DEPARTMENT OF COMPUTER SCIENCE ENGINEERING, SCHOOL OF ENGINEERING AND TECHNOLOGY, SHARDA UNIVERSITY, GREATER NOIDA

COVID-19 Outbreak Prediction using Machine learning.

A project submitted

In partial fulfillment of the requirements

for the degree of Bachelor of Technology

in Computer Science and Engineering

By

VAISHALI CHHONKAR (2019007712)
DISHA BANSAL (2019006264)
KHALILULLAH AHMADZAI (2018002354)
SAYED FAYAZ SADAT (2018011644)

Supervised by:

Dr. Anil Kumar Sagar

May 2022

CERTIFICATE

This is to certify that the report entitled **"Covid-19 Outbreak Prediction Using Machine Learning"** submitted by "Ms. Vaishali Chhonkar (2019007712), and Ms. Disha Bansal (2019006264),Mr. Khalilullah Ahmadzai (2018002354),Mr. Said Fayaz Sadat(2018011644)" to Sharda University, towards the fulfillment of requirements of the degree of **"Bachelor of Technology"** is a record of bonafide final year Project work carried out by him in the **"**Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University**"**. The results/findings contained in this Project have not been submitted in part or full to any other University/Institute for the award of any other Degree/Diploma.

Signature of Supervisor

Name: Prof. (Dr.) Anil Kumar

SagarDesignation: Asst. Prof (CSE)

Signature of Head of Department

Name: Prof. (Dr.) Nitin Rakesh

Place: Greater Noida

Date:

Signature of External Examiner
Date:

ACKNOWLEDGEMENT

A major project is a golden opportunity for learning and self-development. We consider ourselves very lucky and honored to have so many wonderful people lead us through the completion of this project.

First and foremost we would like to thank Dr. Nitin Rakesh, HOD, and CSE who gaveus an opportunity to undertake this project.

My grateful thanks to **Prof. (DR.) Anil Kumar Sagar** for his guidance in our project work.

DR. Anil Kumar Sagar, in spite of being extraordinarily busy with academics, took time out to hear, guide, and keep us on the correct path. We do not know where we would have been without his help.

CSE department monitored our progress and arranged all facilities to make life easier. We choose this moment to acknowledge their contribution gratefully.

Name and signature of Students

Vaishali Chhonkar (2019007712)

Disha Bansal (2019006264)

Khalilullah Ahmadzai (2018002354)

Said Fayaz Sadat (2018011644)

Abstract

Covid19 is the worst highly infectious and most deadly disease faced by the world. Covid19's first case was confirmed in Wuhan China in December 2019 and then expand all around the world. According to the history after every 100 years, One deadliest disease comes but Covid is the highest death rate recorded disease till now in the history of the world there are diseases which came in1800'sthird chorale in 1900's sixth cholera, Hong Kong flu, Asian flu in 2000's aids/HIV and now it is covid19. Covid19 has generated panic across the world a large-scale quarantine lockdown has changed the life of the people. Due to Covid, many people lost their lives and many people lost their employment Covid leads to many factors including employment, political issues, cancellation of manyexams, etc., which has destroyed the lives of people the people who earn daily their survival Covid has killed there earning and hopes. Lockdown plays a very important rolein getting rid of Covid 19 but the lockdown has two effects positive and negative, negative is due to lockdown many people lost their jobs and faced economical and financial issues.Board's exams are canceled first time in history due to which many students' dreams suffered. So here in this prediction, we are using machine learning for predicting the future forecast of this deadliest disease covid19.

Keywords: Covid19 Outbreak, Covid disease, Machine learning, Covid prediction, Corona virus, pandemic, forecasting, Covid outbreak.

Contents

Chapter1: INTRODUCTION

Covid 19 has been the world's greatest severe health hazard since World War II. While keeping the distribution cycle in mind while analyzing a Covid 19 outbreak, as it can help you make critical steps to reduce the prevalence rate. We can effectively allocate financial resources to areas where public health officials are most required (WHO).

The proposed model is called "Covid 19 outbreak Prediction Using Machine learning. “In orderto forecast outbreaks early, we evaluate the data and make estimates for the future. To anticipate future outcomes, here we have taken data from the Git hub and used machine learning algorithms such as linear regression and support vector for future 30 days predictions that help us to be prepared for the next deadly disease that occurs in the future that gives us an idea of the situation. Covid is the world's most serious global health concern. Countries make every effort to treat and test their citizens.

Following the lockdown, several countries are experiencing economic hardship, financial unemployment, and political unrest as is difficult to survive in a shutdown with a lack of resources and hopes. Many exams have been canceled or postponed as a result of the lockdown, and social meeting places have been closed as a result of the lockdown.

Covid outbreak prediction analysis aids in predicting this pandemic and recognizing the spread cycle, which may aid in requiring key measures to reduce the spread rate and ensuring that public health officials may allocate economic resources efficiently to areas of highest need. WHO Covid outbreak prediction using machine learning is the proposed model, and it employs algorithms such as Linear Regression support vector machine classifier SVM.

This aim is to predict the out Break before it's observed that the most effective performance is of the XG boost classifier. However, accuracy may be improved using hyper-parameter turning.

1.1 PROBLEM DEFINITION:

Covid 19 is a deadly virus that was declared pandemic by the WHO in December 2019.As a deadly disease that is contagious, It is very important to investigate the outbreak which the help of data that is collected all over the world. The goal of the project is to predict the outbreak of the disease base on the data that was collected from different sources. The aim is to identify the location or country where the likelihood of increasing cases is high and try to find a solution for the coming outbreak.

This project is also aiming of providing some hidden information about the pandemic which can be used to handle cases in an appropriate way.

Following are some reasons for choosing this project.

1. Identifying how fast the virus is spreading.
2. Show the country where the virus is the most present.
3. Predict the future outbreak of the virus.
4. It can help to predict how accurate a drug is against the virus.

1.2 PROJECT OVERVIEW:

The data was collect from an open source platform which is GitHub for data science and machine learning competitions.

The collected data contain (545 rows and 731 columns)

To understand the data and find patterns, statistical methods were applied. First of all a Univariate, Multivariate and Bivariate analysis was done on the data to understand features to extract knowledge from our features from the data. Secondly, features engineering was applied to features after finding that some features were highly correlated with each other. The feature engineering was done to handle the problem of multicollinearity which can result in a poor model generalization. Thirdly, good features were selected after looking at the correlation and p-values of features and some irrelevant ones were dropped.

Finally, some algorithms were utilized such as support vector machine, and linear regression to train the data on these learning algorithms and the state-of-the-art show model was the best among all the adapted algorithms.

1.3 Hardware specifications:

A Laptop with:

Minimum Requirements	**Hardware**
Processor / CPU	Intel processor/AMD, M1
RAM	4/8/16 GB
INTERNET CONNECTION	WIFI maximum speed limit
Memory / Disk Space	HDD/SSD higher than 500 GB
Other Hardware Components	Keyboard, Mouse, Monitor

Recommended Requirements	**Hardware**
Processor / CPU	Dual core, Intel / AMD / M1 processors
RAM	8 GB
Graphics Adapter	Intel / AMD graphics adapter with 2gb minimum graphics memory
Memory / Disk Space	120 GB SSD or higher
Other Hardware Components	HD display Monitor, keyboard, mouse

1.4 Software Specifications:

Minimum Requirements	**Software / Libraries / Applications**
Operating System	Windows / Linux / MAC (64 bit)
Programming Language	Python 3.6 or greater
IDE	Anaconda Navigator or Google Colab
Notebooks	Jupyter lab / Jupyter Notebook with anaconda or miniconda
Libraries	Numpy, Pandas , Matplotlib, Seaborn, Sklearn, and other required libraries.

Chapter2: Literature Survey

2.1 EXISTING SYSTEM:

Machine learning may be used to anticipate Covid outbreaks in a variety of ways. Many forecasts have been made in the past to predict Covid outbreaks using machine learning. [1] Machine-readable monitoring model electronic learning in machine learning was utilized in the last several predictions.

SIER estimates employing intelligence, CT scan, chest X-ray studies, and AI image is also re-utilized in forecasts generated on the Covid outbreak. Covid 19 is a global disease thatis quickly expanding. Because there are no symptoms, it is extremely difficult to detect. Estimating the number of patients that may be affected in the future by researching theory: This study used Covid-19 to forecast an infectious illness outbreak. [2] This project forecasts what will happen in the future. Machine learning supervised algorithm linear regression and support vector machine were used for 30 days (SVM).

We used real-time data from GitHub to test both of these algorithms. By evaluating several graphs, it was shown that the linear regression technique provides the best prediction rate when compared to the support vector machine. Future forecasts for Covid-19 are based on machine-readable monitoring models. Machine learning (ML) contributes to the outcomes of working to improve decision-making by predicting the future course of action.

ML models are used to predict the number of future patients who will be afflicted by COVID-19. [3] COVID-19 threats have been predicted using standard prediction models such as the total reduction and selection operator (LASSO), support vector support (SVM), linear regression (LR), and big display smoothing (ES). The fact that ES is doing so well in the current field of prediction stems from study findings that show that, given the type and quantity of the database, it is capable of doing so. LR and LASSO also perform a good job of predicting death rates and confirming cases to some extent.

The Covid-19 Case: Examining Rehabilitation Using Reading Theory This study looked at how to predict how long it would take to recover from an infectious disease outbreak. The error learning method is used here, with objection measures such as treatment, segregation, social isolation, and so on, specifically to harmonize control of virus spreads by reducing infection levels if this is effective, and the rate of infection, in the background to reach the top,

Down to follow what is known as the Universal Recovery Curve. [4] The threat of when the rate of infection is slow, infectious diseases are infrequent, a phenomenon known as 'loosening of the curve.

When the rate rises, it should fall as a result of effective countermeasures. [5] Data mining and data analysis are research fields in which data is mined and analyzed. Data In the First Covid-19 Epidemic, Death, Prevention, and Drug Development Data mining of scientific literature records from the Core Science [6]Web Collection was utilized to obtain Covid-19 death facts, immunizations, and vaccines.

The analysis compares records from throughout the Covid-19 study topics, with individual records being evaluated separately. [7] Recurrent neural network-LSTM, ARIMA, and Prophet on the COVID-19 data to forecast the future effects of the pan-demic. Study the impact of some new parameters such as population statistics and life expectancy, etc., in the prediction of COVID-19 spread.

2.2 PROPOSED APPROACH:

COVID-19 outbreak prediction model aims to solve the future occurring problems by today's youth using the present analysis we are examining the future forecast by using machine learning technique. This model is designed so that it results in faster, flexible, and can meet human, and government needs. And it can also fulfill the total needs of providing security, medical support, and early trigger about the rate of getting infected.

The proposed approach focuses more on getting accurate data to identify the affected areas using machine learning algorithms. The objective is to develop an efficient COVID-19 analyzing system following certain major steps starting from research in the focused field, thereby collecting testing data for the implementation.

Firstly, we looked for the data collection then train the data to full fill and check our model requirement followed by the implementation of a confusion matrix in order to detect the efficiency of the algorithm as well. Moving forward, it'll be followed by analyzing the prediction rate and comparing the algorithms which is giving a more accurate scenario. This approach will move towards completion with testing and verification as per the requirements.

2.3 FEASIBILITY STUDY:

The feasibility study helps to understand the project scope and whether the project is feasible with the available technology, budget, and resources.

For this project the biggest challenge was how to get the data set, we cannot talk about machine learning without data. We finalize our decision about the project only when we were 100% sure that we got relevant data.

Different areas for the feasibility study are shown below:

- **Technology**

Whether the project technically feasible or not?

Does the project use state-of-the-art technology?

Can the project cope with future changes in components and technology?

- **Cost / Finance**

Whether the project is financially feasible or not?

Does the project exceed the allotted budget?

Is the outcome of the project freely available?

- **Time**

Can the project be completed within the given time?

- **Resource**

How many resources do the project need in order to be completed?

Whether the required resources available or not?

The major variables used for this study are:-

a) Technical Feasibility

b) Cost Feasibility

a) Technical Feasibility

The technical feasibility is to find out if the project is technologically feasible, that means, the technical requirements of the proposed system and the availability of the required resources. It was found that with the help of available machine learning and preprocessing packages our project is technically feasible we used different packages such as:

- Scikit-learn
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Missingno

b) Cost Feasibility

The purpose of cost feasibility is to determine whether the project can be completed with the given budget. It gives an idea about the cost of money that can be spent on the research in order to meet the end goal of the project.

The project is feasible because it uses open source so all the material that has been utilized in the project is a fruit of open source projects which is freely available at no cost.

2.3.1 Training Data:

Training data were collected alongside the testing data further preprocessing is needed to know better the predictors. Training data helps us to prepare a budget request at some point and it's a proper document for building a business case and justifying budget requests.

The training data is the data that we will use to fit into the machine learning model in order to

Further, see how well it can generalize in unseen data. The training data should contain the label which is the response (target) variable in our it is the case of Regression.

Moreover, this project is based on binary classification which is to determine whether a person has a bank account or not. The training data contain different features and it's our job to determine whether to include that feature or not. The selection of features is done with the help of statistical analysis such as correlation and others.

2.3.2 Testing Data:

For any machine Learning project we should have training data and testing data, the testing data should be similar to the training data in terms of features except that the testing data doesn't have a target column.

After fitting the data to the learning algorithm it is required to check how well the model performs and where the testing data come into use.

The testing data is always less than the training data and that's important for the model to learn from too much experience. Testing data gives the actual performance of the model with which we can check how well the model fits with the unseen data. This evaluation will lead us to more optimization and generalization of the machine learning algorithm.

2.3.3 Predictive Features:

To understand what drives the target outcome, there should be some research or an investigation to get ideas on the data points. Once the quality of understanding of what will fit well, and the target outcome is achieved, further process of data requests can help build a businesscase.

2.3.4 Data Sources:

The source of data can be many. For example, if we have access to internal data then that is the best, or the data that needs work from a DevOps or data engineer is the 15-second best. In case that data is not available internally it might be available externally from a third-party company, organization, or the government. Research and surveys are other most effective data collection methods if enough time and budget are available. Web scraping is another effective data collection method if the data is available on the internet on different websites. We have obtained the data set from the Git hub datasets.

2.4 Risk Management:

Risk management helps to prevent the failure of the proposed system when working with real-world data. This will also help us to get an idea about different types of risks that may occur in the present and future as well.

2.2.4.1 Risk Identification

The first step to managing any risk is to first identify the risk. It consists of finding out potential risks that may occur within a given time.

2.2.4.1.1 Data quality-related

If the collected data was not relevant to solving the problem or the way they collected the data wasn't appropriate.

2.2.4.1.2 Machine learning algorithms related

Failing to use appropriate machine learning algorithms will result in bad performance.

2.2.4.1.3 Data Imbalance related

Training the algorithm on imbalanced data. This risk is a big problem in machine learning known as bias.

2.2.4.1.4 Data leakage related

Data leakage risk, failing to split the data properly into train and test sets during model training will result in bad performance of the model.

This is a serious problem in data analysis and machine learning and it can lead to serious consequences that can affect the generalization dispute. The goal is to build a model that can be generalized in unseen data.

2.2.5 Risk Mitigation

Collection of the data from trusted organizations and online platforms can reduce the risk of data quality

Using the state of the art machine learning algorithms to solve the problem we can prevent using old-fashioned algorithms which may not be able to find the patterns in the data.

Since data imbalance is a common issue in datasets using different oversampling and under sampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) prevents from biasedness of machine learning algorithms

Adopting different data splitting methods such as train-test-split from sklearn before training the model will prevent data leakage.

Chapter 3: System Analysis and Design

3.1 Software Requirement Specification:

3.1 Workflow

The following steps describe our machine learning workflow:

3.1.1 Objective

To predict the accuracy rate using machine learning algorithms.

3.1.2 Gathering Data

Data gathering is the first step in choosing any machine learning project. We should first try
To find the data and the process of finding data requires asking so many questions such as
Is the data we have collected relevant for solving the problem?
Is the data enough to apply machine learning?
Is the data redundant or not if so why?
Is the data appropriate for solving the problem and is the problem solvable with machine learning?

The data set was obtained from the online platform Git Hub.

3.1.3 Prepare Data

Data should be free from redundancy, missing values, and duplications.

We start preparing the data first by removing inconsistency which means as data is coming from different sources it is always prone to inconsistency. Inconsistent data may lead to different problems or even failing to apply machine learning to that particular data set. We tried to find the data set which fits our requirements and helps us to give the best accuracy rate.

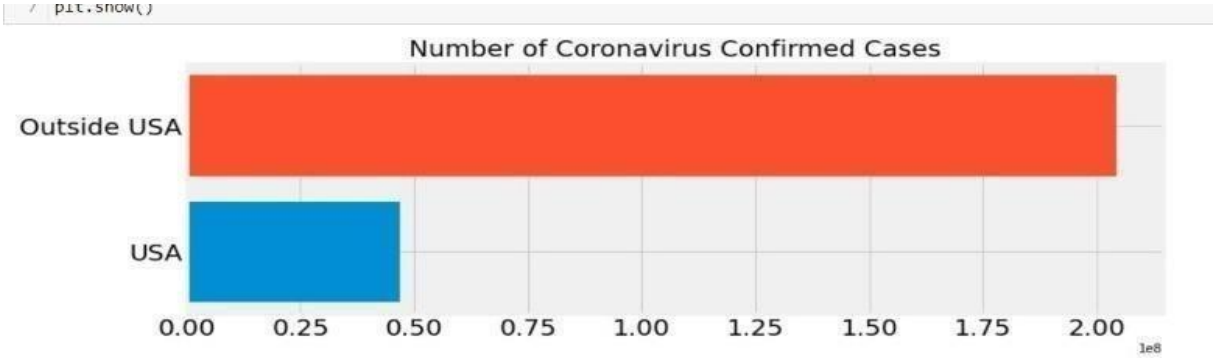


FIGURE 1: Variance in the outside USA and in the USA confirmed cases.

Description- Graph shows the ratio of the cases in comparison with other countries and how much USA population is infected with coronavirus in comparison with other countries. Thegraphs show clearly the level of confirmed cases in the USA which is high but not more than outside countries number of confirmed cases got higher in the USA due to the late information of coronavirus what this virus is all about.

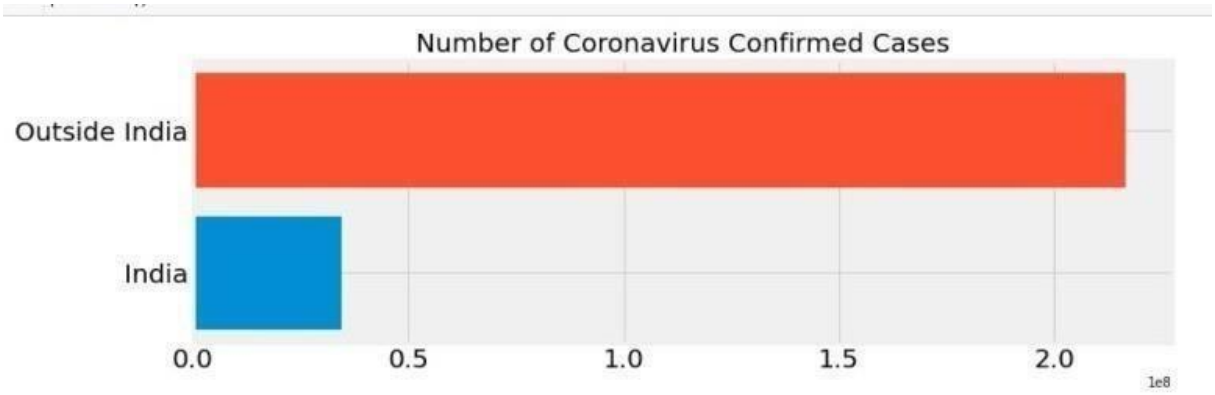


FIGURE 2: Variance in outside India and in India confirmed cases.

Description- This Graph shows the ratio of the cases in comparison with other countries and how much India's population is infected with coronavirus in comparison with other countries. The graphs show clearly the level of confirmed cases in India which is high but not more than outside countries number of confirmed cases got higher in India due to the huge population and gatherings at the begging of the reporting of viruses in India. But also because people were going from one place to another the spread of the virus was high which made the country get infected at a very high level.

3.1.4 Select Algorithm

Most of the time it's very difficult to select a machine learning model without fitting the data into the model. For that purpose, we decided to go with multiple machine learning models.

3.1.5 Model Training

Training the model is also called the model learning process. It is the step on which different algorithms will try to find the patterns or relationships between the features and target. This process may take varying times depending upon the type of algorithms used, size of the training data, pre-processing of the data, etc. After the model training is completed the models will be given different unseen data to check the actual performance.

3.1.6 Model Testing

The data was split into 70% of training and 30% of testing because we have a high sample. So we use that 30% of the population sample to test how accurate our models perform.

Different evaluation metrics were used to evaluate the performance but we consider the f1-score as the main evaluation matrix due to the fact that the data was unbalanced initially. The result shows that our features engineering was good as we got good results in most of the utilized learning algorithms.

3.2 Data Flow Diagram

The Data flow is a pipeline that visualizes how the data will flow starting from the time of collecting the data until the final evaluation.

Figure 1: shows the data flow diagram of the project.

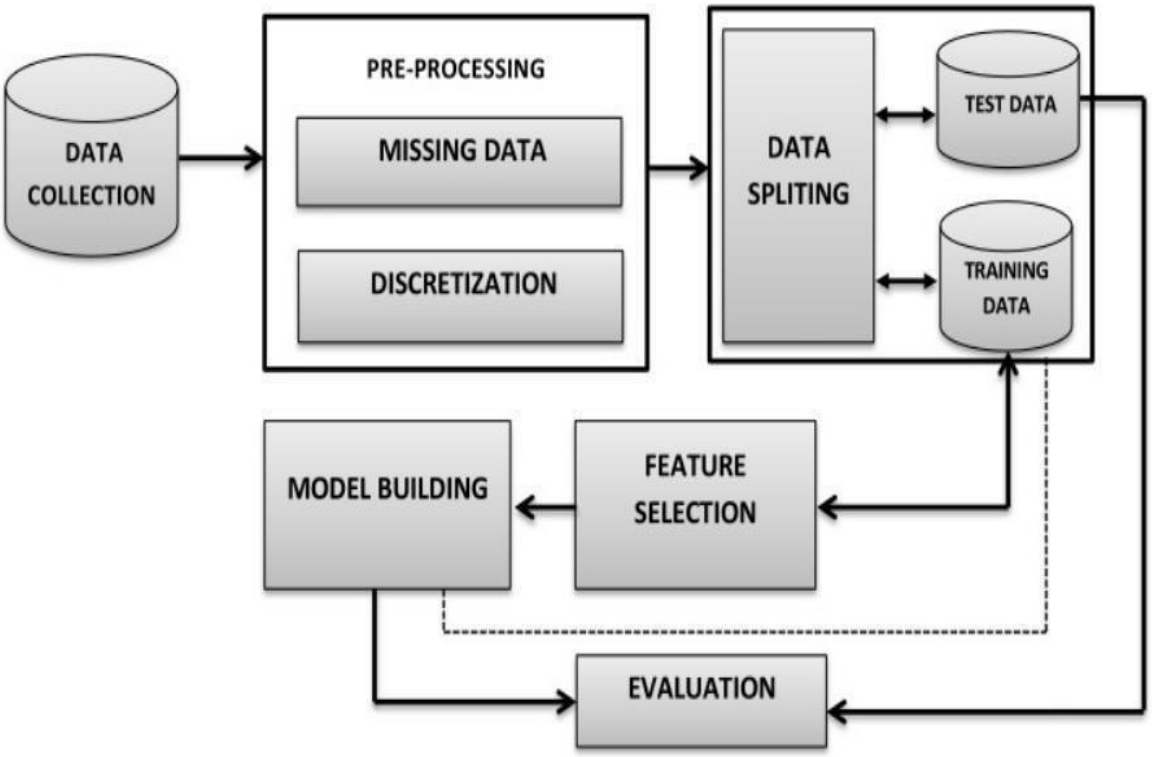


Fig 1: Data Flow diagram for machine learning

Research and exploring

Firstly, we did research on the topic of COVID-19 outbreak prediction. Research and exploring are important as it focuses on the issues that help improve the project's effectiveness.

3.2.1 Data Collection

The data used for the research was retrieved from GitHub which was extracted from various Fin-scope surveys. The data set was in Microsoft Excel in CSV (comma separated variable) format.

3.2.2 Data Preprocessing

Firstly we segregate the data Country, Areas, Regions vise and according to this data, we find out the death, recovery and confirmed cases rate. With help of this scenario, it's easy to predict further the accuracy rate.

Below is the code for confirmed, death, and recovery cases in the world.

```
In [2]: confirmed_cases = pd.read_csv("datasetnew\\COVID-19\\csse_covid_19_data\\csse_covid_19_time_series\\time_series_covid19_conf
        confirmed_cases.head()
```

Out[2]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	10/31/21	11/1/21	11/2/21	11/3/21	11/4/21	11/5/2
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	156250	156284	156307	156323	156363	15639
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	185300	185497	186222	186793	187363	18799
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	206452	206566	206649	206754	206878	20699
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	15516	15516	15516	15572	15618	1561
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	64433	64458	64487	64533	64583	6461

5 rows × 662 columns

```
In [3]: deaths_reports = pd.read_csv("datasetnew\\COVID-19\\csse_covid_19_data\\csse_covid_19_time_series\\time_series_covid19_death
        deaths_reports.head()
```

Out[3]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	10/31/21	11/1/21	11/2/21	11/3/21	11/4/21	11/5/2
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	7280	7281	7281	7284	7284	728
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	2924	2931	2937	2940	2944	294
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	5920	5924	5927	5931	5936	593
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	130	130	130	130	130	13
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	1710	1713	1713	1716	1718	171

5 rows × 662 columns

```
In [4]:
recoverd_cases = pd.read_csv("datasetnew\\COVID-19\\csse_covid_19_data\\csse_covid_19_time_series\\time_series_covid19_recov
recoverd_cases.head()
```

Out[4]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	10/31/21	11/1/21	11/2/21	11/3/21	11/4/21	11/5/2
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	0	0	0	0	0	
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	0	0	0	0	0	
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	0	0	0	0	0	
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	0	0	0	0	0	
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 662 columns

3.2.2.2 Feature Selection:-

Feature selection is the most crucial part of any machine learning project. Selecting the features that are highly important for identifying underlying patterns helps to get better results. In order to select features, we applied some statistical analysis to identify if there is any multicollinearity between independent features.

Code snippets:

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

```
#!pip install numpy
#!pip install pandas
#!pip install matplotlib
#!pip install sklearn
```

```
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.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.svm import SVR
#from sklearn.svm import LinearSVR
from sklearn.metrics import mean_squared_error, mean_absolute_error

import datetime
import operator
plt.style.use('fivethirtyeight')
%matplotlib inline
```

```
for i in dates:
    confirmed_sum = confirmed[i].sum()
    death_sum = deathes[i].sum()
    recovered_sum = recovered[i].sum()


    world_cases.append(confirmed_sum)
    total_deaths.append(death_sum)
    total_recovered.append(recovered_sum)
    total_active.append(confirmed_sum-recovered_sum-death_sum)
    mortality_rate.append(death_sum/confirmed_sum)
    recovery_rate.append(recovered_sum/confirmed_sum)


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

```
print('Outside USA {} cases:'.format(outside_USA_confirmed))
print('USA {} cases:'.format(USA_confirmed))
print('Total {} cases:'.format(USA_confirmed+outside_USA_confirmed))
```

```
Outside USA 204591728 cases:
USA 46744841 cases:
Total 251336569 cases:
```

```
INDIA_confirmed = latest_data[latest_data['Country_Region']=='India']['Confirmed'].sum()
outside_INDIA_confirmed = np.sum(country_confirmed_cases) - INDIA_confirmed

```

```
plt.figure(figsize=(12,4))
plt.barh('India',INDIA_confirmed)
plt.barh('Outside India', outside_INDIA_confirmed)
plt.title('Number of Coronavirus Confirmed Cases', size = 20)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```

3.2.4.1 Cross-Validation

The goal of any machine learning project is to create a model that can generalize on unseen data.

Some machine learning algorithms are prone to overfitting and it's very important to consider the bias-variance trade-off.

A good machine learning model is one that has low bias and low variance.

3.2.4.1.1 Low variance

it tells you the smallest changes in the data set to cause the result to change in the target function.

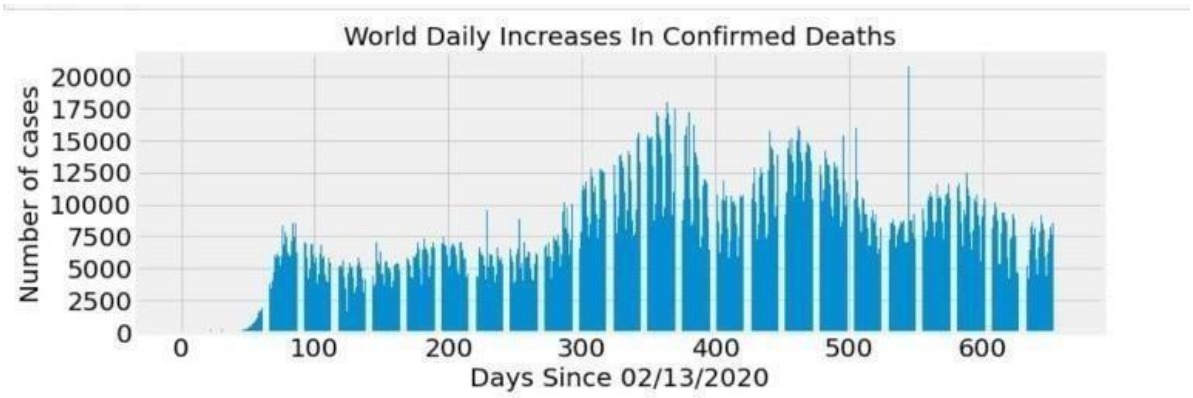


FIGURE 1: Plot daily increase in Confirmed cases

Description-In this graph on the x-axis with a scale of 100 difference with the heading Days since 02/13/2020 and on a y-axis scale of 0.2 difference variation with heading number of cases is shown. This graph represents the number of World daily increases in cases of Corona in the world over time.

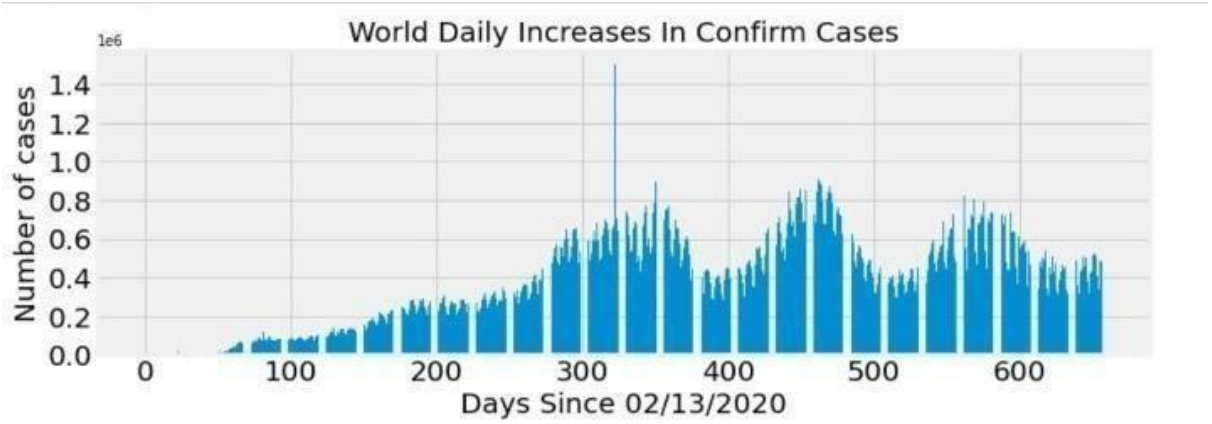


FIGURE 2: Plot daily increase in Confirmed Recoveries.

Description-In this graph on the x-axis with a scale of 100 difference with the heading Days since 02/13/2020 and on a y-axis scale of -0. 2 difference variation with heading number of cases is shown. This graph represents the number of World daily increases confirming the recovery in the world over time. This outcome had come with the data we have taken for the prediction. The recovery for corona patients had come after such a long time in India because of not availability of vaccines and proper cure of this deadly disease that's why therecovery rate is low over the period of time.

3.2.4.1.2 High variance

it will tell the big change has to occur so that the objective function changes in its estimates.

- **How do we detect over-fitting?**

It is very difficult to identify over-fitting however there are some techniques that help us to do so.

To detect over-fitting in our machine learning model, we would need a way to test the model on unseen data.

We often refer to this as Cross-validation whether we want to evaluate the performance of the algorithm on unseen instances. Cross-validation is a variety of techniques that assess the performance of the predictive model generalization capabilities to an independent data set that wasn't introduced to the model.

3.2.2.3 Over-fitting

Over-fitting is when you train your model and evaluate it and get a high accuracy but the model will fail to generalize on unseen data.

Some learning algorithms such as Random forest are prone to over-fitting.

3.2.2.4 Under fitting

Under fitting occurs when the model fails to capture the underlying logic of the data. This meansthat we will get high errors in our training and high errors in our testing phase which is not good and this is referred to as high bias and high variance.

❖ **How do we solve this problem?**

As it's common in machine learning to overfit or underfit during the training or testing phase, to fight against this we can introduce some bias or penalty term to the model and that process is known as regularization techniques such as L1 and L2 regularization.

Over-fitting occurs also due to the complexity of the model so changing the default parameters of the learning algorithm also fights against over-fit. The following is a list of parameters to use.

- Learning Rate.
- Max depth
- Lambda which is the regularization penalty.
- Number of estimators

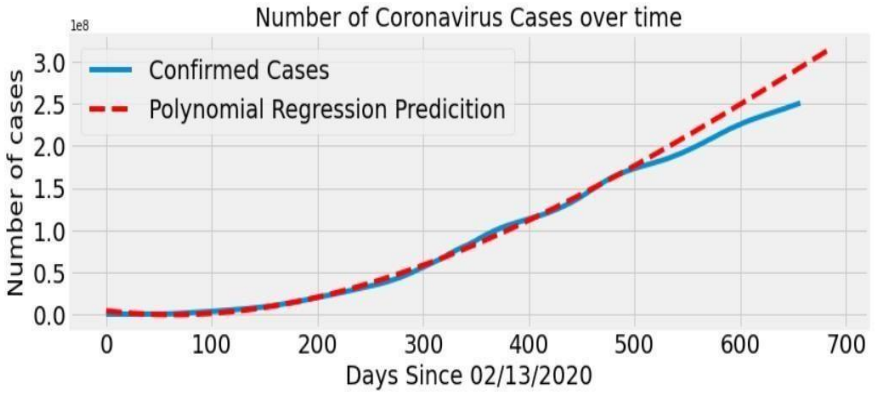
Code snippets of different machine learning algorithms training and predicted graph

.

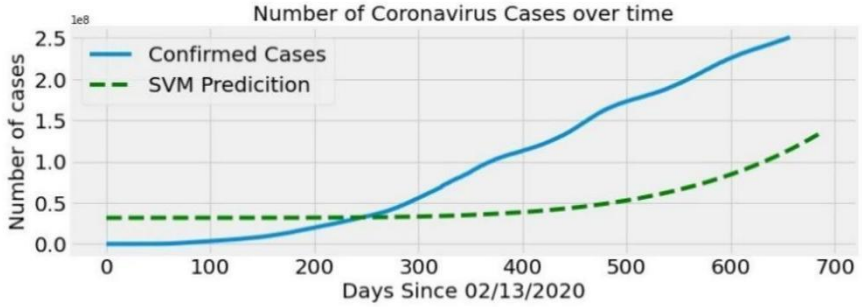
In [164]:
```
def plot_predections(x, y, pred, algo_name, color):
    plt.figure(figsize=(12,4))
    plt.plot(x,y)
    plt.plot(future_forecast, pred, linestyle='dashed', color=color)
    plt.title('Number of Coronavirus Cases over time',size=20)
    plt.xlabel('Days Since '+date_since,size=20)
    plt.legend(['Confirmed Cases', algo_name], prop={'size':20})
    plt.ylabel('Number of cases',size=20)
    plt.xticks(size=20)
    plt.yticks(size=20)
    plt.show()

```

In [165]:
```
plot_predections(adjusted_dates, world_cases, linear_pred, 'Polynomial Regression Predicition', 'red')
```



In [156]:
```
plot_predections(adjusted_dates, world_cases, svm_pred, 'SVM Predicition', 'green')
```



In [175]:
```
def plot_predections(x, y, pred1, pred2, algo_name1, algo_name2):
    plt.figure(figsize=(12,6))
    plt.plot(x,y)
    plt.plot(future_forecast, pred1, linestyle='dashed', color='red')
    plt.plot(future_forecast, pred2, linestyle='dashed', color='green')
    plt.title('Number of Coronavirus Cases over time',size=20)
    plt.xlabel('Days Since '+date_since,size=20)
    plt.legend(['Confirmed Cases',  algo_name1,  algo_name2], prop={'size':20})
    plt.ylabel('Number of cases',size=20)
    plt.xticks(size=20)
    plt.yticks(size=20)
    plt.show()
```

3.2.2.6 Prediction

In our prediction model, COVID-19 outbreak prediction the linear regression (polynomial regression) prediction gives better accuracy than the support vector machine algorithm. In this graph, the prediction rate is clear by seeing this graph we can say that the polynomial regression prediction line plot is near to the outcome of the confirmed case these lines are plotted by using our data set.

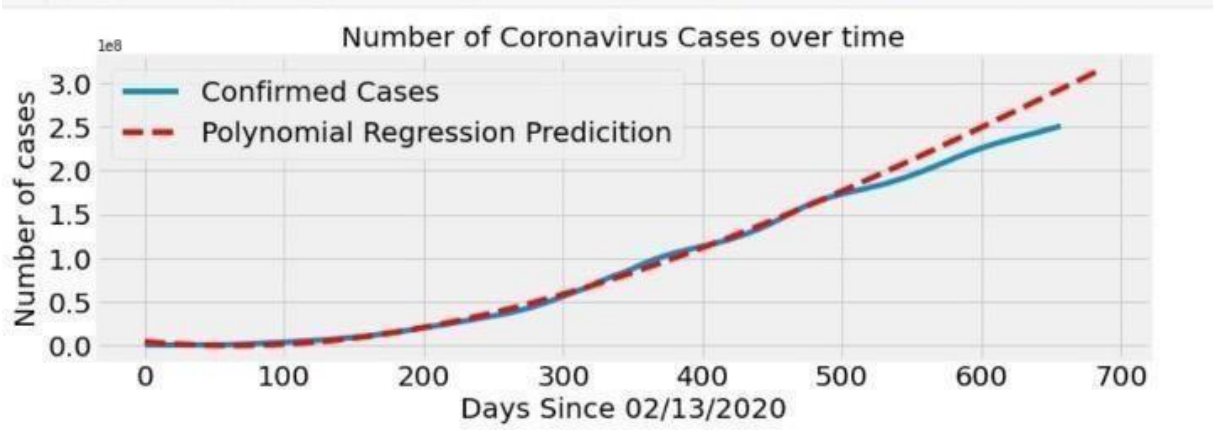


FIGURE 1: Prediction graph in comparison of confirmed cases and polynomial Regression prediction.

Description-In this graph on the x-axis with a scale of 100 difference with heading Days since 02/13/2020 and on the y-axis scale of 0.5 difference variation with heading number of cases is shown. This graph represents the prediction rate of coronavirus confirmed cases in the world over time. This outcome has been analyzed with the data we have taken for the prediction. The recovery for corona patients had come after such a long time in India because of not availability of vaccines and proper cure of this deadly disease that's why the recovery rate is low over the period of time. In this graph, the blue line shows the best fit line of confirmed cases with the help of data taken and the red lines show the prediction rate of polynomial regression prediction.

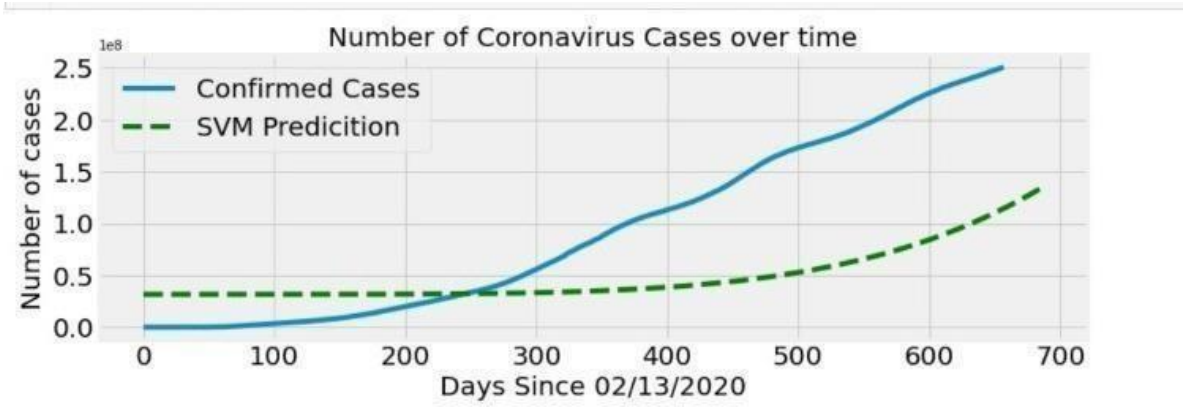


FIGURE 2: Prediction graph in comparison of confirmed cases and SVM.

Description-In this graph on the x-axis the scale of 100 difference with heading Days since 02/13/2020 and on the y-axis scale of 0.5 difference variation with heading number of cases is shown. This graph represents the prediction rate of coronavirus confirmed cases in the world over time in comparison with the SVM.

This outcome has been analyzed with the data we have taken for the prediction. In this graph, the blue line shows the best fit line of confirmed cases with the help of data taken and the green lines show the prediction rate of SVM.

3.3 Design and Test Steps/Criteria:

```
start = '11/09/2021'
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'))
```

```
X_train_confirmed, X_test_confirmed, y_train_confirmed, y_test_confirmed = train_test_split(days_since_1_22,
```

```
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_forecast = poly.fit_transform(future_forecast)
```

```
country_df =pd.DataFrame({'Country Name': unique_countries, 'Number of confirmed cases': country_confirmed_cases,
                          'Number of Deaths':country_death_cases, 'Number of recoveries': country_recovery_cases,
                          'Number of Active Cases': country_active_cases, 'Mortality Rate': country_mortality_rate })

country_df.style.background_gradient(cmap="Blues")
```

Country Name	Number of confirmed cases	Number of Deaths	Number of recoveries	Number of Active Cases	Mortality Rate
US	46744841	758553	0.000000	45986288.000000	0.016228
India	34388579	461849	0.000000	33926730.000000	0.013430
Brazil	21903249	610044	0.000000	21293205.000000	0.027852
United Kingdom	9412185	142556	0.000000	9269629.000000	0.015146
Russia	8727817	244588	0.000000	8483229.000000	0.028024

```
unique_provinces = list(latest_data['Province_State'].unique())
```

```
province_confirmed_cases = []
province_country = []
province_death_cases = []
province_recovery_cases = []
province_mortality_rate = []
```

```
province_df = pd.DataFrame({'Province/State Name':unique_provinces, "Country": province_country,
                            'Number of Confirmed Cases': province_confirmed_cases, 'Number of Deaths' : province_death_cases ,
                           'Number of Recoveries' : province_recovery_cases, 'Mortality Rate' : province_mortality_rate})
province_df.style.background_gradient(cmap='Reds')
```

	Province/State Name	Country	Number of Confirmed Cases	Number of Deaths	Number of Recoveries	Mortality Rate
0	England	United Kingdom	7953228	123800	0.000000	0.015566
1	Maharashtra	India	6619329	140430	0.000000	0.021215
2	Kerala	India	5027318	34362	0.000000	0.006835
3	California	US	4983595	72681	0.000000	0.014584
4	Sao Paulo	Brazil	4414187	152529	0.000000	0.034554
5	Texas	US	4298123	72111	0.000000	0.016777

```
nan_indices = []

for i in range(len(unique_provinces)):
    if type(unique_provinces) == float:
        nan_indices.append(i)

unique_provinces = list(unique_provinces)
province_confirmed_cases = list(province_confirmed_cases)


for i in nan_indices:
    unique_provinces.pop(i)
    province_confirmed_cases.pop(i)
```

```
USA_confirmed = latest_data[latest_data['Country_Region']=='US']['Confirmed'].sum()
outside_USA_confirmed = np.sum(country_confirmed_cases) - USA_confirmed
```

3.4 Testing Process

FIGURE: 1

Description: In this graph on the x-axis and y-axis, the increase of confirmed cases in the world is shown incomparison with the tested data and polynomial regression. [9]The best fit of polynomial regression prediction is shown. In this, the blue lines show the test data variation, and the red lines show the polynomial regression variation.

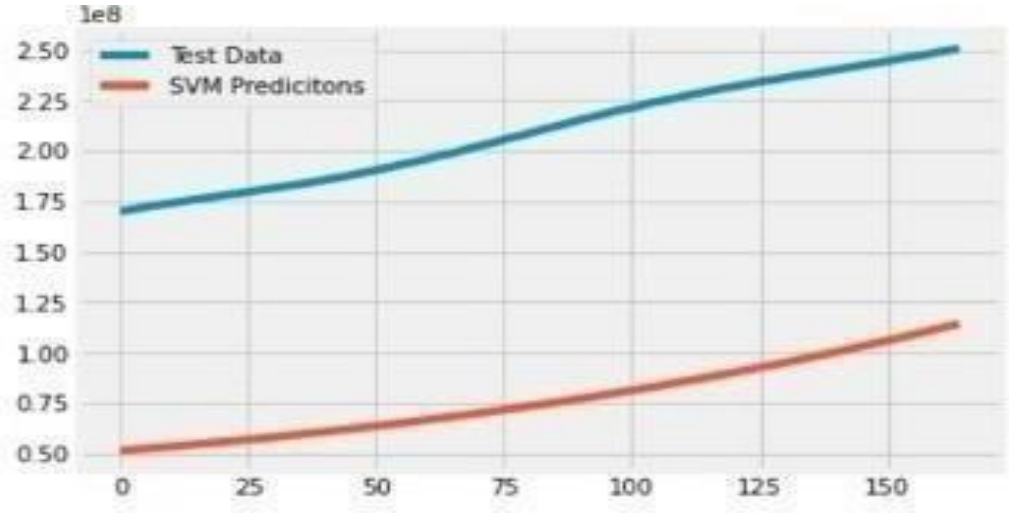


FIGURE: 2

Description: In this graph on the x-axis and y-axis, the increase of confirmed cases in the world is shown in comparison with the tested data and Support Vector Machine (SVM). [10]The best fit of SVM prediction is shown. In this, the blue lines show the test data variation, and the red lines show the SVM variation.

3.4.1 Introduction

The use of testing is to ensure every component of a system is working as required. Testing requires monitoring and investigation of different modules present in the system and provides information regarding the quality of the output or the results produced. It consists of a continuous cycle of observational science and evaluation. For a machine learning system testing involves inquiry of data quality, data size, usage of used algorithms, methods, and the reliability of the produced results.

3.4.1.1 Unit Testing

In unit testing, all the components are evaluated separately. This ensures reliability, security, and the functionality of the system. Each component is tested using different parameters which the system can face in the coming time.

The unit testing functions that will be tested are as follows:

Get the correct input features as input to the machine learning model:

- Preprocessing of the data can be used.
- Separate features and the target variable
- Use of cross-validation

3.4.2 Integration Testing:-

All the modules are combined together and tested as a whole during the integration phase. Modules that are unit-tested are integrated into a single system and the whole system is analyzed and passed through different tests. Integration testing ensures the proper functioning of the complete system at once.

3.3.1 Validation Testing

Validation testing is conducted in order to ensure whether the system or process which we have built satisfies the needs of the users or not. It is basically carried out in order to test the system from the users' perspective.

➢ Methodology

We are using machine learning algorithms to predict the outbreak of covid19 using supervised classification algorithms linear regression as polynomial and support vector machine (SVM) as a classifier. We have created a virtual environment in the anaconda Jupiter notebook we are using python in the 3.8 version for coding [7] it is also used for classification and regression analysis. SVM uses labeled data. Regression analysis is a type of predictive modeling technique that looks at how a dependent and independent variable are related. Regression analysis is used to determine the strength of predictors, forecast an effect, and forecast trends.

Types of Regression analysis:

- Linear regression.

- Logistic regression.

Consider a straight equation line that combines any two variables X and Y and may be written as:

The linear regression algorithm is a sort of regression analysis that involves establishing a link between the dependent and independent variables. It has a linear structure.

$$Y = aX + b$$

Based on the input variables in a linear combination (x). The dependent variable is the one that wishesto be predicted.

The independent variable is the one we're utilizing to predict another variable.

Linear regression is a technique for predicting a variable's value based on the value of another Variable. This model is created with the use of a straight line [8] and a continuous variable. It is determined by the amount of money lost (R).

LINEAR REGRESSION

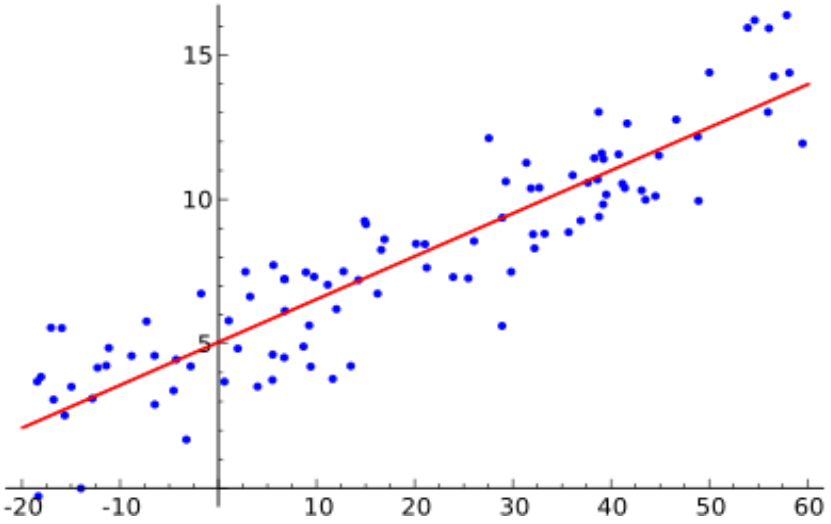


FIGURE: 1

This example employs simple linear regression, where the space between the red line and Each sample point is lowered by the square of the distance.

LOGISTIC REGRESSION

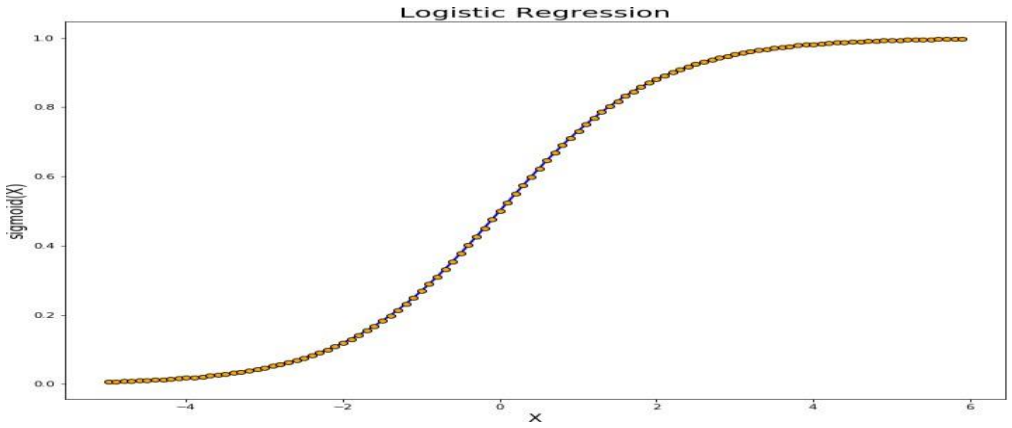


FIGURE: 2

We don't draw a straight line through our data in linear regression. As a result, data were fitted

To assigned curve, an S-shaped bend line.

Pseudo RR2 Equation

Chapter 4: Results/ Outputs

This project "covid19 outbreak prediction using machine learning "is a prediction project. After the prediction and the final testing have been done using ML supervised algorithm which is linear regression and SVM. And after doing the testing with live data which we have taken from GitHub and then comparing both the Algorithm with live data we found that the linear regression algorithm is best for the prediction covid19 outbreak in this research.

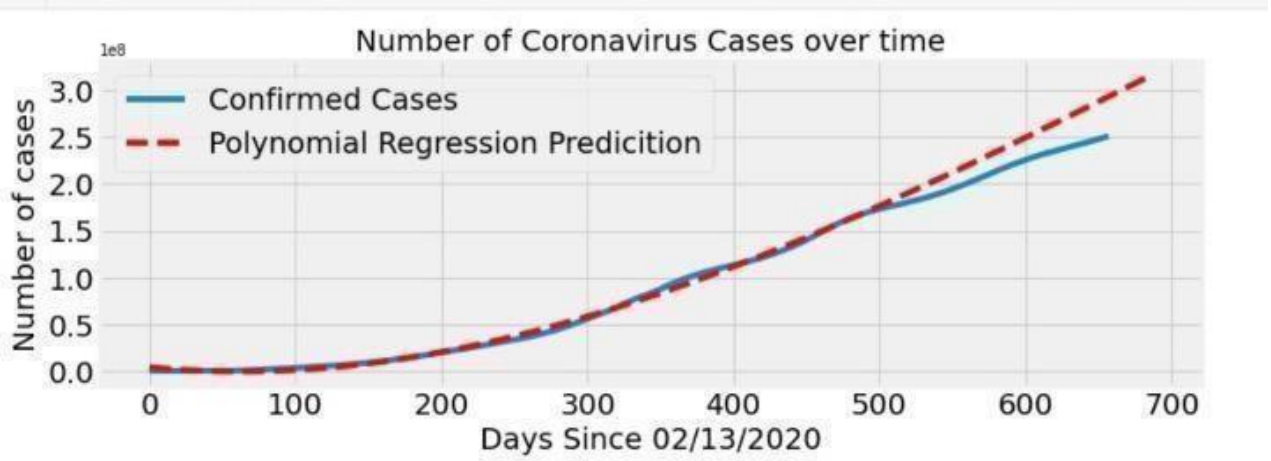


FIGURE 1 Plot in comparison to confirmed cases polynomial regression.

Description- In this graph on the x-axis and y-axis the increase of confirmed cases in the world is shown in comparison with the tested data and polynomial regression the best fit of polynomial regression prediction is shown. In this, the blue lines show the test data variation and the red lines show the polynomial regression variation.

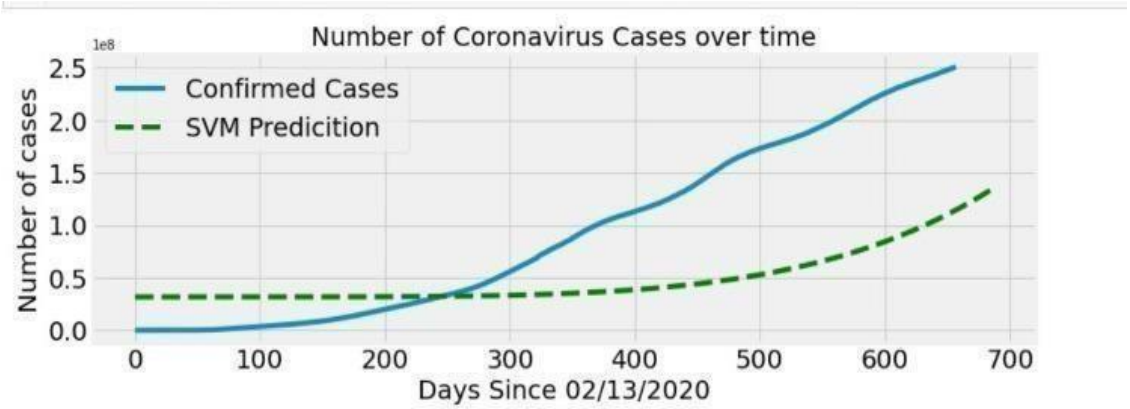


FIGURE 2 Plot in comparison to Confirmed cases and SVM.

Description- In this graph on the x-axis and y-axis the increase of confirmed cases in the world is shown in comparison with the tested data and Support Vector Machine (SVM). The best fit of SVM prediction is shown. In this, the blue lines show the test data variation,and the red lines show the SVM variation.

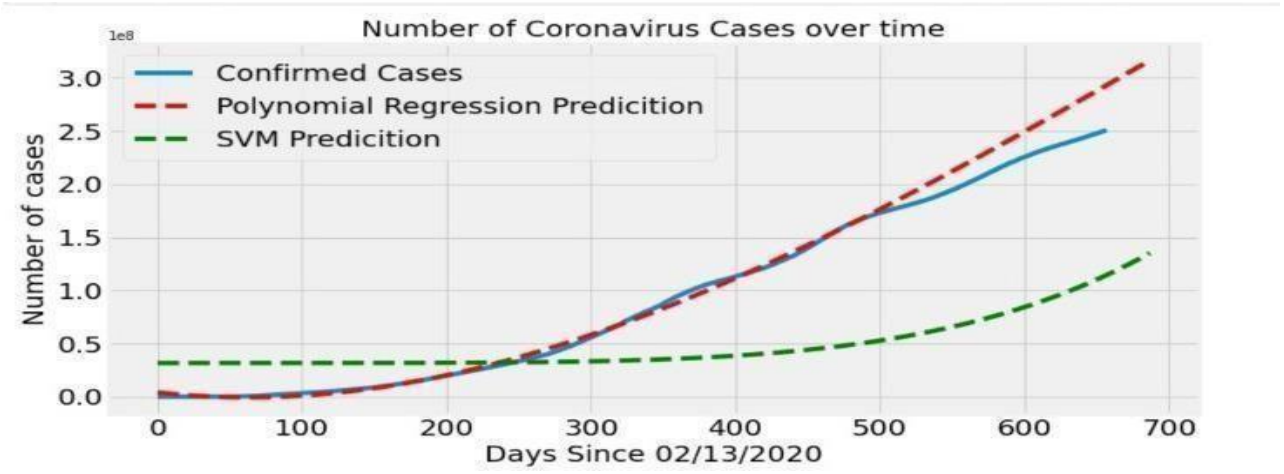


FIGURE 3 Prediction graph in comparison of confirmed cases and polynomial Regression prediction.

Description-In this graph on the x-axis the scale of 100 difference till 700 is taken with Heading Days since 02/13/2020 and on the y-axis scale of 0.5 till 3.0 difference variation Is taken this graph size is 12 on x by 6 on y, with heading number of cases is shown.

This graph represents the prediction rate of corona-virus confirmed cases in the world over time in comparison with the SVM and Polynomial Regression Prediction. This outcome has been analyzed with the data we have taken for the prediction. In this graph, the blue line shows the best fit line of confirmed cases with the help of data taken and the green lines show the prediction rate of SVM and the red line shows the Polynomial Regression Prediction. With the help of this Graph, we can easily analyze the best prediction rate of confirmed cases among these two algorithms (SVM, Polynomial Regression Prediction) in our Project with the help of this graph we can say that Polynomial Regression Prediction gives the best accuracy in prediction because it touches the blue line of confirmed cases more accurately.

Chapter 5: Conclusion

The Coronavirus Disease (COVID-19) Outbreak in India and outside India was studied in this report. COVID-19's transmission in India and worldwide is influenced by a variety of variables, including religious gatherings, which can result in a super-spreader of the virus. Any future gatherings like this will be harmful to people's health. Lockdown has shown to be an excellentmove in India, and it should be expanded much further due to this shutdown the ratio of deathscases is less in India. Many unknown criteria might lead to huge uncertainty in the projection, the prediction is meant to aid the government in making future decisions and dealing with the continuing corona-virus spread in India. As this study anticipated, greater attention should be paid to control measures such as increasing the testing rate, maintaining social distance, and avoiding needless gatherings, wearing a mask. If all of these measures are taken into account, the severity of COVID-19 in India may eventually be reduced to a manageable level. The findings of this study will aid in the planning of healthcare resources and the development of effective prediction of the virus and more knowledge about it. Moreover, we didn't rely on Accuracy only because the class was unbalanced; using accuracy will always result in predicting the majority class.COVID19's transmission in India and worldwide is influenced bya variety of variables that can function as a super spreader of the virus. Many unknown criteriamight lead to huge uncertainty in the projection; the prediction is meant to aid the governmentin making future decisions and dealing with the continuing corona-virus spread in India. The conclusion of this study will aid in the planning of healthcare resources and the development of effective prediction of virus and knowledge about it. The environment is under the control of the COVID-19. These prediction aims are to use ML models to examine the epidemic with data from GitHub. In conclusion, the Polynomial Regression (PR) method is best for predictionin this research. Future work would be trying other boosting algorithms such as Cat boost, LightGBM, and a neural network might be a good try for model improvement. In addition to this, a comparison between oversampling and under sampling will also be a good work to consider.

5.1 Further improvement

Collection of new relevant data by using surveys that include as many countries, as possible.

It's always good to keep improving work and it's a requirement. For this project, further, improvement is using different optimization techniques and working on more machine learning algorithms like Random Forest, KNN, etc., and analyzing which one gives more accurate accuracy than are available.

Can work more on variants of information gathering and differentiate in our forecasting. As Covid is not ended here in 2022, So many changes and many new things are still left to exploreand identify.

Train model with Light GBM and Cat boost as a new boosting algorithm. A further collection of more data and applying neural networks to that data.

Chapter 6: References

[1] International Journal of Computer Information Systems and Industrial Management Applications. ISSN 21507988 Volume 12 (2020).

[2] MIR Labs, www.mirlabs.net/ijcisim/index.html Machine Learning and Deep Learning Covid-19 Project-Final Report Project: Spread Visualization and Prediction of the Novel Coronavirus Disease COVID-19(2019).

[3] Sahas Ramanan A., & Kumar, N. (2020). The network structure of COVID19 spread and the lacuna in India's testing strategy. Available at SSRN3558548 from https://arxiv.org/ftp/arxiv/papers/2003/2003.09715 on 3rd April 2020.

[4] Machine learning-based prediction of COVID-19 diagnosis based on symptomsYazeed Zoabi(2020)

[5] Rajan Gupta, Saibal K. Pal and Gaurav Pandey. A Comprehensive Analysis of COVID19 Outbreak situation INDIA Access from WHO Coronavirus disease (COVID-19) Pandemic (2020).

[6] Janmajay Nayak, Vighnaraaj Naik, Paidi Dinesh,Kanithi Vakula, B. Kameswara Rao, Weiping Ding, Danilo Pelusi. Intelligent system for COVID-19 prognosis: a state-of-the-art survey(2020)

[7] Singh, R., & Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. Ar Xiv preprint arXiv: 2003.12055 on 4th April 2020.

[8] Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gillard. Predicting the growth and trend of the COVID-19 pandemic using ML and cloud computing (2020)

[9] short-Term Prediction of COVID19 Cases Using Machine Learning Models Md. Shahriar Satu Koushik Chandra Howlader, Mufti, Mahmud, Shamim Kaiser, Sheikh Mohammad Shafiullah, Julian M.W.(2020)

[10] Quinn, Salem A. Alyami and Mohammad Ali Moni, Citation: Satu, M.S.; how lade K.C.; Mahmud, M.; Kaiser, M.S.; Shaiful Islam, S.M.; Quinn, J.M.W.; Alyami, S.A.; Moni, M.A(2019)