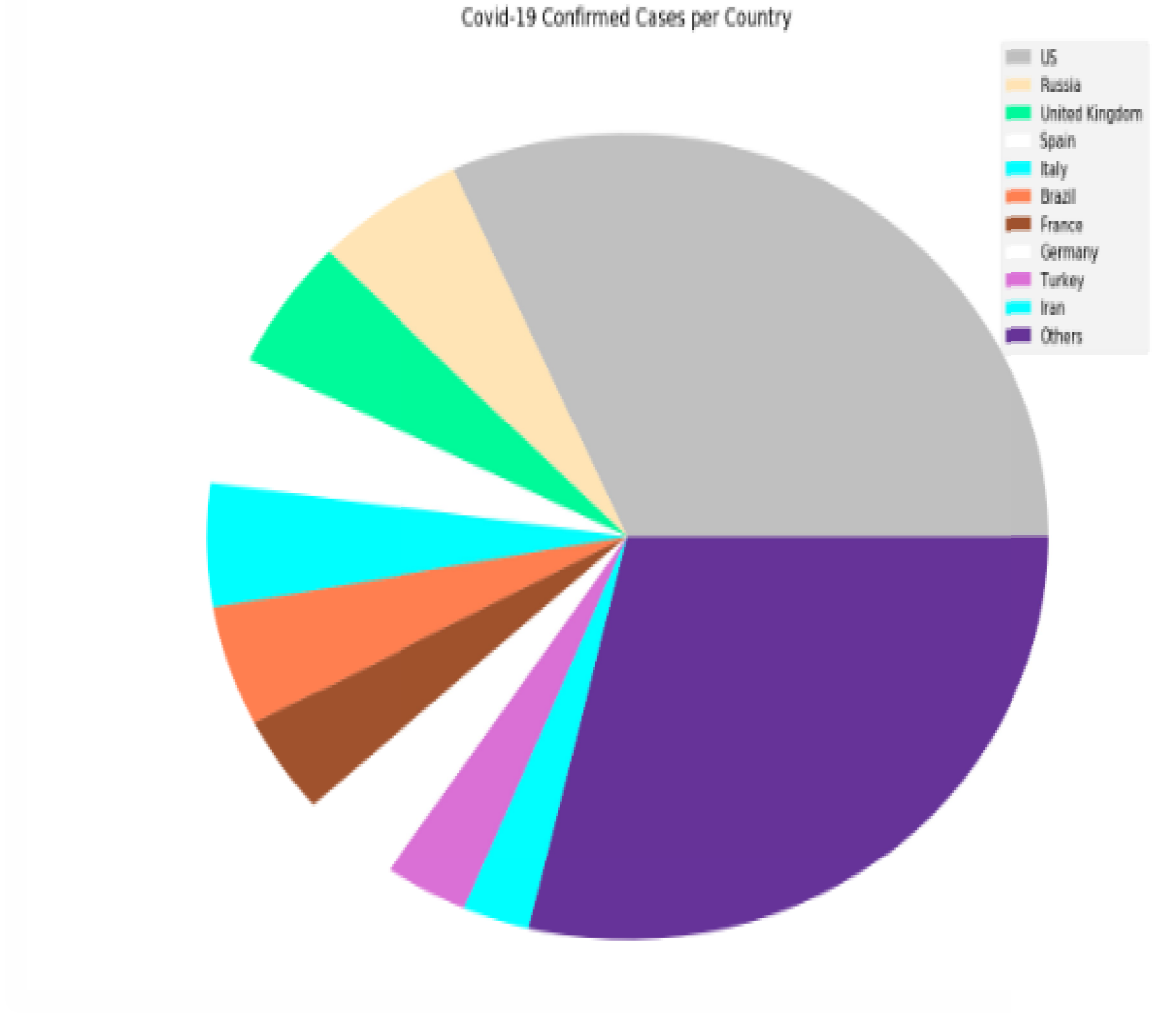


Figure 8 : Pie chart Visualization of data-set

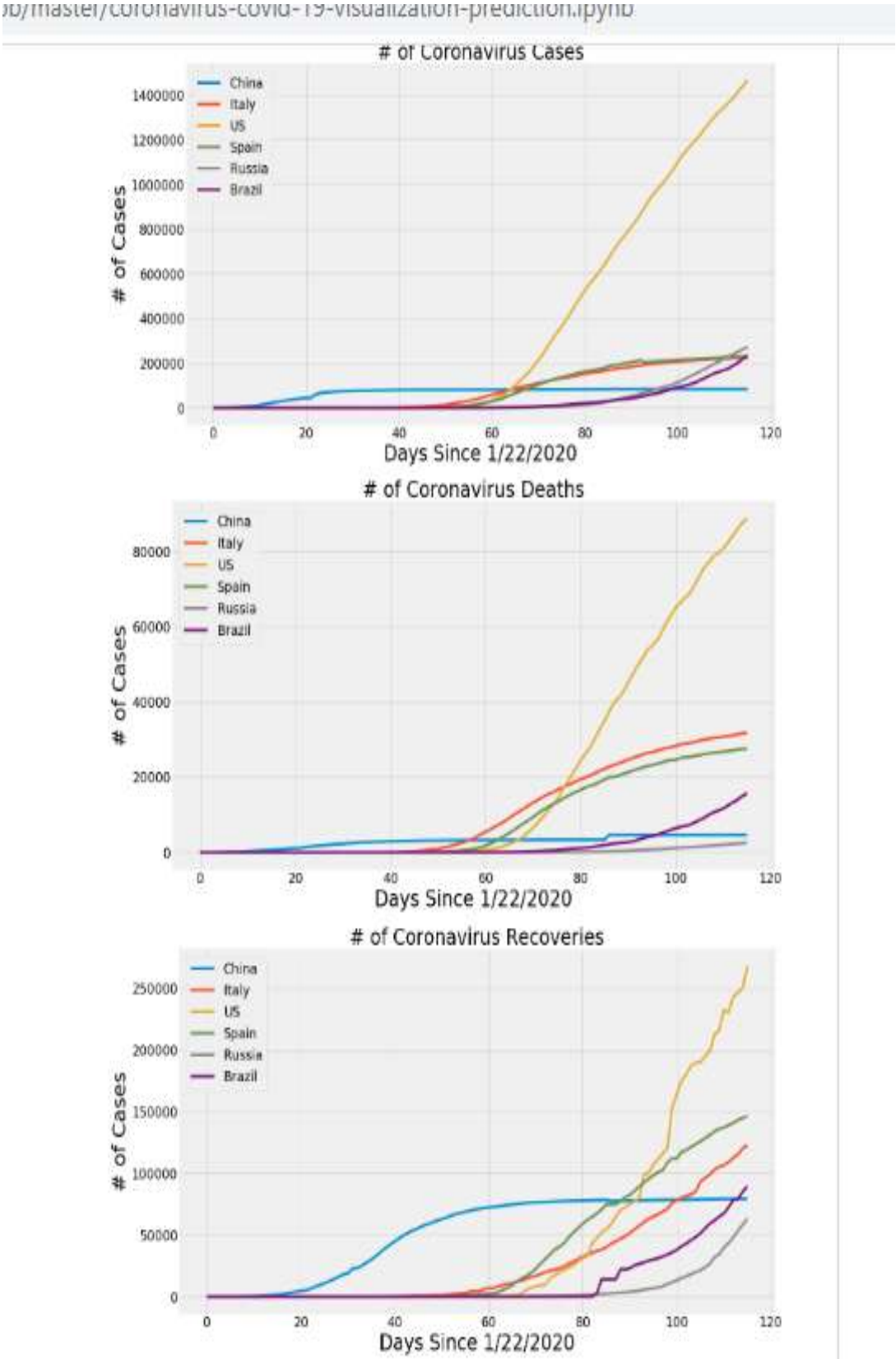


Figure 9 : Drawing the conclusions from visualized data

RESULT ANALYSIS

By analyzing and fitting the data covid-19 data into different models and after visualizing them results were pretty guessable.

We have visualized almost all the countries in the world and have predicted the following scenarios through different predictions.

- The predictions given by the SVM were somewhat accurate(67.98%) till the day 60-63 since 22/1/2020 after that it under performed but surprisingly it got more accurate predictions after around 118 days since 1/22/2020

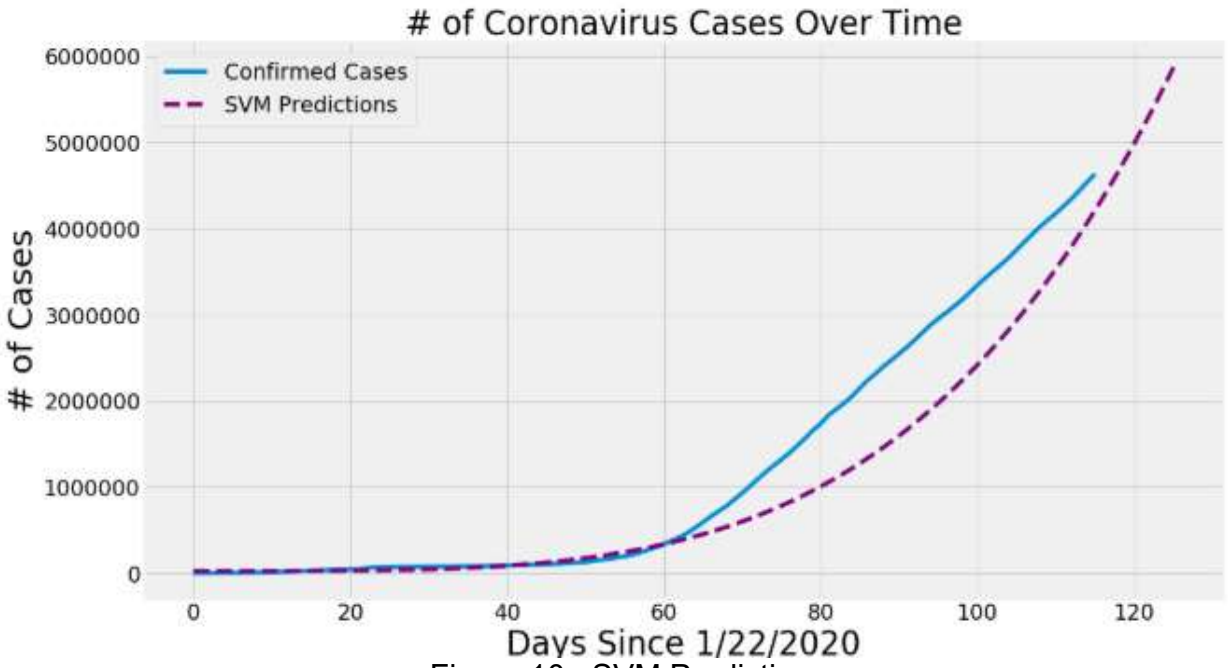


Figure 10 : SVM Prediction

- The predictions given by the Polynomial Regression were having high accuracy(about 78.4%) till day 95-96 since 1/22/2020 but after that the prediction performed with slightly greater error. This may be due to government authority actions like lock down ,quarantining suspected people which helped in reducing the confirmed cases worldwide.

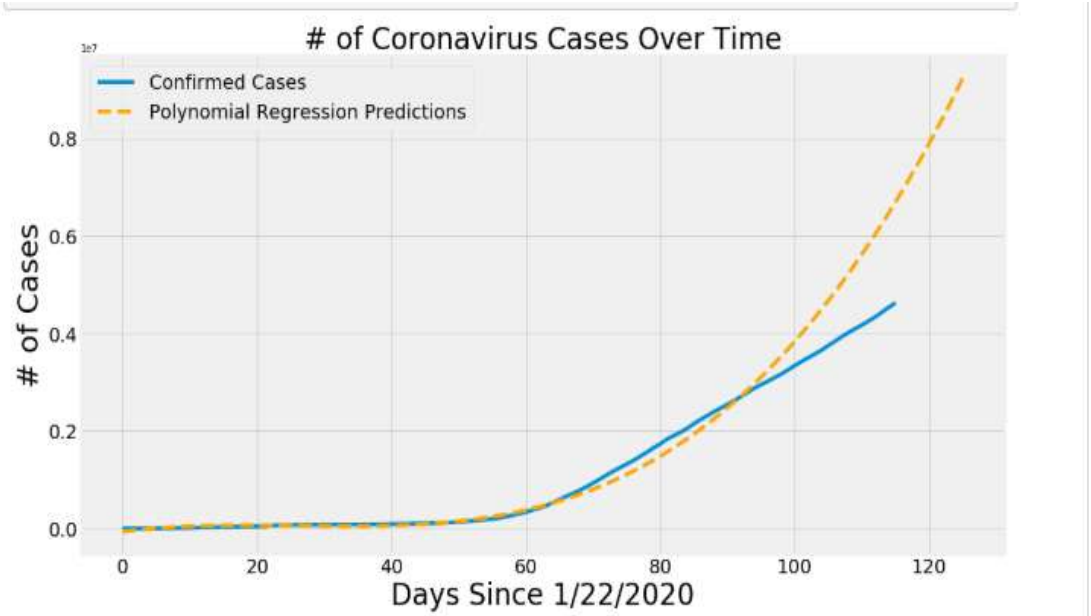


Figure 11 : Polynomial Regression Predictions

- The prediction by the Baysian Ridge Regression has the highest accuracy(79.34%) rate as it showed less or minute error till 100th day since 1/22/2020. And after that the model performed poorly. This may be due to government authority actions like lock down ,quarantining suspected people which helped in reducing the confirmed cases worldwide.

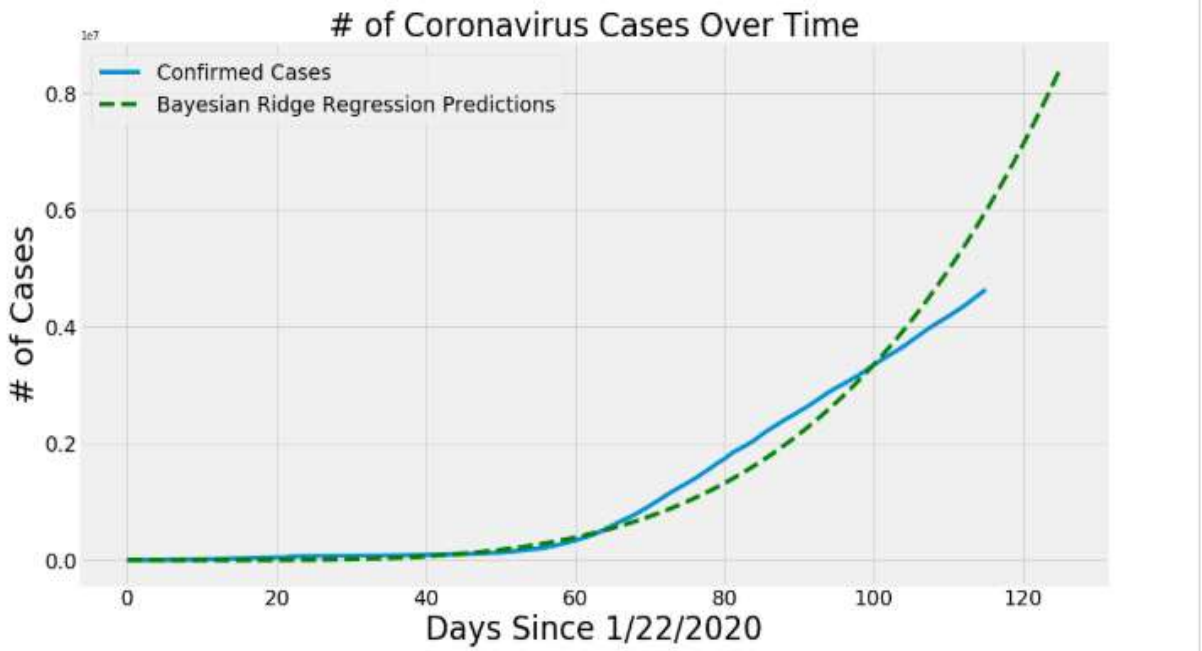


Figure 12 : Bayesian Ridge Regression Prediction

CONCLUSION AND FUTURE SCOPE

The uncertainty surrounding an unknown, novel coronavirus can spark a global alarm, leading a Harvard Professor stating that 40-70% of the global population might be infected in the coming year which matches Chancellor Angela Merkel's warning regarding the effects of the novel coronavirus in Germany [7]. Norman, Bar-Yam and Taleb discuss the systemic risk of pandemics, the existence of fat-tailed processes due to global inter connectivity and the negatively biased estimates of spread, reproduction and mortality rates. On the opposite side, others are arguing about people overly panicking and neglecting the probabilities with the new virus being the first "infodemic" as a result of the hyper-connectivity offered by today's social media . The polarization of the opinions globally can be summarized by the quotes of three renowned personalities:

• Elon Musk: "The coronavirus panic is dumb".
• Nassim Nicholas Taleb: "Saying the coronavirus panic is dumb is dumb".
• Bill Gates: "I hope it's not that bad, but we should assume it will be until we know otherwise".

Regardless of what one's beliefs are, we believe that forecasts and their associated uncertainty can and should be an integral part of the decision-making process, especially in high-risk cases. Apart from the significant public health concerns, the dangers imposed on global supply chains and the economy as a whole are also considerable. Risk-averse people can focus on the worst-case-scenarios and act accordingly. Deciding to discard any formal, statistical forecasts and acting conservatively, still implies an underlying forecasting process, even if this process is not formalized(personal judgment/belief) [8].
This analysis is performed on covid-19 to help and see the prediction of the covid-19 cases on the basis of the data obtained till date. This analysis is not as accurate as we wanted it to be as the Covid-19 is global pendemic and medical professionals and data scientist are still trying to get a cure for it. Till now we have not found abby cure or any 100% accurate predictor machine as this disease can be controlled by social distancing and hygine around you which reduces the chance s of getting infected.
We used three diffent Machine Learning algorithms to predict the cases as soon as possible . We will try to improve this project to its best in future with more resources.For future , the of use more machine learning algorithm , neural networks,deep learning and feature engineering concepts can definitely improve the efficiency of the prediction by the machine.

References:

- [1] Wang V. Coronavirus epidemic keeps growing, but spread in China slows. New York Times. [https://www.nytimes.com/2020/02/18/world/asia/china-coronaviruscases.html?referringSource=articleShare] Accessed: 2020-02-19.
- [2] Medicine Net. Flu kills 646,000 people worldwide each year: Study finds. [https://www.medicinenet.com/script/main/art.asp?articlekey=208914] Accessed: 2020-02-19.
- [3] Hyndman RJ, Koehler AB, Snyder RD, Grose S. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting. 2002;18(3):439–454.
- [4] BBC News. Coronavirus: Up to 70% of Germany could become infected—Merkel [https://www.bbc.co.uk/news/world-us-canada51835856] Accessed: 2020-03-15.
- [5] For basic Information Regarding definitions [https://en.wikipedia.org]
- [6] Data is provided by Hopkins University [https://github.com/CSSEGISandData/COVID-19]
- [7] 2. World Health Organisation(WHO) : 2. World Health Organisation(WHO) [https://www.who.int/emergencies/diseases/novelcoronavirus-2019]
- [8] Learn more from the [https://www.cdc.gov/coronavirus/2019-ncov]