A PROJECT ON

COVID-19 ANALYSIS AND PREDICTION

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYTICS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY

'Plot no R/2', Market yard road,
Behind hotel Fulera,Gultekdi
Pune – 411037.
MH-INDIA

SUBMITTED BY:

ROSHAN JOHN (63181)

UNDER THE GUIDANCE OF

Mrs. Suvrunda Nangare
Faculty Member
Sunbeam Institute of Information Technology, PUNE.

SUNBEAM

CERTIFICATE

This is to certify that the project work under the title ‘Covid-19 Analysis and Prediction’ is done by Roshan John in partial fulfilment of the requirement for the award of Diploma in Big Data Analysis Course.

Mrs. Suvrunda Nangare	**Mrs. Pradnya Dindorkar**
Project Guide	**Course Co-ordinator**

Date:

ACKNOWLEDGMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs Pradnya Dindorkar (Course Coordinator, SIIT, Pune) and Project Guide Mrs Suvrunda Nangare.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Roshan John
DBDA March 2022 Batch
SIIT Pune

Contents

List of Figures

1 Introduction

Infectious diseases are caused by various pathogens that can be transmitted from person to person, animal to animal, or person to animal. They can be transmitted in various ways, and the speed of transmission is fast. Early diagnosis of infectious diseases is crucial, and prevention and control are paramount.

In December 2019, unexplained viral pneumonia was reported in Wuhan, China. The virus, named the 2019 Novel Coronavirus (2019-nCoV) by the World Health Organization (WHO) on January 12, 2020, causes Corona Virus Disease 2019 (COVID-19), and is currently the seventh known species of coronavirus that can infect human beings.

The mortality rate of COVID-19 is approximately 2% to 4%, although this is an extremely early percentage and may change as more information becomes available. Meanwhile, this does not mean that the virus is not serious, it simply means that not everyone infected with it will face the worst outcome. We are currently in a tense state of prevention and control and are concerned about a global pandemic.

There are currently no specific treatments for COVID-19. However, many symptoms can be treated, and treatment must be given according to the clinical status of the patient. At present, owing to the widespread nature of COVID-19, factories are being shut down, schools are being suspended, and people are isolated in their own homes, significantly disrupting daily life. It is therefore extremely important to reasonably predict and analyze the development trend of this pandemic.

1.1 Objectives

The world has been suffering from the disease covid-19 for more than two years. We have seen the number of infected cases and deaths rising as time passes. Here in this project, we plan to predict the future confirmed cases, deaths and recovery using available data. We also plan to visualise this data and get inferences about how this disease is affecting various countries in the world.

In this study, the development trend analysis of the cumulative confirmed cases, cumulative deaths, and cumulative cured cases was conducted from January 22, 2020, to the present using Bayesian ridge regression, polynomial regression and support vector machine (SVM). An SVM with fuzzy granulation was used to predict the growth range of confirmed new cases, new deaths, and new cured cases.

1.2 Dataset Information

1.2.1 Dirty Data

time_series_covid19_confirmed_global.csv

time_series_covid19_deaths_global.csv

time_series_covid19_recovered_global.csv

These datasets provide the daily confirmed cases, deaths and recovered cases.

Rows: Countries and Provinces

Columns: The daily corresponding value from 22 January 2020 till the present date

1.2.2 Cleaned Data

countrywise.csv

It has 5 columns

Country/Region: Names of countries

Date: The dates starting from 22-01-2020 till the present

Confirmed: Confirmed Cases of that day

Deaths: Confirmed Deaths of that day

Recovered: Recovered Cases of that day

datewise.csv

It has 4 columns

Date: The dates starting from 22-01-2020 till the present

Confirmed: Confirmed Cases of that day

Deaths: Confirmed Deaths of that day

Recovered: Recovered Cases of that day

worldometer_data.csv

This is the dataset scrapped from Worldometer. It has columns.

The columns are Countries, Total Cases, Total Deaths, Total Recovered, Total Active, Total Cases, and Population.

All the columns contain the latest total data

2 Problem Definition and Algorithm

2.1 Problem Definition

The problem is quite straightforward. We are given data on the daily confirmed covid cases, deaths and recovery. We have to split this data into training data and testing data and perform the necessary algorithms and make a model. We want to fit a model to the training data that can forecast those cases as accurately as possible. Our metric of interest will be the Mean Absolute Error and R2 score value. We will also have to use the data to build various visualizations like graphs, charts, tables etc to get some inferences regarding the effects of covid-19 in various countries. This can be done using the various visualization libraries in python as well as using the tool Power Bi.

2.2 Algorithm Definition

SVM: SVM is a classic model that can be used not only for classification but also for regression. We do not delve into its theoretical derivation here, however. Unlike with a classification problem, the output of the regression problem is no longer a discrete value, but a continuous value. In reality, it is often impossible to accurately predict the value of COVID-19, and to this end, it is particularly important to predict the development trend and change space for the important parameters of this disease. Without considering other factors, time is an important independent variable affecting COVID-19.

Polynomial Regression: In polynomial regression, the relationship between the independent variable x and the dependent variable y is described as an nth-degree polynomial in x. Polynomial regression, abbreviated E(y —x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y. It usually corresponded to the least-squares method. According to the Gauss Markov Theorem, the least square approach minimizes the variance of the coefficients. This is a type of Linear Regression in which the dependent and independent variables have a curvilinear relationship and the polynomial equation is fitted to the data; we'll go over that in more detail later in the article. Machine learning is also referred to as a subset of Multiple Linear Regression. Because we convert the Multiple Linear Regression equation into a Polynomial Regression equation by including more polynomial elements.

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 be drawn from a probability distribution rather than estimated as a sin.

3 Experimental Evaluation

3.1 Methodology

The objective of this project is to predict the daily cases related to covid-19. The dataset is taken from github, which is updated daily and has 3 CSV files for confirmed covid cases, deaths and recovered. The data is merged to obtain one master data file and then the data preprocessing is carried out.

1. **Loading raw data**

```
conf_df = pd.read_csv(’time_series_covid19_confirmed_global.csv’)
deaths_df = pd.read_csv(’time_series_covid19_deaths_global.csv’)
recv_df = pd.read_csv(’time_series_covid19_recovered_global.csv’)


conf_df_long = conf_df.melt(id_vars=[’Province/State’, ’Country/Region’],value_vars=dates,
var_name=’Date’, value_name=’Confirmed’)
deaths_df_long = deaths_df.melt(id_vars=[’Province/State’, ’Country/Region’ value_vars=dates,
var_name=’Date’, value_name=’Deaths’)
recv_df_long = recv_df.melt(id_vars=[’Province/State’, ’Country/Region’value_vars=dates,
var_name=’Date’, value_name=’Recovered’)


full_table = pd.merge(left=conf_df_long, right=deaths_df_long, how=’left’,on=[’Province/State’,
’Country/Region’, ’Date’])
full_table = pd.merge(left=full_table, right=recv_df_long, how=’left’,on=[’Province/State’,
’Country/Region’, ’Date’])


print(full_table.shape)
full_table.head()
```

2. **Preprocessing**

The datatype of ’Date’ Column is changed from object to datetime

```
full_table[’Date’] = pd.to_datetime(full_table[’Date’])
```

The missing values are replaced by zeroes

```
full_table[’Confirmed’] = full_table[’Confirmed’].fillna(0)
full_table[’Deaths’] = full_table[’Deaths’].fillna(0)
full_table[’Recovered’] = full_table[’Recovered’].fillna(0)
full_table.isna().sum()
```

After the missing values are filled grouping is done

```
datewise = full_table.groupby(’Date’).agg(”Confirmed”: ”sum”, ”Deaths”:”sum”,
”Recovered”: ”sum”).reset_index()
countrywise = full_table.groupby([’Country/Region’, ’Date’]).agg(”Confirmed”:
”sum”, ”Deaths”: ”sum”, ”Recovered”: ”sum”).reset_index()
```

3.2 Flow Diagram

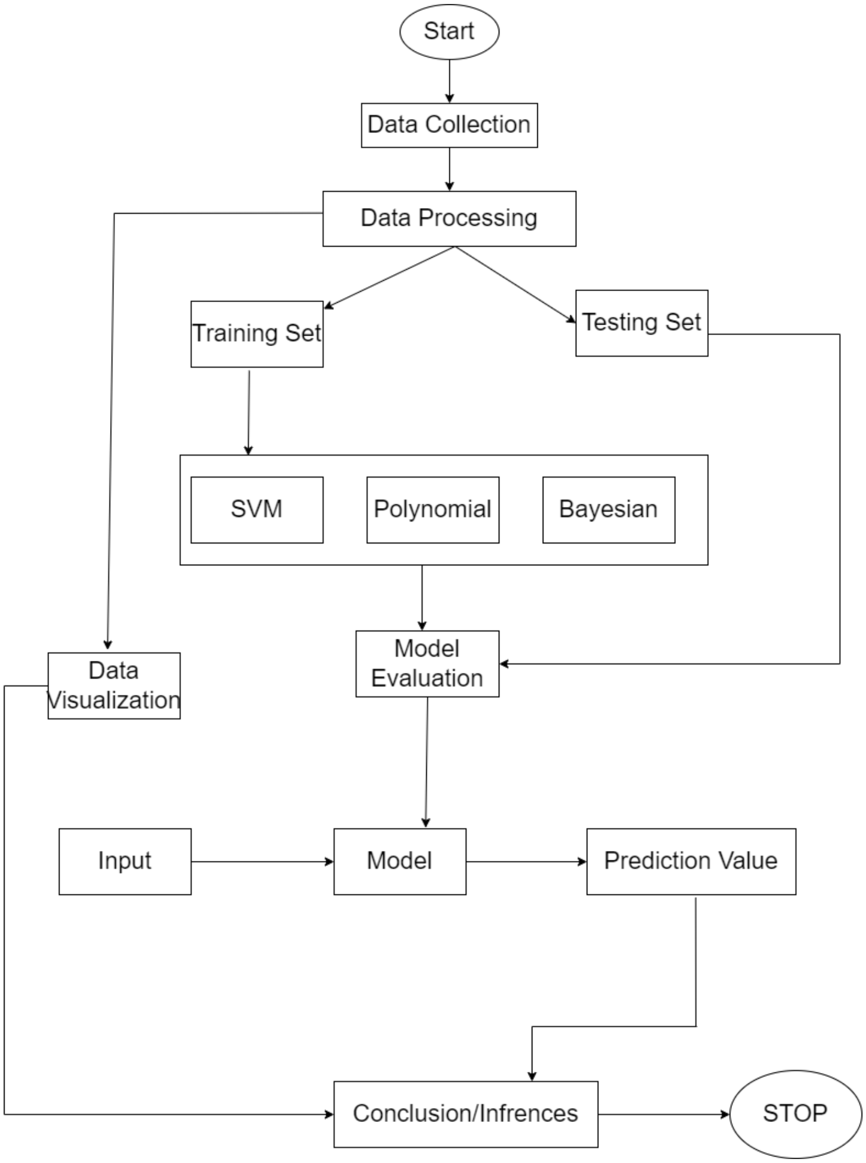

Figure 1: Flow Diagram

3.3 Exploratory Data Analysis

From the given bar graph(Figure 2) we can see that the total cases in the USA are significantly higher than in other countries, being more than double second-placed India. India, Brazil, France, Germany, UK, and Italy all have also a high number of confirmed cases. For the rest of the countries, the number of confirmed cases is close to each other.

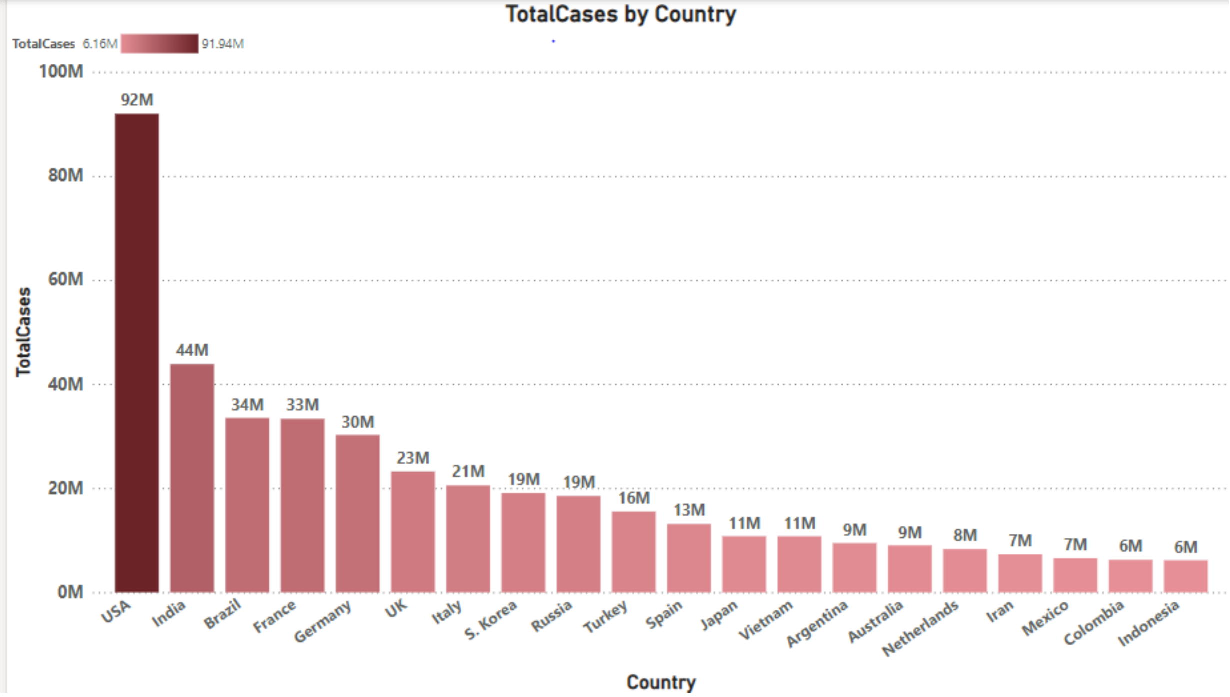

Figure 2: Total Cases by top 20 Countries

Similar to the total number of confirmed cases, the US tops the chart(Figure 3) in the total deaths. Unlike the confirmed cases, India has a lesser number of deaths compared to Brazil even though the confirmed cases were higher in India. Russia, Mexico and Peru also place in a higher position in deaths compared to their compared cases.

This graph(Figure 4) is very similar to confirmed cases. The Us leads, followed by India, Brazil, France, Germany, UK, Italy, South Korea, Russia, Turkey and Spain. We can see

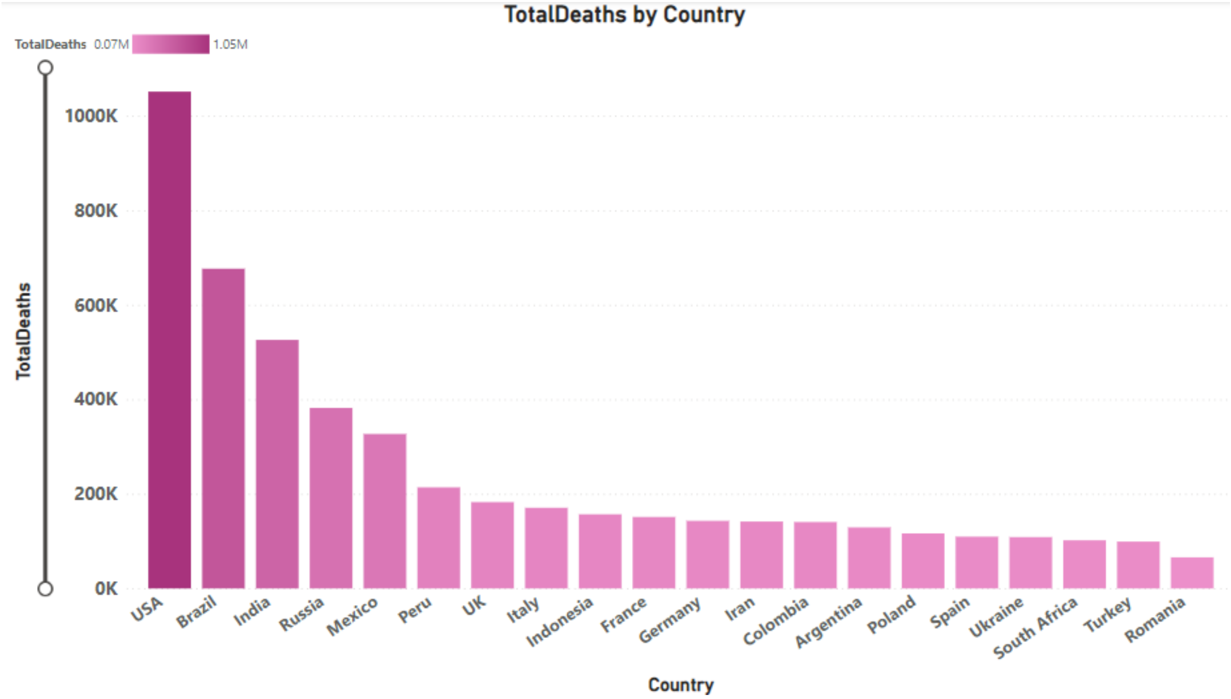

Figure 3: Total Deaths by top 20 Countries

that even though the confirmed cases are very high, the recovery rate is also very high. This shows that the disease is not that fatal. With proper healthcare, it can be tackled well.

This chart(Figure 5) shows that the countries are conducting a lot of tests to find out covid cases and quarantine those affected. Massive amounts of tests are conducted in US, India, UK, Spain, Russia, France, Italy, Austria, UAE etc. This is very important as the tests help us in finding the infected and lets us prevent further spreading of the disease.

This chart(Figure 6) is the pie-chart representation of the total confirmed cases, total deaths, total recovered, and total tests conducted country-wise for the top 20 countries for each metric. This chart shows us the percentage of each country was affected in the respective metric for the top 10 countries

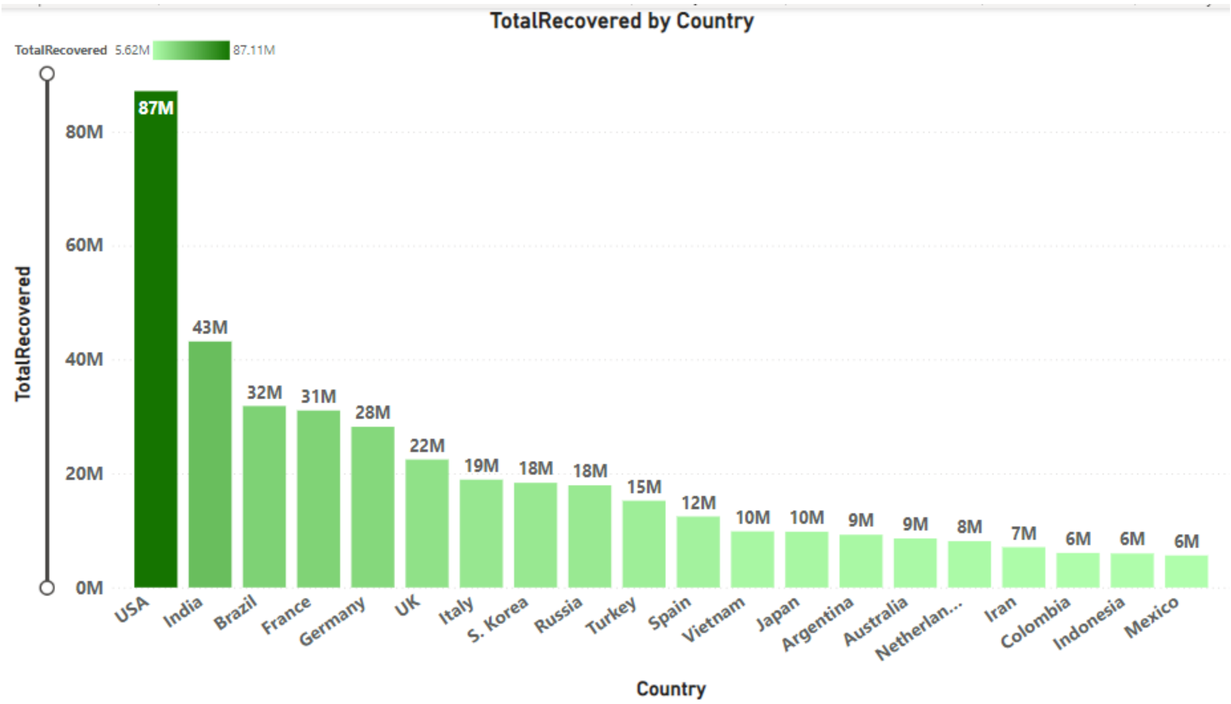

Figure 4: Total Recovered by top 20 Countries

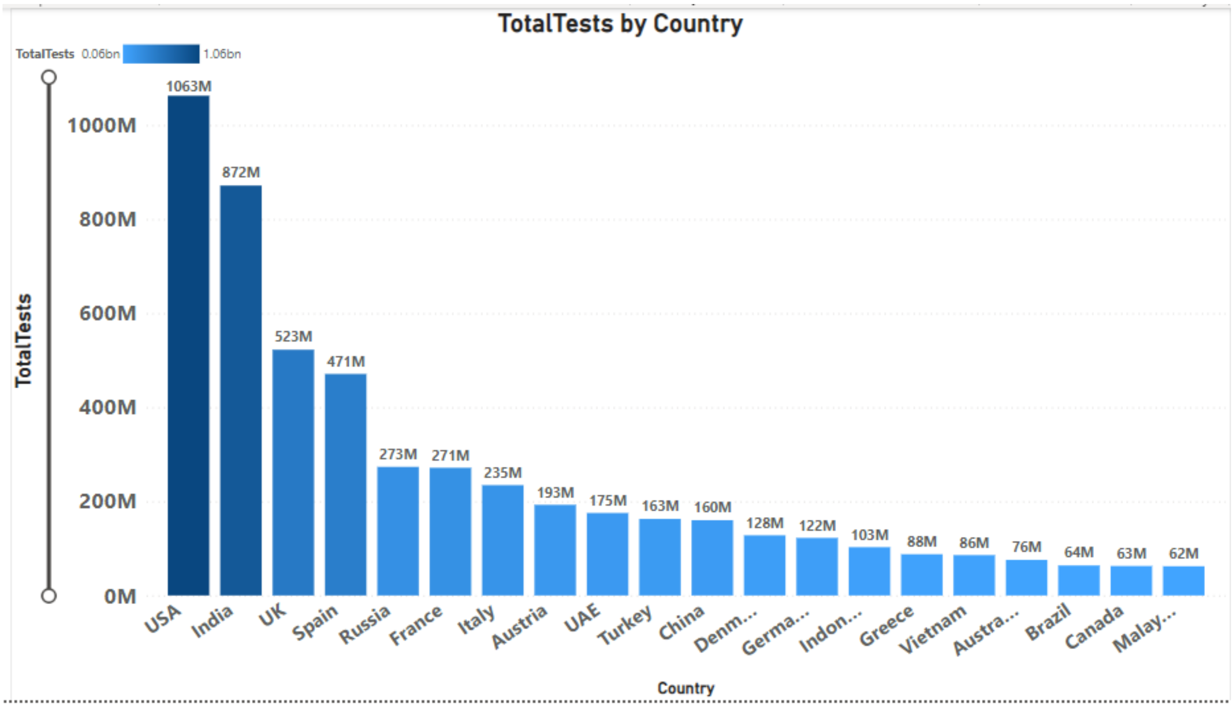

Figure 5: Total Tests by top 20 Countries

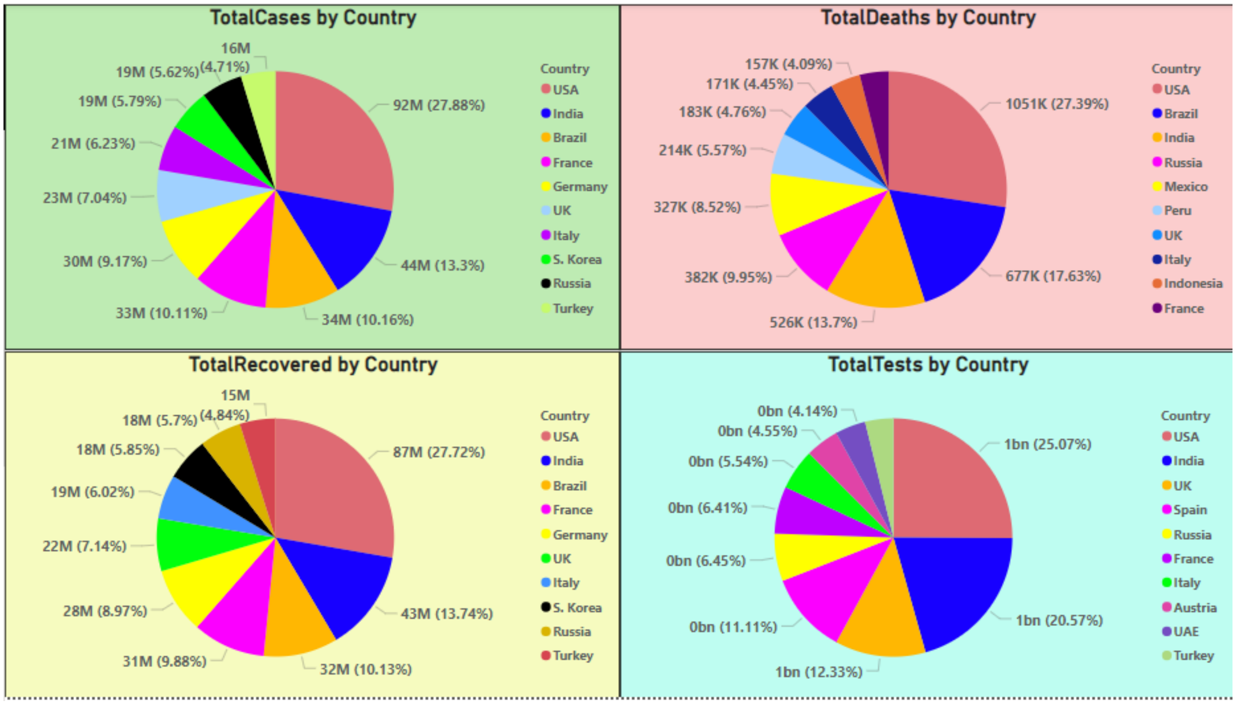

Figure 6: Pie Chart of the top! 0 Countries of Various metrics

We also have a doughnut chart(Figure 7 for the data regarding each continent. We can see that for total cases Europe leads, followed by Asia, North America, South America, Africa and Australia. For total deaths, we have Europe, North America, Asia, South America, Africa and Australia in order.

In the case of total recoveries, we have the order of Europe, Asia, North America, South America, Africa and Australia. In case of total tests conducted we have the order of Europe, Asia, North America, South America, Africa and Australia

This is the line chart(Figure 8) to show the progress of the disease in each quarter-from Jan 2020 to the present. We can see from the chart that the number of confirmed cases was steadily rising from the beginning to around Jan 2022. After Jan 2022 there was a quick rise in the number of cases for a few months. After a few months, it has gone back

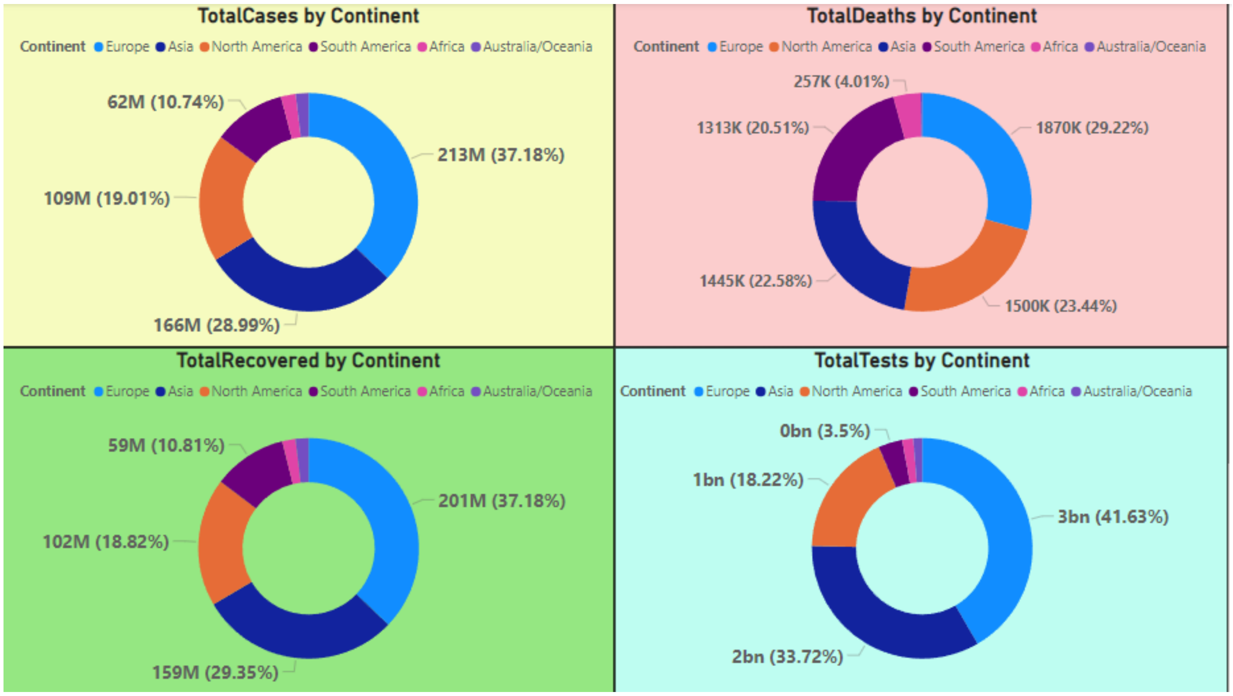

Figure 7: Donut Chart of all Continents of Various metrics

to a steady rise.

The Death rate is more predictable. It was initially rising slowly. After a few months, the slope was increasing slightly. Ever since march 2022, the slope has nearly plateaued.

This chart(Figure 9 is a stacked bar chart which shows total cases, total deaths, total recovered and total tests in the same bar. Here we can easily compare the performance of the various countries. For instance, India has conducted a lot of tests concerning the cases compared to Brazil where the number of tests conducted was less compared to cases. This shows India did well tackling the issue. Some other countries who did well are the USA, Spain, UK and Indonesia. Countries like Brazil, Germany, Argentina, Mexico and Ukraine did not do well in terms of tests done in ratio to confirmed cases

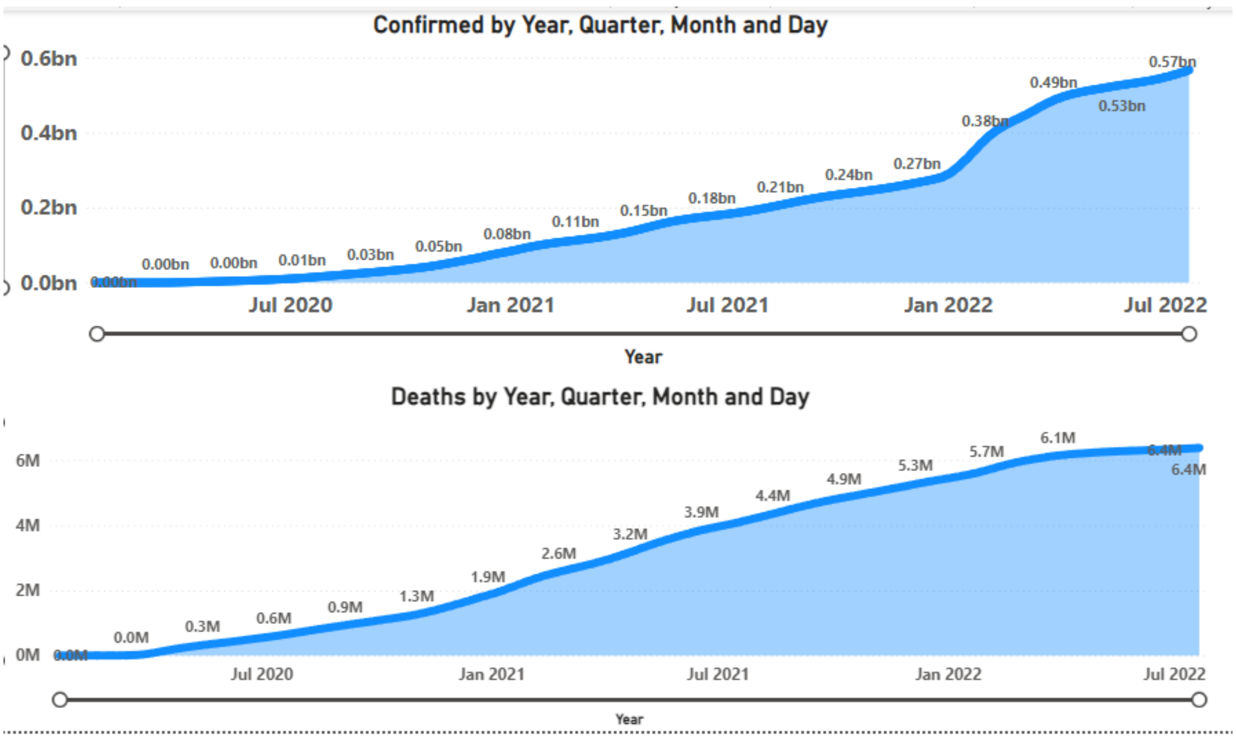

Figure 8: Quarterly Change of Cases and Deaths

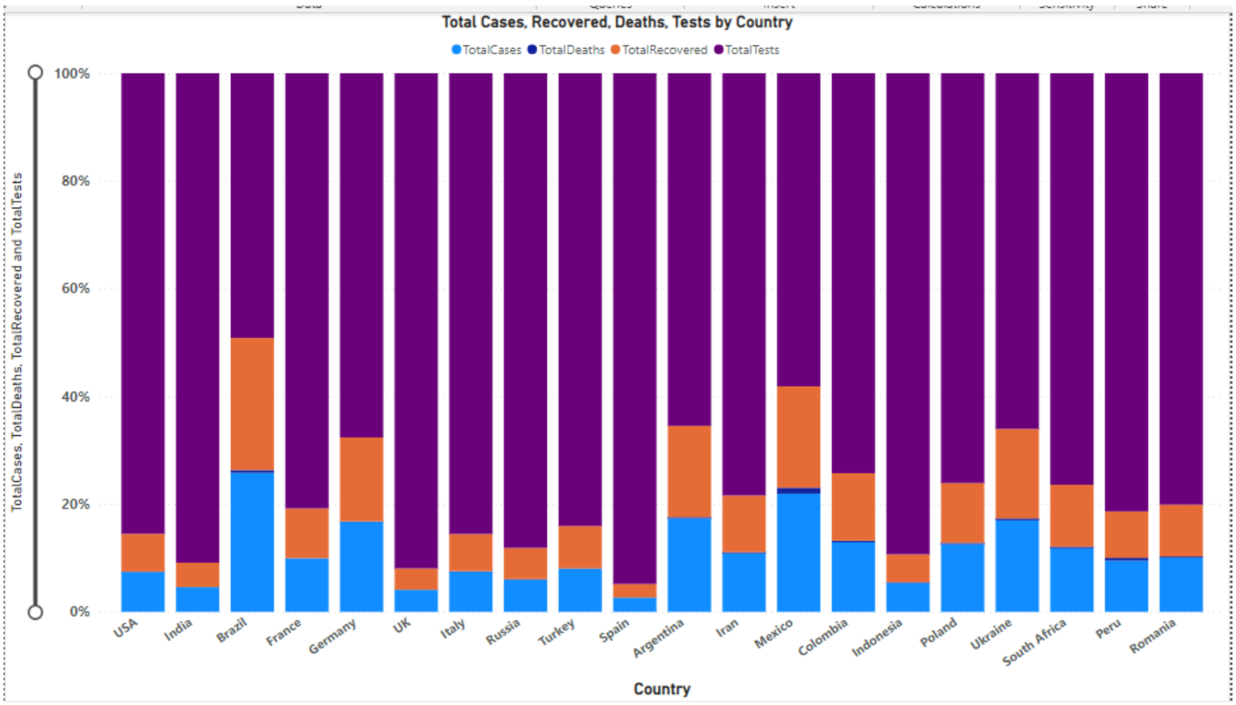

Figure 9: Stacked Chart on Cases, Deaths and Recovered

4 Results and Discussion

4.1 Results

Polynomial regression, support vector machine, and Bayesian ridge regression machine algorithms were used to predict the future cases and deaths of covid-19 patients. Among the given algorithms SVM algorithm was the best performing one as it provided the highest R2 score.

```
from sklearn.svm import SVR
svm_model = SVR()
svm_model.fit(x_train, y_train)
y_prediction = svm_model.predict(x_test)
R2 = r2_score(y_test, y_prediction)
print(f”R2 = {R2}”)
R2 Score : 0.98
```

5 GUI

GUI is made using the Flask framework. Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework-related tools.

6 Future Work And Conclusion

6.1 Future Work

The emergence of COVID-19 has been a heavy burden. During the past few months, however, significant progress has been made in the diagnosis and treatment of this disease. In this study, machine learning was used to predict and analyze the development trend and growth range of COVID-19. The results using data from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) showed that the SVM can better predict the development trend of cumulative deaths and cumulative cured cases, whereas the prediction error of cumulative confirmed cases is relatively large. In comparison, the Bayesian model has a worse predictive effect on cumulative confirmed cases, deaths and cured cases owing to the aperiodic data. Moreover, the prediction results of the two models used in this study are unstable because models tend to fall into a local optimum for data with irregular growth. For the prediction of the growth range of confirmed new cases, new deaths, and new cured cases, the SVM was shown to be effective. However, the predicted average values are larger than the actual average values, which indicates that the model is less robust to data with large fluctuations and still needs to be improved. In future research, we will improve the models based on the aforementioned problems and continue to improve the generalizability of the models.

6.2 Conclusions

The study shows that there is high recovery rate than infected and death rates. Taking the present situation as a new normal way of living they are showing negligence toward the COVID-19 Virus. Likewise, this project helped every one of us to be familiar with our home districts, the situation of the COVID pandemic, and different software. Moreover, it taught us a lot about the essence of planning, and cooperation during the team works. This type of project can be the best way of collecting, analyzing and visualizing the COVID-19 cases and their impact on different aspects of the community.

These are some of the inferences and conclusions we obtained.

- Covid has affected all countries regardless of it being a developed, developing or underdeveloped country

- The US and Brazil were the most affected countries in terms of confirmed cases and deaths. Moreover, the test done ratio of Brazil was low compared to many other countries.

- India has done a good job in tackling covid. Even though a significant number of the population was infected, preventive measures were done to treat the infected. Relatively, a higher number of tests were done which lowered the deaths.

- Small countries like Italy, Germany, and France were also affected severely even though their population was low.

- It was also noted that there was a spike in the number of covid cases in Jan 2022. It was because there was a relaxation of covid safety protocols during that time in many parts of the world. The mutation of the virus into another variant also contributed to this.