Prediction and Forecasting Of Worldwide Corona Virus (COVID-19) Outbreak

Harsh Soni, Deepak Kumar, and Tanuj Kumar

***Abstract*—What will be the global impact of the novel corona virus (COVID-19)? Answering this question requires accurate forecasting the spread of confirmed cases as well as analysis of the 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. 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 sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic 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 decision-making and provides objective forecasts for the confirmed cases , active cases , deaths occurred and recovered cases of COVID-19 worldwide on the continental map Data Science and Machine Learning Techniques.**

***Index Terms*—Corona Virus (COVID-19),SARS-CoV-2, Data Visualization, Accuracy, Data-Set, Curve Fitting, Infodemic.**

I. 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.

There is a lot of stress on the part of administration and health officials for accommodating patients with possible symptoms of COVID-19. So, for that some prediction tools must be used to know about the number of cases in coming days for making preparations at the administrative level [2].

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.

II. Literature Survey

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 [3]. 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.

Harsh Soni , Deepak Kumar ,Tanuj Kumar are student in the Department of Information Technology in Indian Institute Of Information Technology Bhopal, India. — Harsh Soni's e-mail: masterharshsoni@gmail.com — Deepak Kumar's e-mail: dkprajapat1212@gmail.com — Tanuj Kumar's e-mail: Tanujbathadiya@gmail.com

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. 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.

III. Software and Tools Description

A. 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. .

B. Python IDLE

IDLE is consolidated development environment for editing and executing python2 version or python 3 version programs. We obtain output of the program as a result.

C. 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.

D. Matplotlib

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

E. Scikit-learn

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

IV. Methodology Used

For developing this data science project we used the methodology which is widely used by data science community which follow 6 stages follows 3

A. 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.

B. 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 [6].

C. Data Preperation

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.

D. Modelling

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.

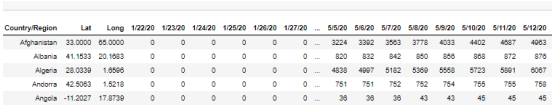
E. Evaluation

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.

F. Deployment

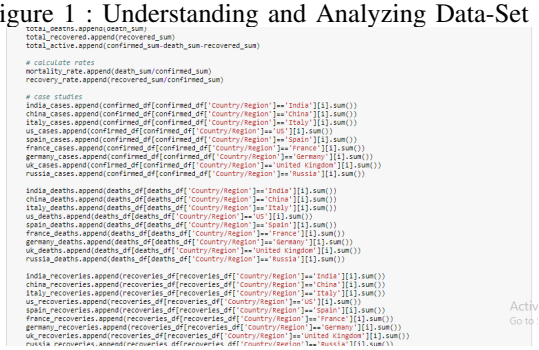
In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions,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



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	3392	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



```
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'][1].sum())
china_cases.append(confirmed_df[confirmed_df['Country/Region']=='China'][1].sum())
italy_cases.append(confirmed_df[confirmed_df['Country/Region']=='Italy'][1].sum())
us_cases.append(confirmed_df[confirmed_df['Country/Region']=='US'][1].sum())
spain_cases.append(confirmed_df[confirmed_df['Country/Region']=='Spain'][1].sum())
france_cases.append(confirmed_df[confirmed_df['Country/Region']=='France'][1].sum())
germany_cases.append(confirmed_df[confirmed_df['Country/Region']=='Germany'][1].sum())
uk_cases.append(confirmed_df[confirmed_df['Country/Region']=='United Kingdom'][1].sum())
russia_cases.append(confirmed_df[confirmed_df['Country/Region']=='Russia'][1].sum())

india_deaths.append(deaths_df[deaths_df['Country/Region']=='India'][1].sum())
china_deaths.append(deaths_df[deaths_df['Country/Region']=='China'][1].sum())
italy_deaths.append(deaths_df[deaths_df['Country/Region']=='Italy'][1].sum())
us_deaths.append(deaths_df[deaths_df['Country/Region']=='US'][1].sum())
spain_deaths.append(deaths_df[deaths_df['Country/Region']=='Spain'][1].sum())
france_deaths.append(deaths_df[deaths_df['Country/Region']=='France'][1].sum())
germany_deaths.append(deaths_df[deaths_df['Country/Region']=='Germany'][1].sum())
uk_deaths.append(deaths_df[deaths_df['Country/Region']=='United Kingdom'][1].sum())
russia_deaths.append(deaths_df[deaths_df['Country/Region']=='Russia'][1].sum())

india_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='India'][1].sum())
china_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='China'][1].sum())
italy_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Italy'][1].sum())
us_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='US'][1].sum())
spain_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Spain'][1].sum())
france_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='France'][1].sum())
germany_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Germany'][1].sum())
uk_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='United Kingdom'][1].sum())
russia_recoveries.append(recoveries_df[recoveries_df['Country/Region']=='Russia'][1].sum())
```

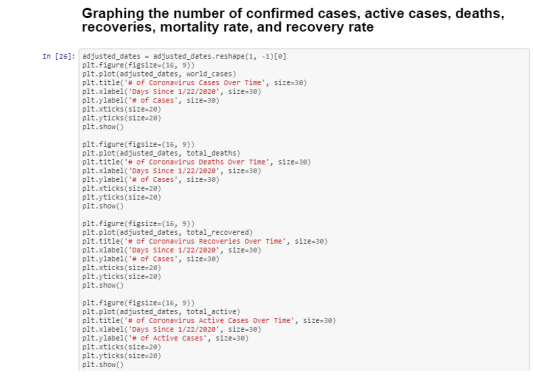
Figure 2 : Preparation of the Data for Machine Use

```
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 example 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

```
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



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

business understanding encounter to discuss the viability after deployment. The journey continues.

V. Implementation And Coding

We focus on the cumulative daily figures aggregated globally of the three main variables of interest: confirmed cases, deaths and recoveries. These were retrieved by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University 4. To forecast confirmed cases of COVID-19, we adopt simple time series forecasting approaches. We produce forecasts using models from the exponential smoothing family. This family has shown good forecast accuracy over several forecasting competitions and is especially suitable for short series. Exponential smoothing models can capture a variety of trend and seasonal forecasting patterns (such as additive or multiplicative) and combinations of those.

A. Using Analytic Approach

The data that has been collected from the sources contain various attributes , the confirmed data frame consist of country column with respect to it a time series of covid-19 cases are labeled. Using this data we predicted the covid-19 cases worldwide. We have used three different models under machine learning and data science as followed:-

1) SVM MODEL : 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)[5].

2) 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].

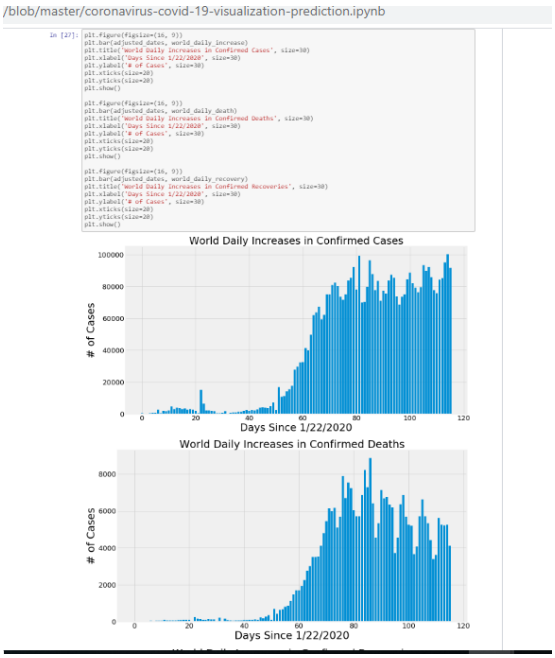


Figure 6 : Visualization of Data Set

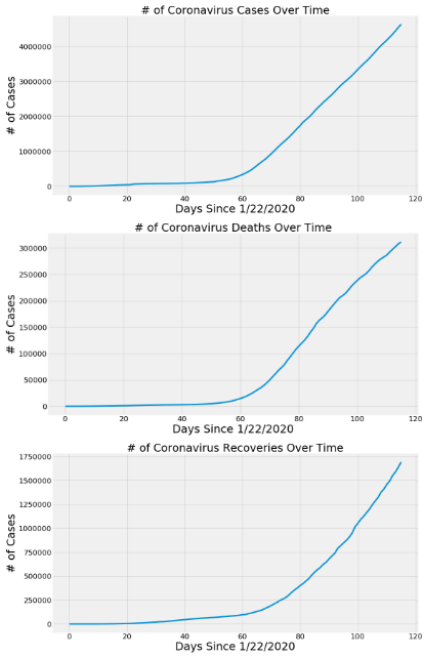


Figure 7 : Visualization of Data-Set

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 valueWe 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

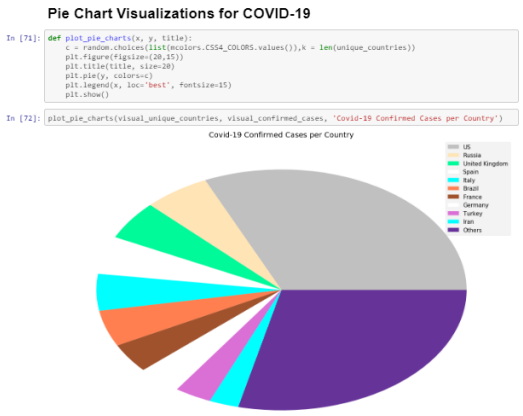


Figure 8 : Pie chart Visualization of data-set

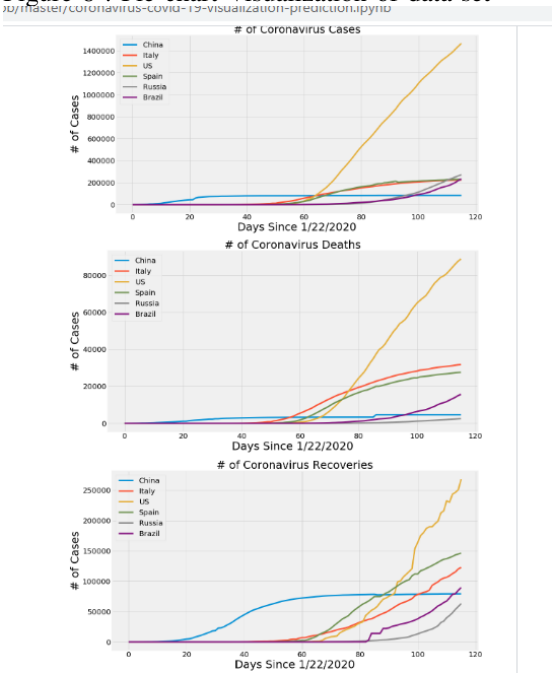


Figure 9 : Drawing the conclusions from visualized data

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.

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 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

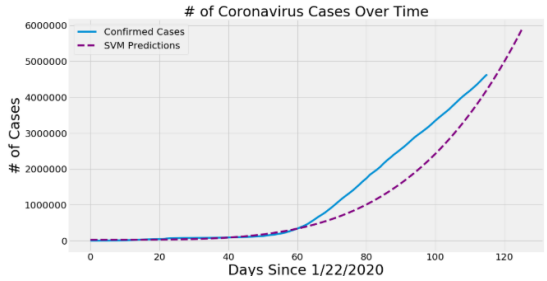


Figure 10 : SVM Prediction

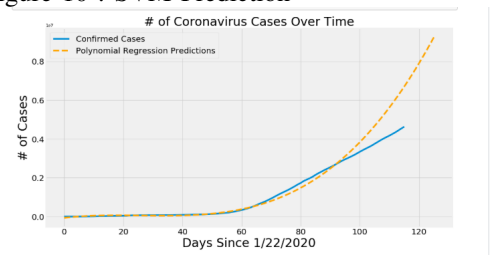


Figure 11 : Polynomial Regression Predictions

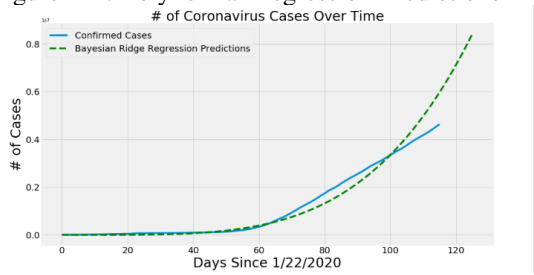


Figure 12 : Bayesian Ridge Regression Prediction

- The predictions given by the Polynomial Regression were having high accuracy 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.

- The prediction by the Bayesian Ridge Regression has the highest accuracy 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.

ACKNOWLEDGMENT

This research paper was mentored by Dr. Priyank Jain , Dr. Nitesh K. Bharadawaj and Dr. yadunath Pathak. We will like to thank our classmates from Indian Institute of Information Technology, Bhopal as they supported us technically whenever we needed that.Last but not the least i would like to thank my family who supported me day and night for the successfull completion of the project.

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-coronavirus-cases.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-canada-51835856] 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/novel-coronavirus-2019]

[8] Learn more from the [https://www.cdc.gov/coronavirus/2019-ncov]