

Prediction of COVID-19 Outbreak using Machine Learning techniques

INTERDISCIPLINARY PROJECT (S23AIPROJ)

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree in biotechnology

Submitted by

S.A ABDUR RAHMAN(Reg. No. 39230050)



SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY (DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600 119

MAY - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOTECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Syed Ali Abdur Rahman who have done the Project work as a Individual who carried out the project entitled **"PREDICTION OF COVID-19 OUTBREAK USING MACHINE LEARNING TECHNIQUES"** under my/our supervision from _____ to _____

Internal Guide

Dr.RAMESH KUMAR V

External Guide

Dr.P.AJITHA

**Head of the Department
(Dr.RAMESH KUMAR V)**

Submitted for Viva voce Examination held on 29/04/2022

Internal Examiner

Dr.RAMESH KUMAR V

External Examiner

Dr.P.AJITHA

DECLARATION

I **SYED ALI ABDU RAHMAN** hereby declare that the Project Report entitled
“**PREDICTION OF COVID-19 OUTBREAK USING MACHINE LEARNING TECHNIQUES**”
done by me under the guidance of **Dr.RAMESH KUMAR V** and **Dr. P. AJITHA** at
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY is submitted in
partial fulfillment of the requirements for the award of Bachelor of Engineering /
Technology degree in **B.TECH BIOTECHNOLOGY**

DATE: 29-04-2022

PLACE: CHENNAI

A handwritten signature in black ink, appearing to read 'Syed Ali Abdu Rahman', written in a cursive style.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. -----**, Dean, School of **BIO AND CHEMICAL ENGINEERING** and **Dr.RAMESH KUMAR V**, Head of the Department, Dept. of **BIOTECHNOLOGY** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guides **Dr.RAMESH KUMAR V** and **Dr. P. AJITHA** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **BIOTECHNOLOGY** who were helpful in many ways for the completion of the project.

ABSTRACT

Machine learning (ML)-based prediction systems have proven to be useful in predicting acceptable outcomes and improving decision-making, based on the long-term sequence of events. ML models are used in a variety of applications and application domains that need to be identified and prioritized for unfavorable circumstances of a threat. The Novel Coronavirus epidemic, or the COVID-19 has had a global impact in diverse parts of the world, causing a large number of deaths.

This is still a huge public health warning, and it ranks among the best pandemics in history. To improve our forecasts, we must identify the variables that may have an association with COVID-19 spread.

This project aims to provide a better understanding of how different Machine Learning models are used in real-world situations. This study focuses on developing and comparing Machine Learning models for predicting COVID-19, and analyzing the current trend or pattern of Covid-19 transmission in India.

LIST OF FIGURES :

Figure 4.1.1. 7

Figure 4.1.2. 7

Figure 4.1.3. 8

Figure 4.1.4 8

Figure 4.3.1. 9

Figure 4.4.1.10

Figure 4.4.2.11

Table 4.4.3. 11

Table 4.4.4. 12

Figure 4.4.5.12

Figure 5.2.114

Figure 5.2.214

Table 5.2.3. 15

Table 5.2.4. 15

Table 5.2.5. 16

Table 5.2.6. 16

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES AND TABLES	iv
CHAPTER 1.	1
Introduction	1
CHAPTER 2.	4
Literature Survey	4
CHAPTER 3.	6
3.1 Aim of the Project.	6
3.2 Scope of the Project.	6
CHAPTER 4.	7
4.1 Collection and Importing of Data.	6
4.2 Description of Data.	9
4.3 Pre-Processing and Modification of Data.	9
4.4 Exploratory Data Analysis	10
4.5 Building the Algorithm.	13
CHAPTER 5	14
Results and Discussion	14
CHAPTER 6	17
Conclusions	17
REFERENCES	18

CHAPTER 1

INTRODUCTION

1. Coronaviruses

Coronaviruses are a large family of viruses known to cause illnesses ranging from common colds to more serious illnesses such as respiratory illnesses in the Middle East Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) . These two The disease is transmitted by the coronavirus known as MERSCoV and SARSCoV. SARS was first observed in China in 2002 and MERS was first observed in Saudi Arabia in 2012. Arabia . The latest virus observed in Wuhan, China is called SARSCOV2, Causes the coronavirus. Unexplained pneumonia found in Wuhan, China was reported for the first time December 31, 2019, World Health Organization (WHO) Country Office in China . Since then, the number of cases of coronavirus has increased. Number of dead. The coronavirus spread from one city to the whole country in just 30 days. It was designated as COVID 19 by the World Health Organization (WHO) on February 11 . Since this COVID 19 spreads from person to person, electronic devices equipped with artificial intelligence can play an important role in preventing the spread of this virus. when The role of epidemiologists in healthcare has expanded and the proliferation of electronics Health data is also increasing . Increased availability of eHealth services Data represent key opportunities in healthcare, both in discovery and practice An application for improving healthcare . This data can be used for training machines Learning algorithms to improve decisions related to disease prediction. As of May 16, 2020, a total of 44,25,485 COVID19s have been registered. The total number of deaths is 3.02.059 . COVID 19 has spread all over the world Approximately 213 countries and regions affected . The increase in the number of cases of infected coronavirus quickly outweighed the medical resources available in the hospital. As a result, the medical system was heavily burdened . Because it is limited Resource availability and medical outcome time lag in hospitals Test, this is a typical situation where healthcare professionals need to provide appropriate treatment patient. Due to the rapid increase in the number of coronavirus tests Routine testing is not possible due to time and cost.

2. Machine learning

Machine Learning is a subset of Artificial Intelligence(AI) and was evolved from pattern recognition where the data can be structured for the understanding of the users. Recently, many applications have been developed using Machine Learning in various fields such as healthcare, banking, military equipment, space etc. Currently, Machine Learning is a rapidly evolving and continuously developing field. It programs computers using data to optimize their performance. It learns the parameters to optimize the computer programs using the training data or its past experiences. Using the data, it can also predict the future. Machine Learning also helps us in building a mathematical model using the statistics of the data. The main objective of Machine Learning is that it learns from the feed data without any interference of humans that is, it automatically learns from given data(experience) and gives us the desired output where it searches the trends/patterns in the data. It is broadly classified into four types:

- Supervised Machine Learning.
- Unsupervised Machine Learning.
- Semi-Supervised Machine Learning.
- Reinforcement Machine Learning.

2.1 Supervised Machine Learning

Supervised Learning is a Machine Learning model that is built to give out predictions. This algorithm is performed by taking a labelled set of data as input and also known responses as output to learn the regression/classification model. It develops predictive models from classification algorithms and regression techniques.

Classification predicts discrete responses. Here, the algorithm labels by choosing two or more classes for each example. If it is done between two classes then it is called binary classification and if it is done between two or more classes then it is called multi- class classification. Applications of classification includes hand writing recognition, medical imaging etc.

Regression predicts continuous responses. Here, the algorithms returns a statistical value. For example, a set of data is collected such that the people are happy when considered the amount of sleep. Here, sleep and happy are both variables. Now, the analysis is done by making predictions. The types of popular regression techniques are:

- Linear regression.
- Logical regression.

2.2 Unsupervised Machine Learning

Unlike the supervised learning, there is no supervisor here and we only have input data. Here, the basic aim is to find certain patterns in the data that occur more than others. According to the statistics, it is called density estimation. One of the methods for the density estimation is called clustering. Here, the input data is formed into clusters or groupings. Here, the assumptions are made such that the clusters are discovered which will match reasonably well with a classification. This is a datadriven approach that works better when provided with sufficient data. For example, the movies in Netflix.com are suggested based on the principal of clustering of movies where several similar movies are grouped based on customer's recently watched movie list. It mostly discovers the unknown patterns in the data but most of the time these approximations are weak when compared with the supervised learning

2.3 Semi-supervised Machine Learning

The name "semi-supervised learning" comes from the fact that the data used is between supervised and unsupervised learning. Semi-supervised algorithm has the tendency to learn both from labelled and unlabelled data. Semi-supervised machine learning gives high accuracy with a minimum annotation work. Semi-supervised machine learning uses mostly unlabelled data together combined with labelled data to give better classifiers. As less annotation work is enough to give good accuracy, humans have less work to do here.

2.4 Reinforcement Machine Learning

Reinforcement learning learns its behaviour from a trial and error method in a dynamic environment. Here, the problem is solved by taking an appropriate action in a certain situation to maximize the output and to obtain the acquired results. In Reinforcement Learning, there is presentation of the input or output data. Instead, when the desired action is chosen, the agent is immediately told the reward and the next state are not considering the long terms actions. For the agent to act optimally it should have the knowledge about states, rewards, transitions and actions actively. Formally, the model consists of :

- a discrete set of environment states, S ;
- a discrete set of agent actions, A ;

Chapter – 2

Literature survey

Title	Findings
1. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms, Narinder Singh Pun, Sanjay Kumar Sonbhadra, Sonali Agarwal	This article proposes to utilize the machine learning and deep learning models with the aim for understanding its everyday exponential behaviour along with the prediction of future reachability of the COVID-2019 across the nations
2. Significant Applications of Machine Learning for COVID-19 Pandemic, Shashi Kushwaha, Shashi Bahl, Ashok Kumar Bagha, Kulwinder Singh Parmar, Mohd Javaid, Abid Haleem and Ravi Pratap Singh	This paper studied how machine learning algorithms and methods can be employed to fight the COVID-19 virus and the pandemic. It further discusses the primary machine learning methods that are helpful during the COVID-19 pandemic.
3. Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning, Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Prathamesh Daphal, Mahesh Warang & Ninad Mehendale	This article refers to the proposed prediction models used by us can help stakeholders to be prepared in advance for any sudden rise in outbreak to ensure optimal management of available resources.
4. Artificial intelligence and machine learning to fight COVID-19, Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar Patricia B. Munroe, Bina Joe and Xi Cheng	In this article large-scale data of COVID-19 patients can be integrated and analyzed by advanced machine learning algorithms to better understand the pattern of viral spread, further improve diagnostic speed and accuracy, develop novel effective therapeutic approaches, and potentially identify the most susceptible people based on personalized genetic and physiological characteristics
5. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach Gergo Pinter, Imre Felde, Amir Mosavi Pedram Ghamisi, Richard Gloaguen	This study proposes a hybrid machine learning approach to predict the COVID-19, and we exemplify its potential using data from Hungary.
6. Comparative study of machine learning methods for COVID-19 transmission	This paper presents a comparative study of machine learning methods for

forecasting Abdelkader,Dairia,FouziHarroub, AbdelhafidZeroual, Mohamad MazenHittaweb,YingSunb	COVID-19 transmission forecasting.
7. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, Shreshth Tuli,Shikhar Tuli,Rakesh Tuli, Sukhpal Singh Gill	This study applies an improved mathematical model to analyse and predict the growth of the epidemic. An ML-based improved model has been applied to predict the potential threat of COVID-19 in countries worldwide.
8. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review , Samuel Lalmuanawmaa, Jamal Hussain Lalrinfela Chhakchhuak	This paper aims to comprehensively review the role of AI and ML as one significant method in the arena of screening, predicting, forecasting, contact tracing, and drug development for SARS-CoV-2 and its related epidemic.
9. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset, L. J. Muhammad, Ebrahim A. Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty & I. A. Mohammed	This paper includes alternatives that may reduce the huge burden on not only limited healthcare systems but also the economic sector; the most promising include harnessing non-clinical techniques such as machine learning, data mining, deep learning and other artificial intelligence.
10. A Deep Learning Method to Forecast COVID-19 Outbreak, Satyabrata Dash, Sujata Chakravarty, Sachi Nandan Mohanty, Chinmaya Ranjan Pattanaik & Sarika Jain	In this paper, support vector regression (SVR) and deep neural network method have been used to develop the prediction models.

Chapter – 3

Aim and Scope of the present investigation

3.1 Aim:

The aim of this thesis is to predict the possibilities of cases in next 10 days a person has COVID-19 or not, using machine learning techniques. The prediction is performed using dataset of the countries with covid cases. The goal is to identify how many possible cases might occur in next 10 days.

3.2 Scope:

The prediction model with highest accuracy can be incorporated into a mobile application or a website, to help people to know how many cases are being increased. The model can also be further enhanced to use the patient's current data to predict new cases for next 10 days

CHAPTER 4

METHODOLOGY AND ALGORITHMS USED

4.1 Collection of data and Importing

The dataset used for this project was directly obtained from Kaggle, an open source repository. This are the following datasets are obtained:

time_series_covid19_confirmed - Excel

Syed Ali

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Cut

Copy

Format Painter

Clipboard

Calibri

11

Fig.4.1.1 Dataset of confirmed cases from 22 Jan 2020

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	02-01-2020	02-02-2020	02-03-2020	02-04-2020	02-05-2020	02-06-2020	02-07-2020
Thailand	Thailand	15	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Japan	Japan	36	138	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Singapore	Singapore	1.2833	103.8333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nepal	Nepal	28.1667	84.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Malaysia	Malaysia	2.5	112.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
British Columbia	Canada	49.2827	-123.121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
New South Wales	Australia	-33.8688	151.2093	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Victoria	Australia	-37.8136	144.9631	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cambodia	Cambodia	11.55	104.9167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sri Lanka	Sri Lanka	7	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Germany	Germany	51	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Finland	Finland	64	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
United Arab Emirates	United Arab Emirates	24	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Philippines	Philippines	13	122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
India	India	21	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Italy	Italy	43	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sweden	Sweden	63	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Spain	Spain	40	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
South Australia	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Belgium	Belgium	50.8333	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Egypt	Egypt	26	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
From Diamond P	Australia	35.4437	139.638	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lebanon	Lebanon	33.8547	35.8623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iraq	Iraq	33	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Oman	Oman	21	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig.4.1.2 Dataset of deaths from 22 Jan 2020

Province/State	Country/R	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	02-01-2020	02-02-2020	02-03-2020	02-04-2020	02-05-2020	02-06-2020	02-07-2020
Thailand		15	101	0	0	0	0	0	2	2	5	5	5	5	5	5	5	5	5	5
Japan		36	138	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Singapore		1.2833	103.8333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nepal		28.1867	84.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Malaysia		2.5	112.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
British Columbia	Canada	49.2827	-123.121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
New South Wales	Australia	-33.8688	151.2093	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2
Victoria	Australia	-37.8136	144.9631	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cambodia		11.55	104.9167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sri Lanka		7	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Germany		51	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Finland		64	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
United Ara		24	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Philippines		13	122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
India		21	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Italy		43	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sweden		63	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Spain		40	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
South Australia	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Belgium		50.8333	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Egypt		26	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
From Diamond Princes	Australia	35.4437	139.638	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lebanon		33.8547	35.8623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iraq		33	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Oman		21	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig.4.1.3 Dataset of recovered cases from 22 Jan 2020

The dataset was imported into Google Colab as a csv file. The following important libraries were imported in Google Colab :

```

# Importing all the important libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import random
import math
import time
from sklearn.model_selection import RandomizedSearchCV, train_test_split
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, mean_absolute_error
import datetime
import operator
plt.style.use('seaborn')
%matplotlib inline

[ ] # loading all the three datasets:

confirmed_cases = pd.read_csv('/content/sample_data/time_series_covid-19_confirmed.csv')

[ ] deaths_reported = pd.read_csv('/content/sample_data/time_series_covid-19_deaths.csv')

[ ] recovered_cases = pd.read_csv('/content/sample_data/time_series_covid-19_recovered.csv')

```

Fig 4.1.4 Screenshot of Google Colab where libraries is being imported

- **import numpy as np**- The as np portion of the code then tells Python to give NumPy the alias of np. This allows you to use NumPy functions by simply typing np.
- **import pandas as pd**- It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a NumPy matrix array.

- **import matplotlib.pyplot as plt-** is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- **import matplotlib.colors as mcolors-** This function is used to specify the color.
- **from sklearn.model_selection import RandomizedSearchCV, train_test_split-** implements a “fit” and a “score” method. It also implements “score_samples”, “predict”, “predict_proba”, “decision_function”, “transform” and “inverse_transform” if they are implemented in the estimator used.

4.2 DESCRIPTION OF DATA

There are 3 datasets being used here:

1. Number of Confirmed cases
2. Number of Deaths
3. Number of Recovered cases

All this three data sets includes:

- a) Province/State
- b) Country/Region
Both of them includes the places around the word
- c) Dates starting from 22-Jan-2020

4.3 PRE-PROCESSING AND MODIFICATION OF DATA

The data variables were modified to make statistical computations on them, easier and more reliable

```
[ ] #converting all the integers to datetime for better visualization

start='1/22/2020'
start_date = datetime.datetime.strptime(start, '%m/%d/%Y')
future_forecast_dates = []
for i in range(len(future_forecast)):
    future_forecast_dates.append((start_date + datetime.timedelta(days=i)).strftime('%m/%d/%Y'))

# For visualization with the latest data of 15th of march

latest_confirmed = confirmed_cases[dates[-1]]
latest_deaths = deaths_reported[dates[-1]]
latest_recoveries = recovered_cases[dates[-1]]

[ ] latest_confirmed

0      114
1      839
2      226
3         1
4      428
...
445     1
446     1
447     1
448     1
449     1
Name: 3/15/20, Length: 450, dtype: int64
```

Fig 4.3.1 Modifying Dataset

Here conversion of all integers into datetime for better visualization is taking place along with the latest data of 15th of march by which latest confirmed cases, deaths, recoveries can be known.

4.4 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. So let us see how the data has been explored:

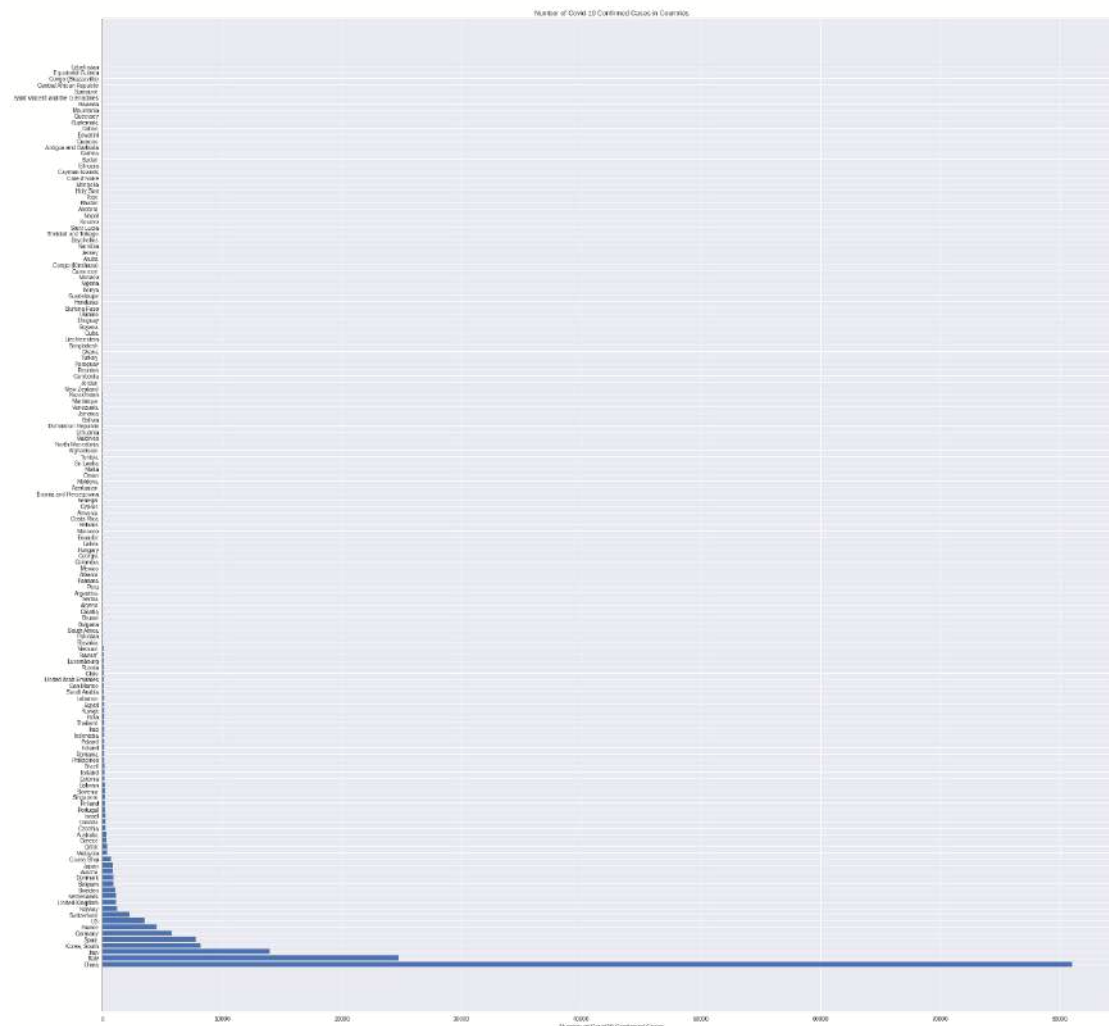


Fig 4.4.1 Plotting a bar graph to see the total confirmed cases across different countries

As the horizontal bar graph clearly shows that the number of cases in China is increasing very high as compared to the other countries and is said to be having high deaths rates in comparison to other regions.

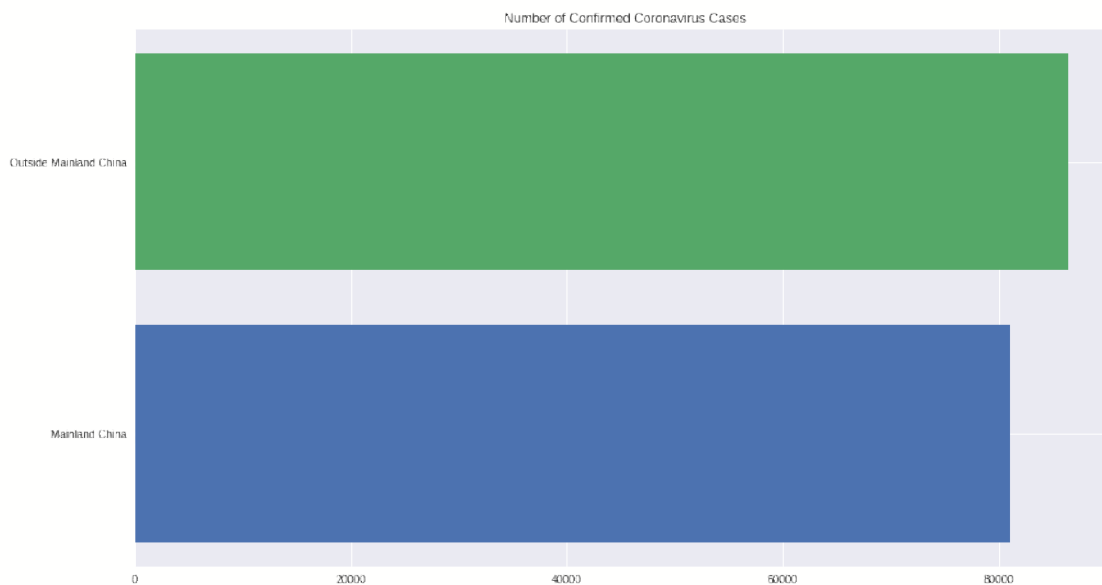


Fig 4.4.2 shows the confirmed corona cases outside Mainland China vs Mainland China

This graph shows the confirmed cases are to be more in outside mainland China than in mainland.

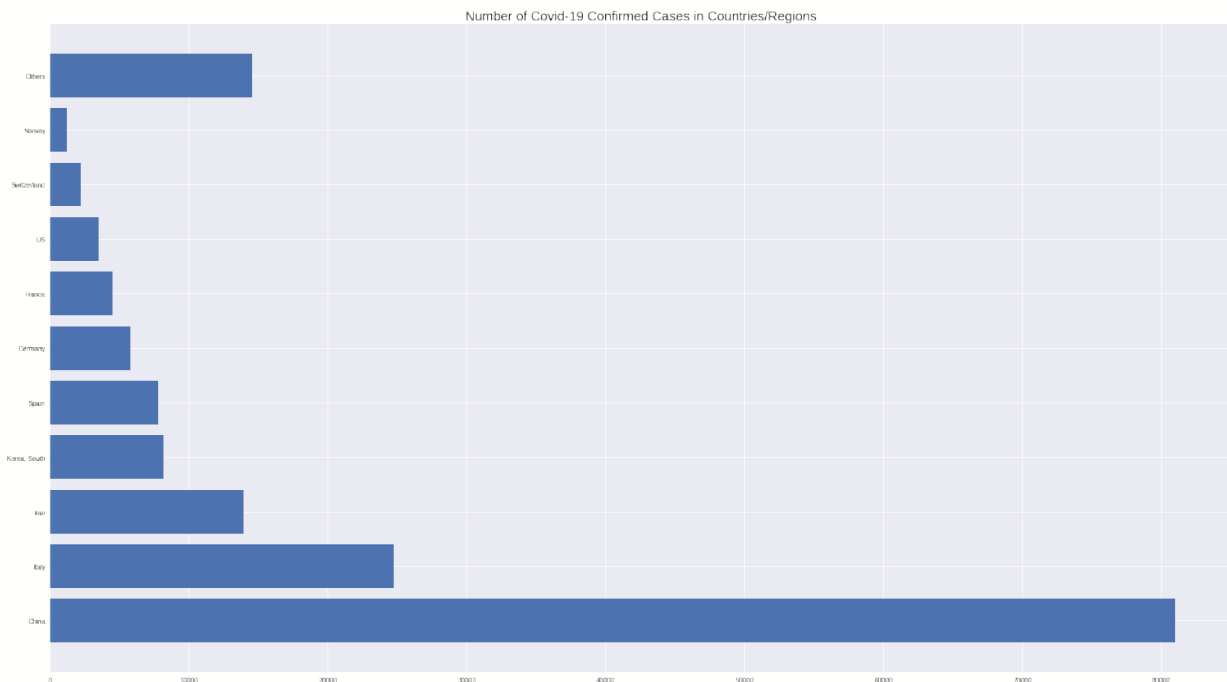


Fig 4.4.3 shows the confirmed cases in unique countries

This graph depicts the cases for unique countries where the China is the highest corona cases country then others

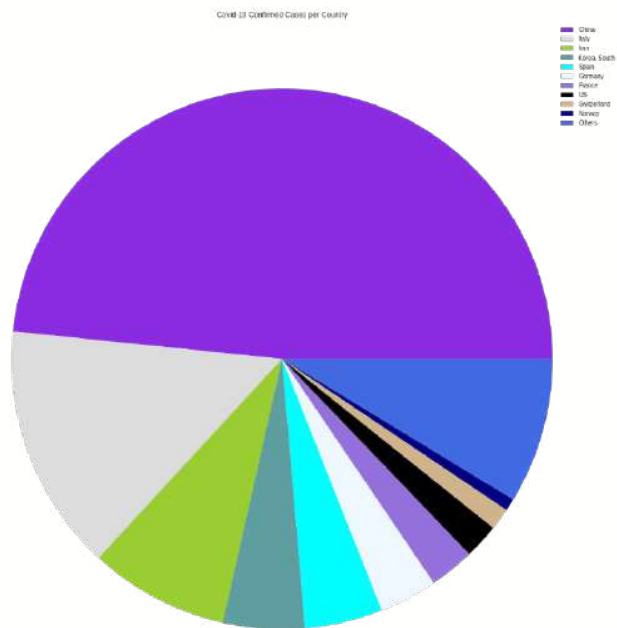


Fig:4.4.4 Pie chart shows the no of Covid-19 Cases per country

As we can see that china(violet) shows the highest composition

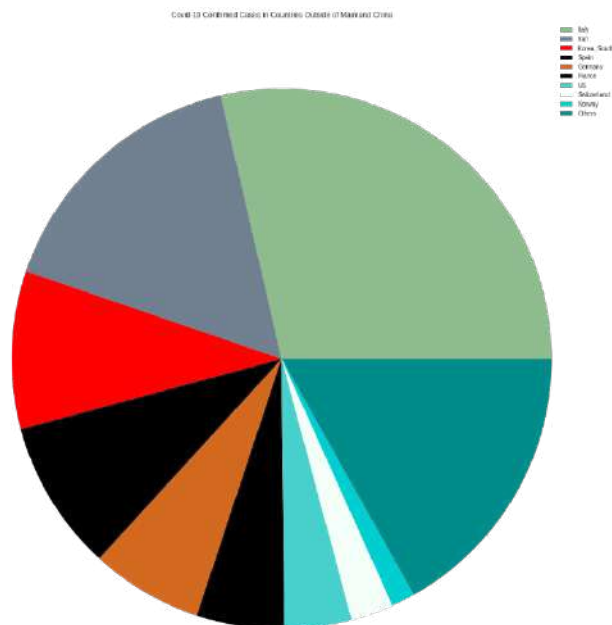


Fig 4.4.5 This pie chart shows the no. of cases out of mainland china

As taking the survey outside mainland China, Italy suppose to have Second highest cases after Mainland China

4.5 BUILDING THE ALGORITHM

One of the preliminary steps of model fitting is to first split the data into test and train sets. A vector “y”, holding the binary outcomes under world cases. The target variable Cases before march 15 was removed from the dataset, so it finally contained only the predictor variables. The dataset was then scaled, in order to maintain uniform variance across all predictor variables. The scaled dataset was split into test and train sets. The “y” vector holding the world cases variable’s values was also split into test and train sets and converted into factors for easy classification.

4.5 MODEL USED LINEAR REGRESSION

```
X_train_confirmed,X_test_confirmed,y_train_confirmed,y_test_confirmed = train_test_split(days_since_1_22,world_cases,test_size=0.15,shuffle=False)
```

```
linear_model = LinearRegression(normalize=True, fit_intercept=True)
```

```
linear_model.fit(X_train_confirmed, y_train_confirmed)
```

```
test_linear_pred = linear_model.predict(X_test_confirmed)
```

```
linear_pred = linear_model.predict(future_forecast)
```

CHAPTER 5

5.1 RESULTS AND DISCUSSION

The accuracy of the Regression Model is **99.7%**.

5.2 The modified Linear Regression model's performance can be visualized here:

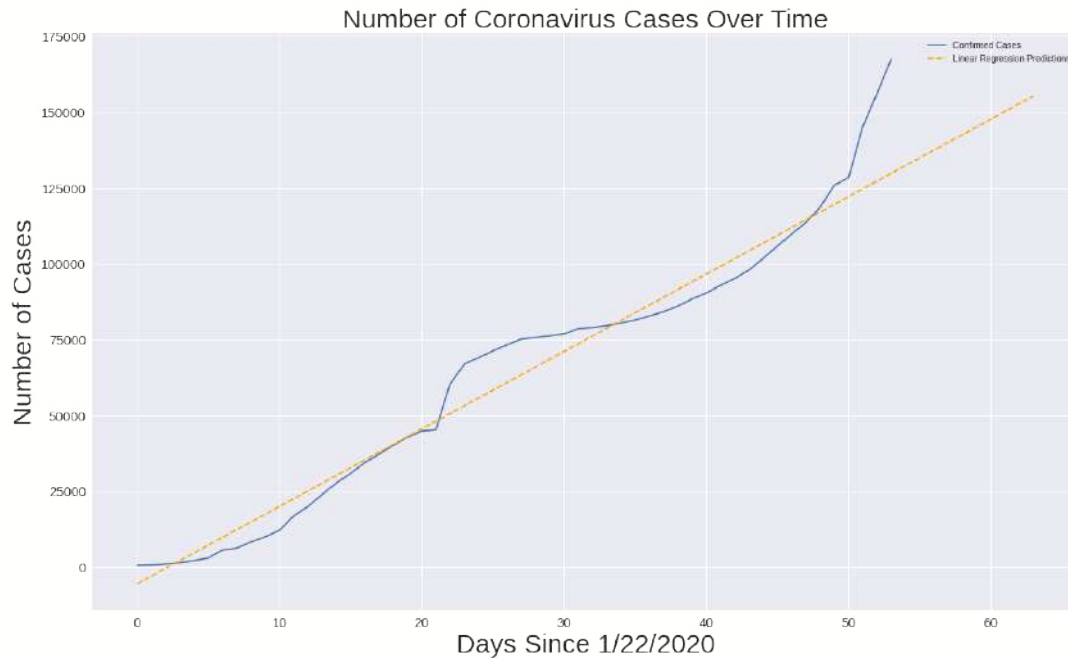


Fig 5.2.1 This graph shows us the Number of cases vs Days since 1/22/2020

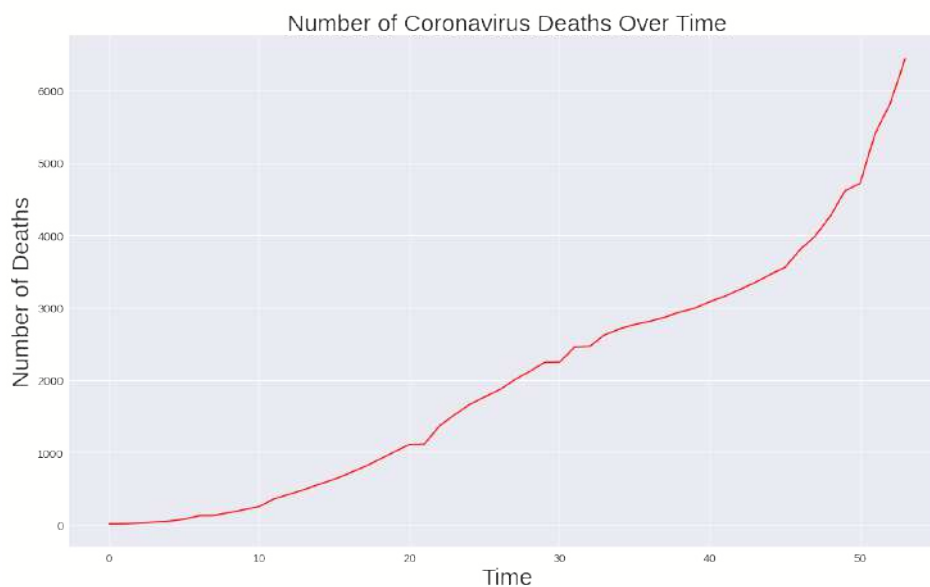


Fig 5.2.2 This graph shows us the number of deaths over period of time

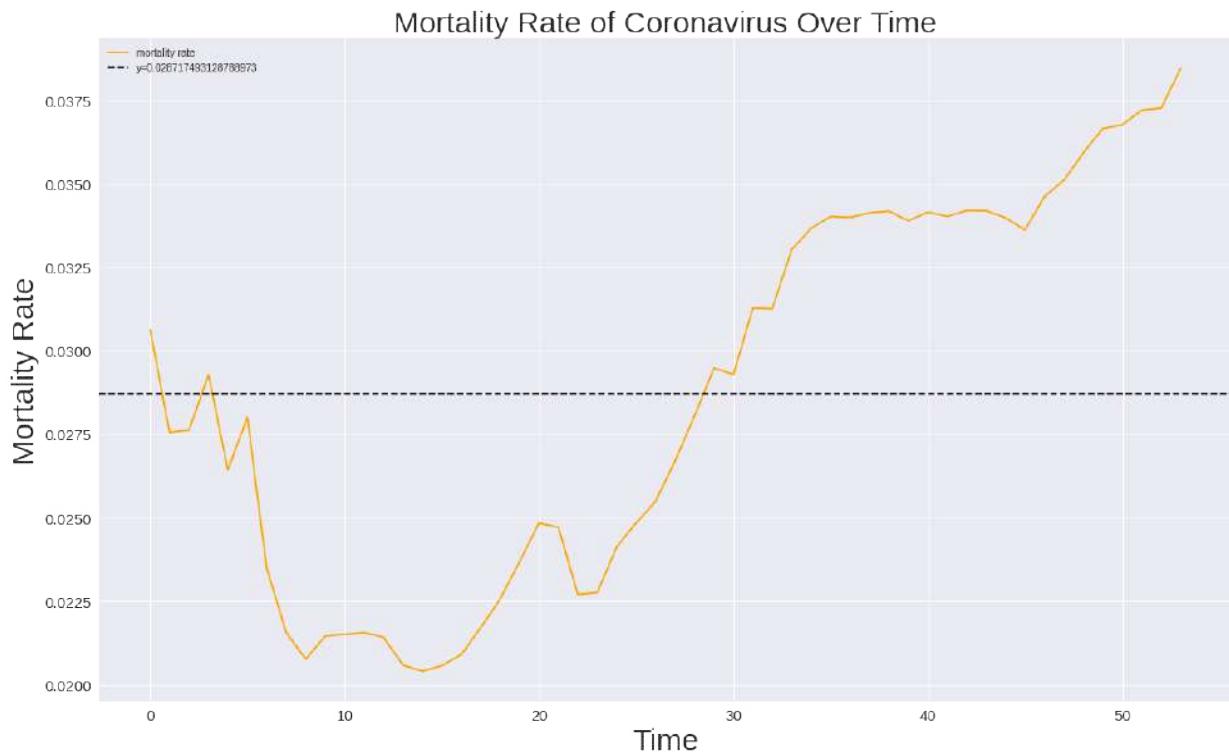


Fig 5.2.3 this graph shows us the Mortality rate over a period of time

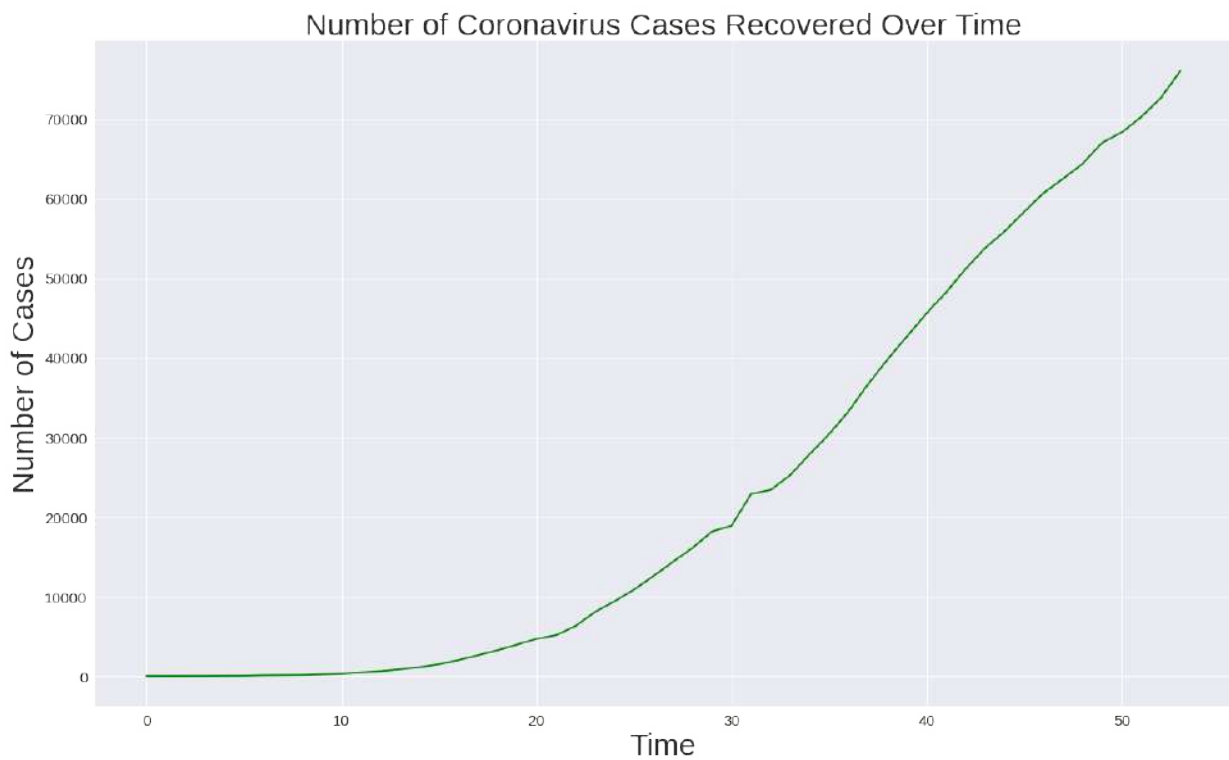


Fig 5.2.4 This graph represents recovered corona cases over a time period

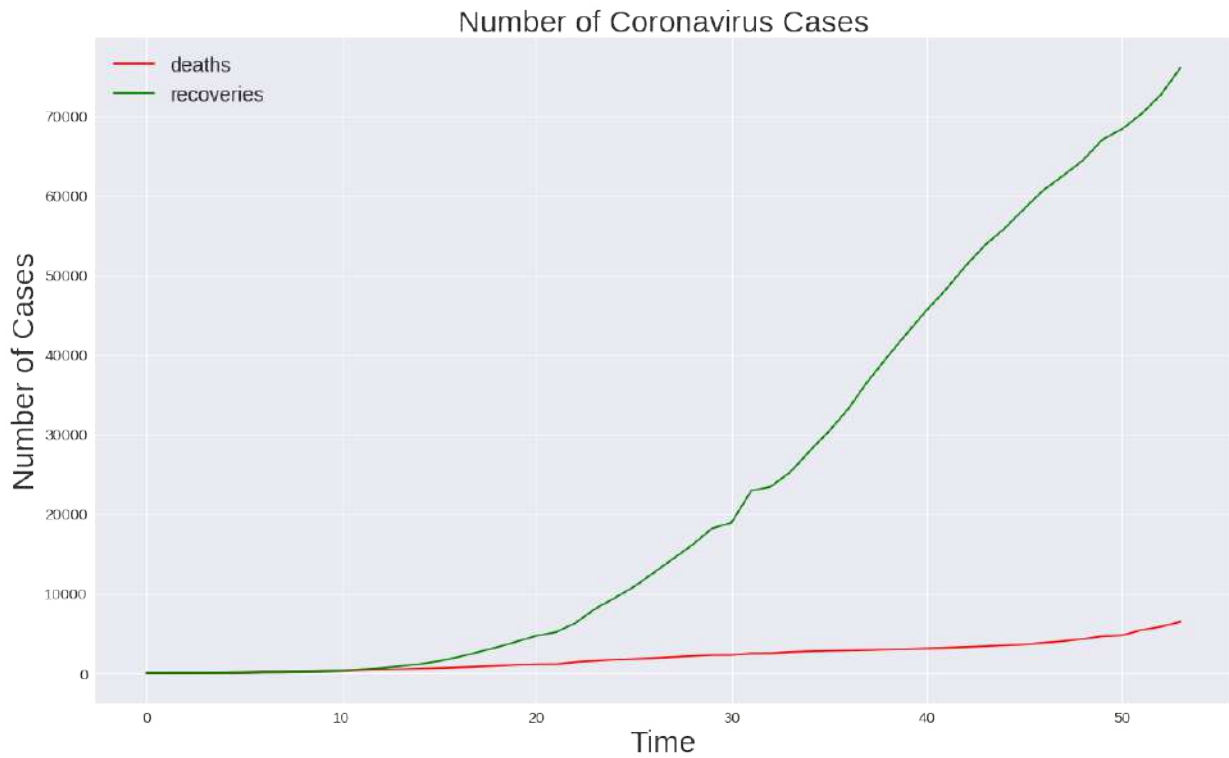


Fig 5.2.5 this shows the variation between deaths and recoveries over a period of time

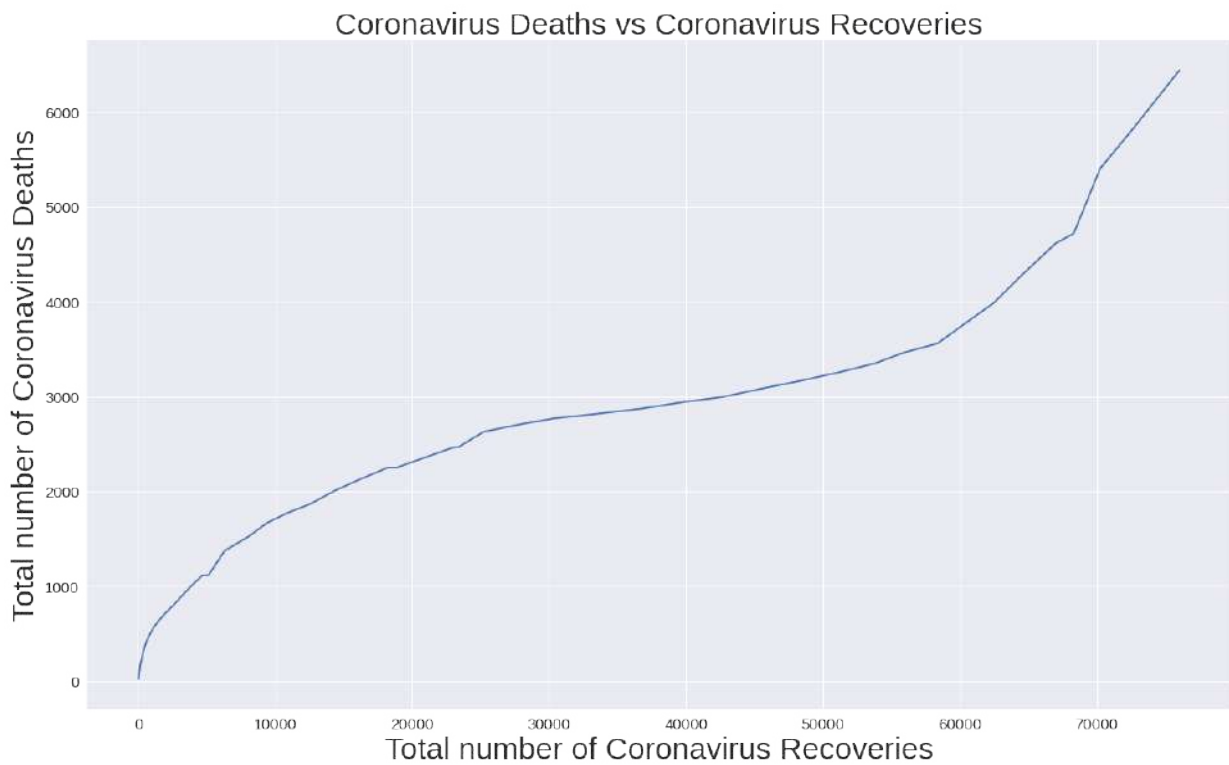


Fig 5.2.6 this shows the deaths and recoveries in single parameter

CHAPTER 6

CONCLUSIONS

In this project, a systematic literature review has been conducted to identify the suitable algorithm for prediction of COVID-19. There was accurate model found to summarize one algorithm as the suitable technique for prediction. The selected algorithms were trained with the datasets. To evaluate the accuracy of machine learning models, algorithm is trained with record sets of varying number of patients. Using accuracy performance metric, the trained algorithms were assessed. After result analysis, Linear Regression showed best prediction accuracy. The trained algorithms were also assessed to find the features that affect the prediction of COVID-19 in patients. There is a lot of scope for Machine Learning in Healthcare. For Future work, it is recommended to work on calibrated and ensemble methods that could resolve quirky problems faster with better outcomes than the existing algorithms.

REFERENCES

1. Pun, N.S., Sonbhadra, S.K. and Agarwal, S., 2020. COVID-19 epidemic analysis using machine learning and deep learning algorithms. *MedRxiv*.
2. Kushwaha, S., Bahl, S., Bagha, A.K., Parmar, K.S., Javaid, M., Haleem, A. and Singh, R.P., 2020. Significant applications of machine learning for COVID-19 pandemic. *Journal of Industrial Integration and Management*, 5(04), pp.453-479.
3. Khakharia, A., Shah, V., Jain, S., Shah, J., Tiwari, A., Daphal, P., Warang, M. and Mehendale, N., 2021. Outbreak prediction of COVID-19 for dense and populated countries using machine learning. *Annals of Data Science*, 8(1), pp.1-19.
4. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B. and Cheng, X., 2020. Artificial intelligence and machine learning to fight COVID-19. *Physiological genomics*, 52(4), pp.200-202.
5. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. and Gloaguen, R., 2020. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics*, 8(6), p.890.
6. Dairi, A., Harrou, F., Zeroual, A., Hittawe, M.M. and Sun, Y., 2021. Comparative study of machine learning methods for COVID-19 transmission forecasting. *Journal of Biomedical Informatics*, 118, p.103791.
7. Tuli, S., Tuli, S., Tuli, R. and Gill, S.S., 2020. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, 11, p.100222.
8. Lalmuanawma, S., Hussain, J. and Chhakchhuak, L., 2020. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, 139, p.110059.
9. Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C. and Mohammed, I.A., 2021. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science*, 2(1), pp.1-13.
10. Dash, S., Chakravarty, S., Mohanty, S.N., Pattanaik, C.R. and Jain, S., 2021. A deep learning method to forecast COVID-19 outbreak. *New Generation Computing*, 39(3), pp.515-539.

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY**



PREDICTION OF COVID-19 OUTBREAK USING MACHINE LEARNING TECHNIQUES”

Internal Guide

**Dr. Ramesh Kumar V
Dept of Biotechnology**

**SYED ALI ABDUR RAHMAN (39230050)
Dept of Biotechnology**

External Guide

**Dr. P Ajitha
Dept of IT**

Problem Statement:

- The Novel Coronavirus epidemic, or the COVID-19 has had a global impact in diverse parts of the world, causing a large number of deaths.
- This is still a huge public health warning, and it ranks among the best pandemics in history. To improve our forecasts, we must identify the variables that may have an association with COVID-19 spread.

Solution for the problem statement

- This project aims to provide a better understanding of how different Machine Learning models are used in real-world situations. This study focuses on developing and comparing Machine Learning models for predicting COVID-19, and analyzing the current trend or pattern of Covid-19 transmission in world.
- The prediction is performed using dataset of the countries with covid cases. The goal is to identify how many possible cases might occur in next 10 days.

1. Data collection

time series covid-19 deaths

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Conditional Formatting Styles Cell Styles Cells Editing

Province/State Country/Region Lat Long 1/22/20 1/23/20 1/24/20 1/25/20 1/26/20 1/27/20 1/28/20 1/29/20 1/30/20 1/31/20 02-01-2020 02-02-2020 02-03-2020 02-04-2020 02-05-2020 02-06-2020 02-07-2020

1	Thailand		13	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Japan		36	138	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Singapore		1.2833	103.8333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Nepal		28.1667	84.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Malaysia		2.5	112.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	British Columbia	Canada	49.2827	-123.121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	New South Wales	Australia	-33.8688	151.2093	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Victoria	Australia	-37.8136	144.9631	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Cambodia		11.55	104.9167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Sri Lanka		7	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Germany		51	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Finland		64	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	United Arab		24	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Philippines		13	122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	India		21	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Italy		43	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Sweden		63	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Spain		40	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	South Australia	Australia	-34.9285	138.0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Belgium		50.8333	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Egypt		26	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	From Diamond	Australia	35.4437	139.638	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Lebanon		33.8547	35.8623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Iraq		33	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	Oman		21	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27																				

time series covid-19 deaths

Dataset of confirmed cases from 22 Jan 2020

Fig.4.1.2 Dataset of deaths from 22 Jan 2020

time_series_covid_19_recovered - Excel

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Help

Tell me what you want to do

Out

Copy

Format Painter

Clipboard

Font

alignment

Number

General

Conditional Formatting

Format as Table

Call Styles

Insert

Delete

Format

Cells

AutoSum

Fill

Clear

Sort & Filter

Find & Select

Editing

A1

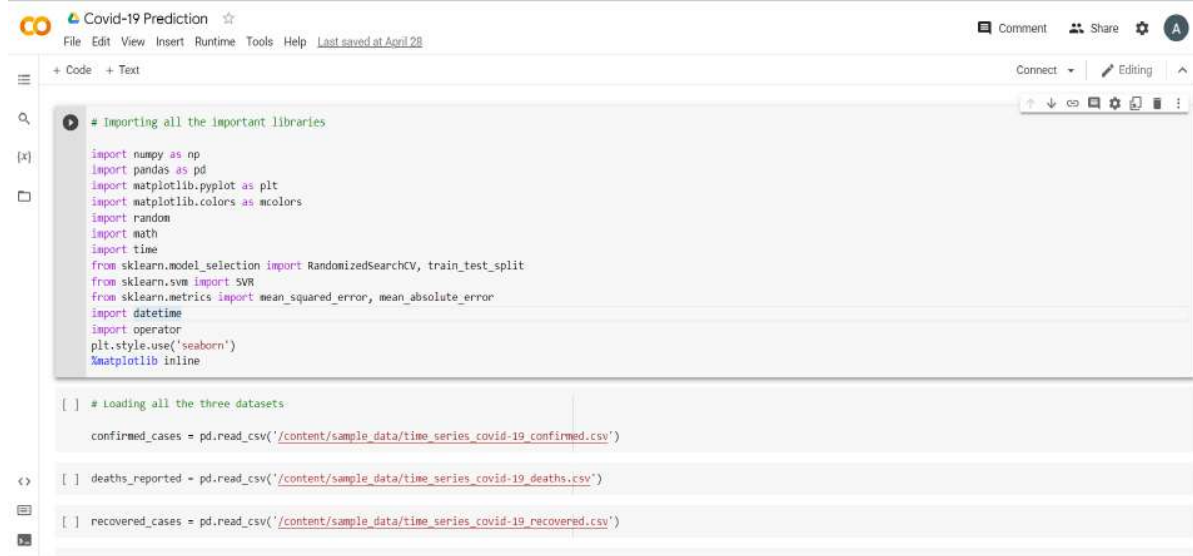
Province/State

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Province/State	Country/H Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	02-01-2020	02-02-2020	02-03-2020	02-04-2020	02-05-2020	02-06-2020	02-07-2020	
2	Thailand	55	101	0	0	0	0	0	2	2	5	5	5	5	5	5	5	5	5	5	
3	Japan	36	138	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	
4	Singapore	1.2833	103.8333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	Nepal	28.1667	84.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	Malaysia	2.5	112.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	British Columbia	Canada	49.2827	-123.121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	New South Wales	Australia	-33.8688	151.2093	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	
9	Victoria	Australia	-37.8136	144.9631	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	Cambodia	11.55	104.9167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	Sri Lanka	7	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	Germany	51	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	Finland	64	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	United Ara	24	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	Philippines	13	122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	India	21	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	Italy	43	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	Sweden	63	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	Spain	40	-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	South Australia	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	Belgium	50.8333	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	Egypt	26	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	From Diamond Princess Australia	35.4437	139.638	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	Lebanon	33.8547	35.8623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	Iraq	33	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	Oman	21	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

time_series_covid_19_recovered

Fig.4.1.3 Dataset of recovered cases from 22 Jan 2020

2. Data Processing and Modification



This screenshot shows a Google Colab notebook titled "Covid-19 Prediction". The code is in a single cell and imports several libraries: numpy, pandas, matplotlib.pyplot, matplotlib.colors, random, math, time, sklearn.model_selection, sklearn.svm, sklearn.metrics, datetime, and operator. It also sets the plot style to 'seaborn' and enables inline plotting. The code then loads three datasets: confirmed_cases, deaths_reported, and recovered_cases, each as a pandas DataFrame from a CSV file in the /content/sample_data directory.

```
# Importing all the important libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import random
import math
import time
from sklearn.model_selection import RandomizedSearchCV, train_test_split
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, mean_absolute_error
import datetime
import operator
plt.style.use('seaborn')
%matplotlib inline

[ ] # Loading all the three datasets

confirmed_cases = pd.read_csv('/content/sample_data/time_series_covid19_confirmed.csv')

[ ] deaths_reported = pd.read_csv('/content/sample_data/time_series_covid19_deaths.csv')

[ ] recovered_cases = pd.read_csv('/content/sample_data/time_series_covid19_recovered.csv')
```

Screenshot of Google Colab where libraries is being imported



This screenshot shows the same Google Colab notebook with additional code. The first cell continues with converting the start date to a datetime object and creating a list of future forecast dates. The second cell imports the latest confirmed, deaths, and recovered cases for the date 3/15/2020. The third cell displays the 'latest_confirmed' data as a table.

```
[ ] #Converting all the integers to datetime for better visualization

start='3/22/2020'
start_date = datetime.datetime.strptime(start, '%m/%d/%Y')
future_forecast_dates = []
for i in range(len(future_forecast)):
    future_forecast_dates.append((start_date + datetime.timedelta(days=i)).strftime('%m/%d/%Y'))

# For visualization with the latest data of 15th of March

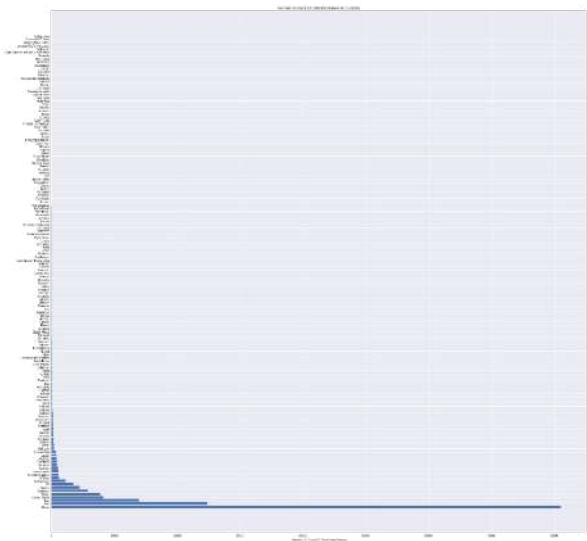
latest_confirmed = confirmed_cases[dates[-1]]
latest_deaths = deaths_reported[dates[-1]]
latest_recoveries = recovered_cases[dates[-1]]

[ ] latest_confirmed

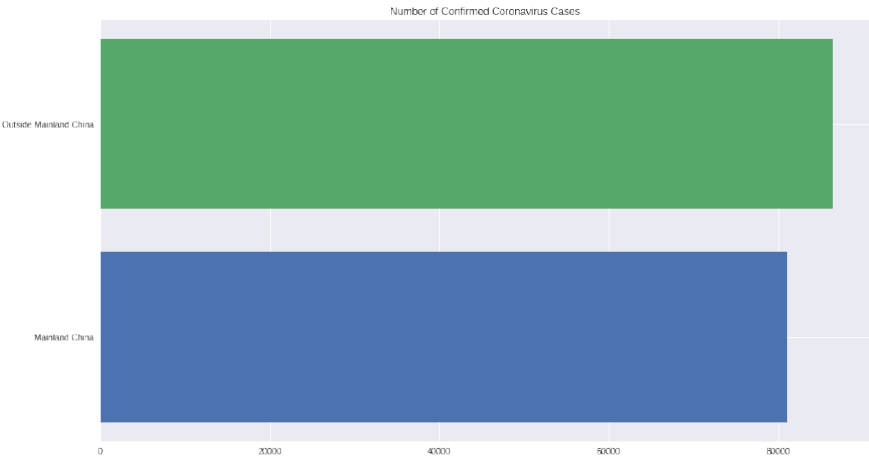
0      114
1      839
2      226
3         1
4      428
...
445      1
446      1
447      1
448      1
449      1
Name: 3/15/20, Length: 450, dtype: int64
```

Modifying Dataset

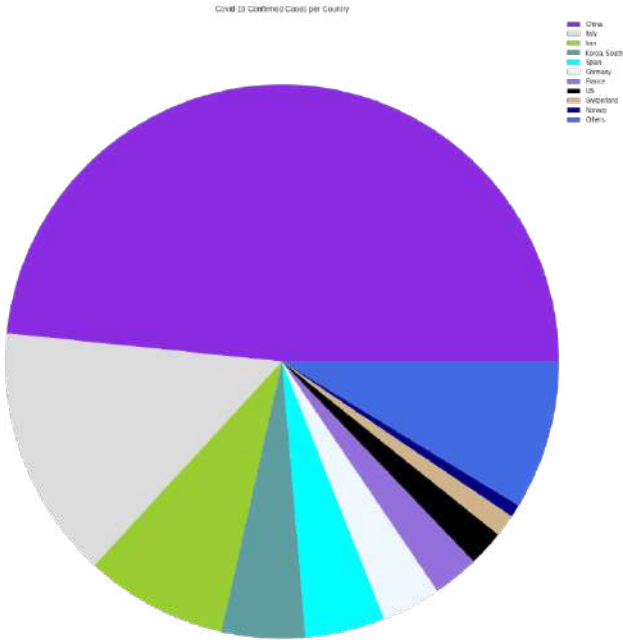
3. Exploratory Data Analysis



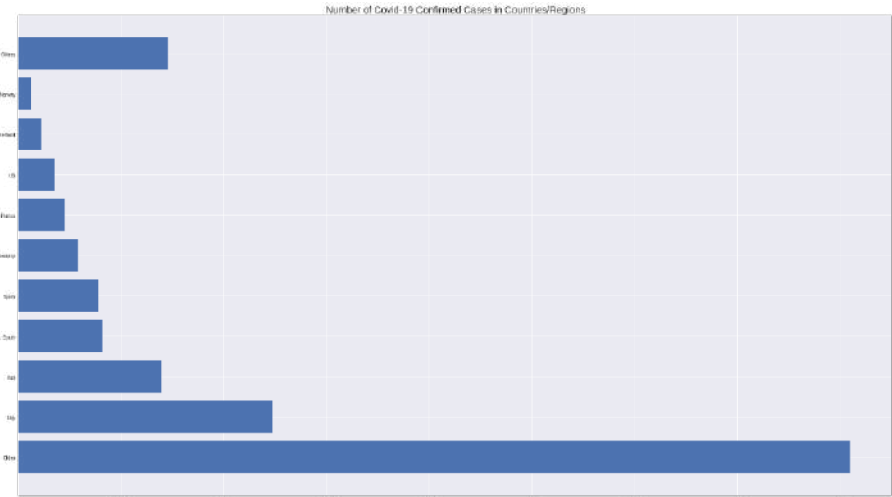
bar graph to see the total confirmed cases a cross different countries



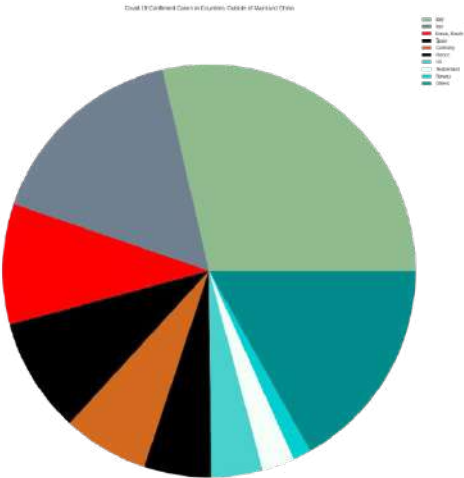
shows the confirmed corona cases outside Mainland China vs Mainland China



Pie chart shows the no of Covid-19 Cases per country



shows the confirmed cases in unique countries



pie chart shows the no. of cases out of mainland china

3. Building the Algorithm

```
X_train_confirmed,X_test_confirmed,y_train_confirmed,y_test_confirmed = train_test_split(days_since_1_22,world_cases,test_size=0.15,shuffle=False)
```

```
linear_model = LinearRegression(normalize=True, fit_intercept=True)
```

```
linear_model.fit(X_train_confirmed, y_train_confirmed)
```

```
test_linear_pred = linear_model.predict(X_test_confirmed)
```

```
linear_pred = linear_model.predict(future_forecast)
```

```
[ ] X_train_confirmed,X_test_confirmed,y_train_confirmed,y_test_confirmed = train_test_split(days_since_1_22,world_cases,test_size=0.15,shuffle=False)

Double-click (or enter) to edit

# Using linear regression model to make predictions
from sklearn.linear_model import LinearRegression
linear_model = LinearRegression(normalize=True, fit_intercept=True)
linear_model.fit(X_train_confirmed, y_train_confirmed)
test_linear_pred = linear_model.predict(X_test_confirmed)
linear_pred = linear_model.predict(future_forecast)
print('MAE:', mean_absolute_error(test_linear_pred, y_test_confirmed))
print('MSE:', mean_squared_error(test_linear_pred, y_test_confirmed))
print('RMSE:', np.sqrt(mean_squared_error(test_linear_pred, y_test_confirmed)))

MAE: 11905.537037037033
MSE: 307996364.0100404
RMSE: 17549.82518602336
FutureWarning: 'normalize' was deprecated in version 1.0 and will be removed in 1.2.
If you wish to scale the data, use Pipeline with a StandardScaler in a preprocessing stage. To reproduce the previous behavior:

from sklearn.pipeline import make_pipeline

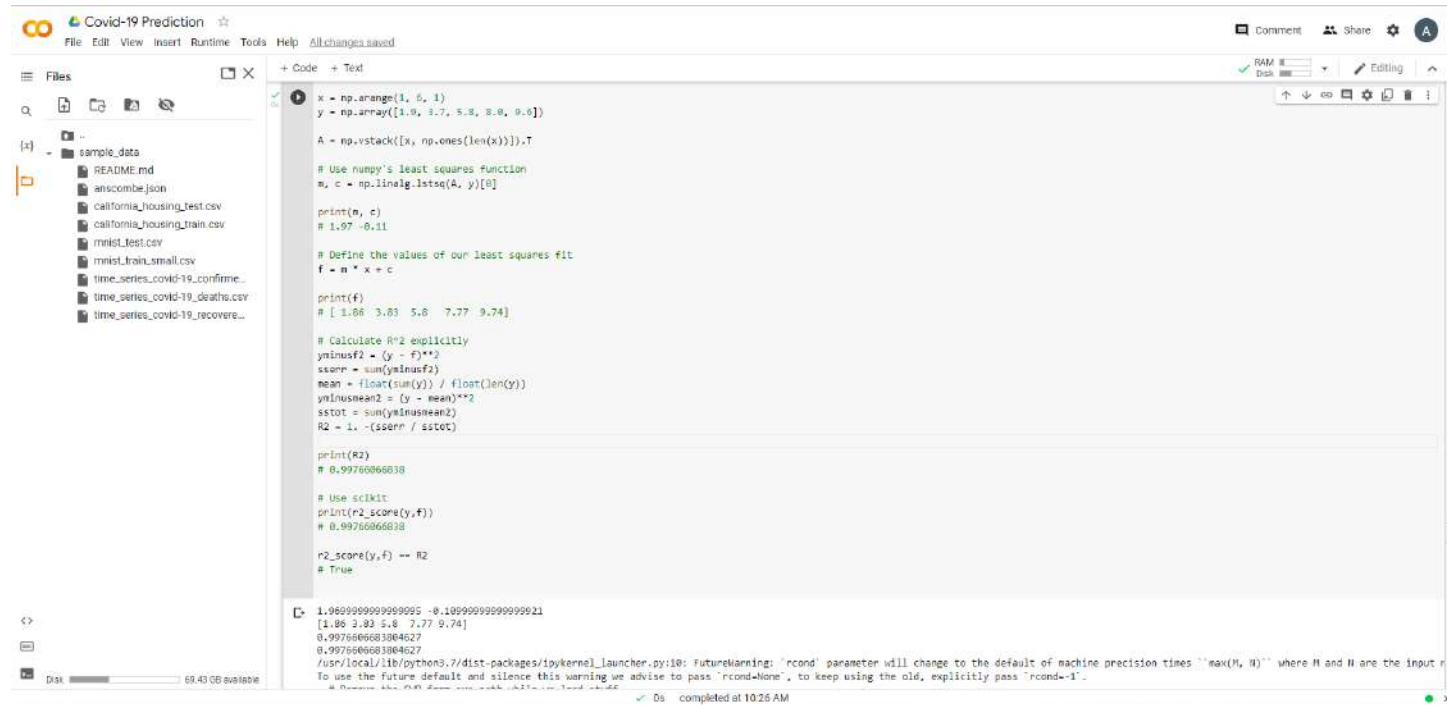
model = make_pipeline(StandardScaler(with_mean=False), LinearRegression())

If you wish to pass a sample_weight parameter, you need to pass it as a fit parameter to each step of the pipeline as follows:

kwargs = {'s[0] + '_' + sample_weight': sample_weight for s in model.steps}
model.fit(X, y, **kwargs)

FutureWarning:
```

RESULTS



The screenshot shows a Jupyter Notebook interface with the title "Covid-19 Prediction". The left sidebar displays a file explorer with a folder named "sample_data" containing several CSV files. The main area shows a code cell with the following Python code:

```
x = np.arange(1, 6, 1)
y = np.array([1.0, 3.7, 5.8, 8.0, 9.0])

A = np.vstack([x, np.ones(len(x))]).T

# Use numpy's least squares function
m, c = np.linalg.lstsq(A, y)[0]

print(m, c)
# 1.97 -0.11

# Define the values of our least squares fit
f = m * x + c

print(f)
# [ 1.86  3.83  5.8  7.77  9.74]

# Calculate R^2 explicitly
yminusf2 = (y - f)**2
sserr = sum(yminusf2)
mean = float(sum(y)) / float(len(y))
yminusmean2 = (y - mean)**2
sstot = sum(yminusmean2)
R2 = 1. - (sserr / sstot)

print(R2)
# 0.99766666838

# Use sklearn
print(r2_score(y, f))
# 0.99766666838

r2_score(y, f) == R2
# True
```

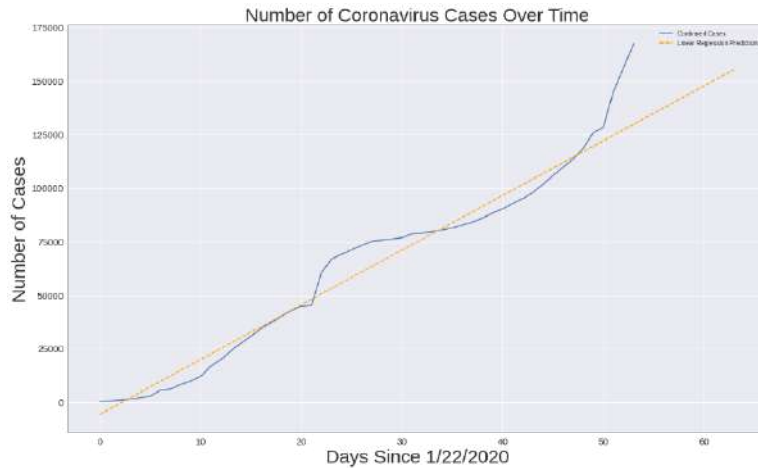
The output of the code cell is displayed below the code:

```
1.06099999999999995 -0.109999999999999921
[ 1.86  3.83  5.8  7.77  9.74]
0.9976666683804627
0.9976666683804627
```

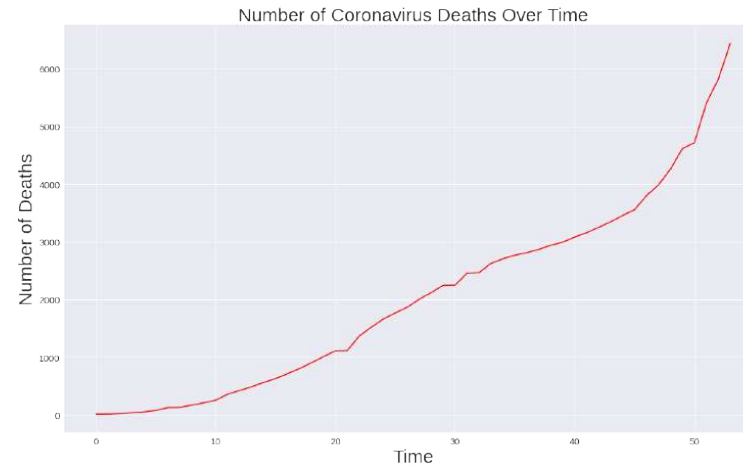
At the bottom of the notebook, a status bar indicates "completed at 10:25 AM".

The accuracy of the Regression Model is **99.7%**.

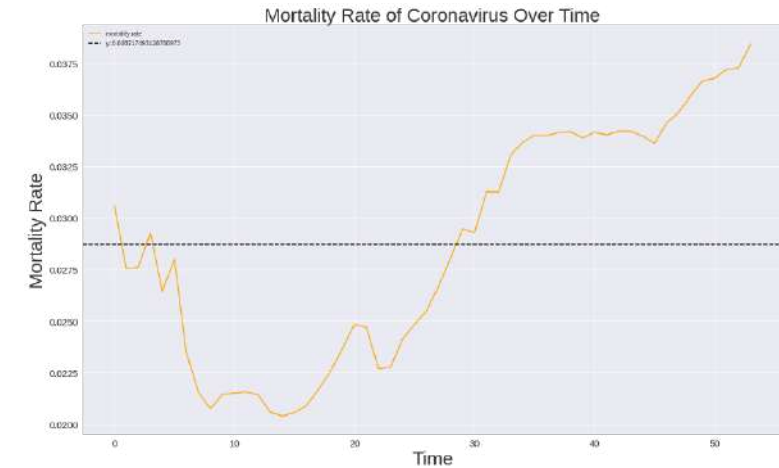
The modified Linear Regression model's performance can be visualized here:



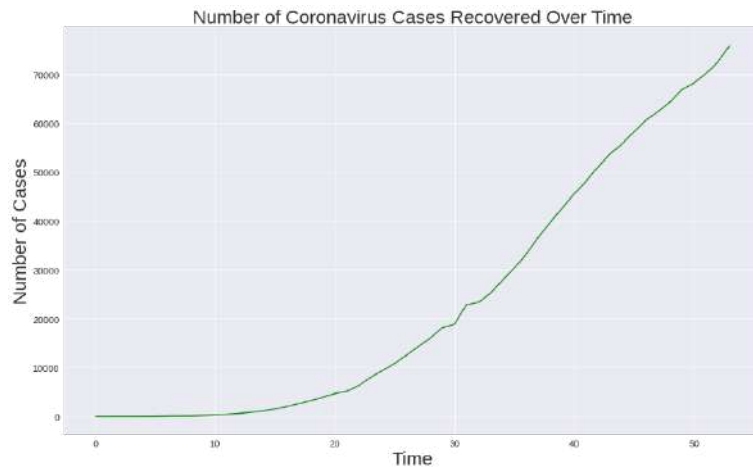
This graph shows us the Number of cases vs Days since 1/22/2020



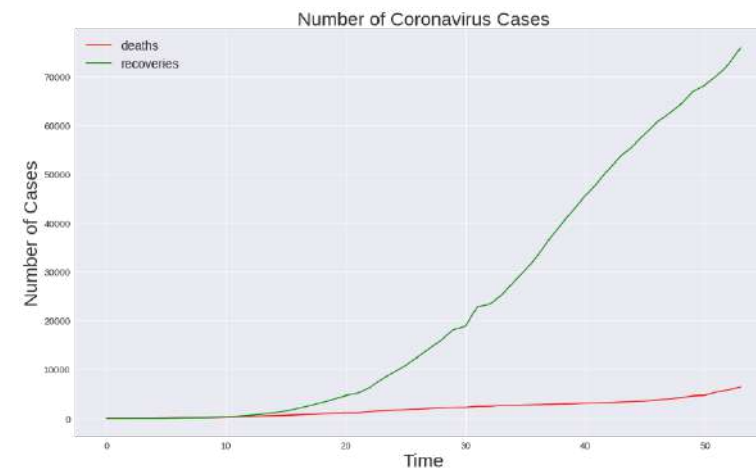
This graph shows us the number of deaths over period of time



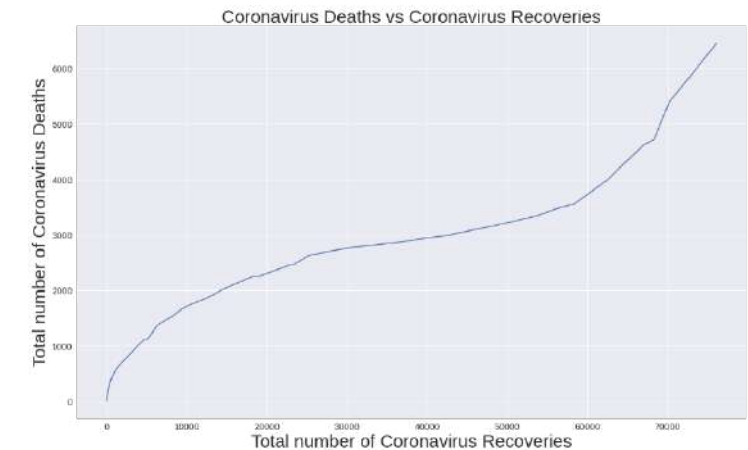
This graph shows us the Mortality rate over a period of time



This graph represents recovered corona cases over a time period



this shows the variation between deaths and recoveries over a period of time



this shows the deaths and recoveries in single parameter

Demo Video performing inputs

