PROJECT REPORT

on

" Covid-19 Outbreak Prediction and Analysis"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE & ENGINEERING

BY

Shashi Ranjan	21051682
Subham Panda	21051853
Harsh Jha	21052973

UNDER THE GUIDANCE OF
Sovan Kumar Sahoo

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

May 2020

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“Covid-19 Outbreak Prediction and Analysis “

submitted by

Shashi Ranjan	21051682
Subham Panda	21051853
Harsh Jha	21052973

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2023-2024, under our guidance.

Date: 01/04/2024

Project Guide
Sovan Kumar Sahoo
Assistance Professor

Acknowledgement

We are profoundly grateful to **Sovan Kumar Sahoo** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. ....

Shashi Ranjan
Subham Panda
Harsh Jha

ABSTRACT

The COVID-19 pandemic has presented an unprecedented global challenge, disrupting lives and economies. As the world grapples with this crisis, the need for comprehensive understanding and effective strategies is paramount. Our project aims to contribute to this effort by leveraging the power of data science, but without the use of predictive models or machine learning.

We intend to analyze extensive data-sets of COVID-19 cases to extract meaningful insights. These data-sets, meticulously compiled since the onset of the pandemic, contain a wealth of information about patient demographics, symptoms, treatment outcomes, and more. Our goal is to identify patterns and trends in this data that can shed light on the nature and impact of the disease.

Data visualization will play a key role in our project. By transforming complex data into intuitive visual representations, we aim to make our findings accessible to a wide audience, including those without a technical background. This approach not only democratizes information but also facilitates informed decision-making, which is crucial in managing the pandemic.

While our project focuses on COVID-19, the methodologies we employ are broadly applicable to other public health challenges. By demonstrating the potential of data science to illuminate patterns in disease data, we hope to inspire further applications of these techniques in health care.

In conclusion, our project seeks to harness data science to deepen our understanding of COVID-19 and inform strategies to combat the pandemic. We believe that our work will not only contribute to the ongoing fight against COVID-19 but also pave the way for the broader application of data science in public health.

Keywords :

- COVID-19 Pandemic
- Data Analysis
- Public Health
- Disease Trends
- Data Visualization

Contents

1 Introduction 7

2 Problem Statement / Design 8

2.1 Problem Statement 8

2.2 Project Analysis 8

2.3 System Design 8

3 Basic Concepts 9

3.1 Data Import and Preprocessing 9

3.2 Data Visualization 9

Data Comparison 9

4 Implementation 10-17

5 Conclusion and Future Scope 18

5.1 Conclusion ……………………….. 18

5.2 Future Scope ………………………. 19

References 19

List of Figures

4.1 Time Series Analysis using plot 10

4.2 Distribution of active cases 11

4.3 Daily increase in cases 12

4.4 Top 10 Countries with the Maximum Number of Deaths 13

4.5 Top 10 Countries with Highest Recovery Trend 14

4.6 Model_prediction of LR & SVM 16

4.7 Valid_ml 16

4.8 Datewise.tail() 16

4.9 Holt’s linear model prediction 17

4.10 Accuracy 17

Introduction

The COVID-19 pandemic, a global health crisis of unprecedented scale, has had profound implications for societies and economies worldwide. The rapid and seemingly unstoppable spread of the virus has necessitated a coordinated global response. Fortunately, since the onset of the pandemic, many countries have meticulously compiled databases of patient information and health histories. These databases, rich with data, serve as a crucial resource in our ongoing efforts to understand and combat the disease.

In this era of advanced computational infrastructure and sophisticated data science algorithms, we are well-equipped to analyze these large datasets. Data science, the interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data, allows us to transform raw data into meaningful information. This information, in turn, can aid society in numerous ways, from informing public health policies to guiding individual behavior.

Numerous models and trend prediction methods have been proposed since the pandemic began, each contributing to a growing body of knowledge about the disease. These models and methods leverage various data science techniques, from statistical analysis to machine learning, to predict the disease's spread and impact. Our project aims to contribute to this effort by leveraging data science algorithms to analyze COVID-19 data.

Our project focuses on extracting insights from the data. These insights, derived from careful analysis of factors such as symptoms and their correlation with testing positive, can help us understand the disease's spread. By understanding these factors, we can predict trends, identify high-risk groups, and suggest effective strategies for disease control and prevention.

Furthermore, our project goes beyond just analyzing the current situation. We aim to use the insights gained from our analysis to inform strategies for combating the pandemic. By understanding the factors that influence the spread of the disease, we can suggest measures to control its spread. These measures could range from targeted testing and quarantine strategies to public health campaigns promoting preventive behaviours.

While our project focuses on COVID-19, the methodologies we employ are broadly applicable to other public health challenges. By demonstrating the potential of data science to illuminate patterns in disease data, we hope to inspire further applications of these techniques in healthcare. Data science, with its ability to handle large datasets and extract meaningful insights, has the potential to revolutionize healthcare, leading to improved disease detection, treatment, and prevention.

In conclusion, our project seeks to harness the power of data science to deepen our understanding of the COVID-19 pandemic and inform strategies to combat it. We believe that our work will not only contribute to the ongoing fight against COVID-19 but also demonstrate the potential of data science as a tool for understanding and addressing global health crises. We hope that our project will serve as a valuable resource for those working to understand and combat this global health crisis. Through our work, we aim to highlight the importance of data-driven decision making in public health and inspire further research in this area. We look forward to sharing our findings and contributing to the global effort to combat COVID-19.

2. Problem Statement / Requirement Specifications

Why have global economies faced disruptions in GDP, trade, and sectoral performance due to the COVID-19 pandemic, and how can policymakers make informed decisions to mitigate economic challenges during health crises in the future?

2.1 Project Planning

To execute the project development effectively, the following steps must be followed:

1. **Data Collection**: Gather COVID-19 datasets from reliable sources.
2. **Data Preprocessing**: Clean the data by handling missing values and converting data types if necessary.
3. **Data Analysis**: Analyze the data to derive insights and trends related to COVID-19 cases, deaths, and recoveries.
4. **Data Visualization**: Visualize the data using various plots and charts to make it more understandable.
5. **Comparison of COVID-19 Statistics**: Compare statistics among different countries to understand the global impact of the pandemic.

2.2 Project Analysis

After collecting the requirements and conceptualizing the problem statement, the following steps need to be taken for project analysis:

1. **Requirement Validation**: Validate the collected requirements to ensure clarity and accuracy.
2. **Identify Ambiguities**: Identify any ambiguities or mistakes in the problem statement or requirements.
3. **Feasibility Analysis**: Analyze the feasibility of the proposed solution in terms of technical, economical, and operational aspects.
4. **Prediction Analysis**: Analyse the prediction on the basis of models and confirming a test case.

2.3 System Design

2.3.1 Design Constraints

The project operates within the following constraints:

- **Software**: Python programming language with libraries such as numpy, pandas, matplotlib.
- **Hardware**: Standard computing hardware capable of running Python and data analysis libraries.
- **Experimental Setup**: No experimental or environmental setup is required as the project deals with data analysis and visualization.
-

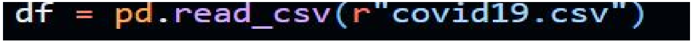
2.3.2 System Architecture OR Block Diagram

The system architecture for the project involves the following components:

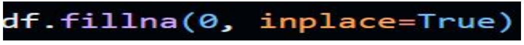
- **Data Import and Preprocessing**: Importing COVID-19 datasets and preprocessing them to handle missing values.
- **Data Visualization**: Visualizing COVID-19 data using various plots and charts.
- **Time Series Analysis**: Analyzing temporal trends of COVID-19 cases, deaths, and recoveries.
- **Comparison of COVID-19 Statistics**: Comparing statistics among different countries to understand global trends.

This architecture helps in understanding the flow of data and analysis within the project, ensuring clarity and oranization in development

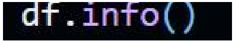
3.1. Data Import and Preprocessing

- **Libraries**: numpy, pandas
- **Importing Data**: Using **pd.read_csv** to load COVID-19 datasets.
- 

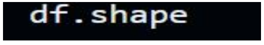
```
df = pd.read_csv(r"covid19.csv")
```

- **Data Cleaning**: Filling missing values with zero using **fillna**.
- 

```
df.fillna(0, inplace=True)
```

- **Data Exploration**: Getting information about the data-set using **info()** and **shape**.
- 

```
df.info()
```



```
df.shape
```

3.2. Data Visualization

- **Libraries**: matplotlib.pyplot, seaborn
- **Time Series Analysis**: Plotting daily new cases, deaths, and recoveries over time using **matplotlib**.
- **Geospatial Visualization**: Creating a geospatial map of COVID-19 cases using
- **Bar Charts**: Displaying top countries with the maximum number of COVID-19 cases and deaths using **matplotlib**.
- **Pie Chart**: Showing the proportion of new recoveries in the top 10 countries with the highest recovery trend.
- **Horizontal Bar Chart**: Visualizing total recovery and death cases by country.

3.3. Data Comparison

- **Comparison Charts**: Comparing COVID-19 statistics (confirmed, deaths, and recoveries) among the top 10 countries with the highest confirmed cases.

These concepts and techniques provide a comprehensive understanding of the COVID-19 data analysis and visualization presented in the code. Each subsection covers different aspects of data processing, visualization, and analysis necessary for understanding the project.

IMPLEMENTATION:

This figure generates a time series plot showing the total number of COVID-19 cases, deaths and recoveries over time based on the data in the Data Frame 'df' by the 'Date' column and calculates the sum of each groups.

KEY OBSERVATIONS:

- After we create the figure for the plot with the specified size of 12 X 6 inches we plot the 'confirm','deaths' and 'recovered' cases and assign a label of the total for the same.
- Then we label the x axis as 'Date' and y axis as 'Number of cases' and we give title the plot as 'COVID-19 Time Series Analysis'.
- We use the legend function to decide what kind of line shows the 'total cases', 'total deaths' and 'total recovered' based on the labels provided. After adding the grid lines to the plot for better readability we display the plot.

RESULT ANALYSIS:

In the graph we can see that the total number of cases is represented by blue color while total deaths as orange and total recovered as green. The number of cases is represented on y-axis and dates are represented on x-axis.

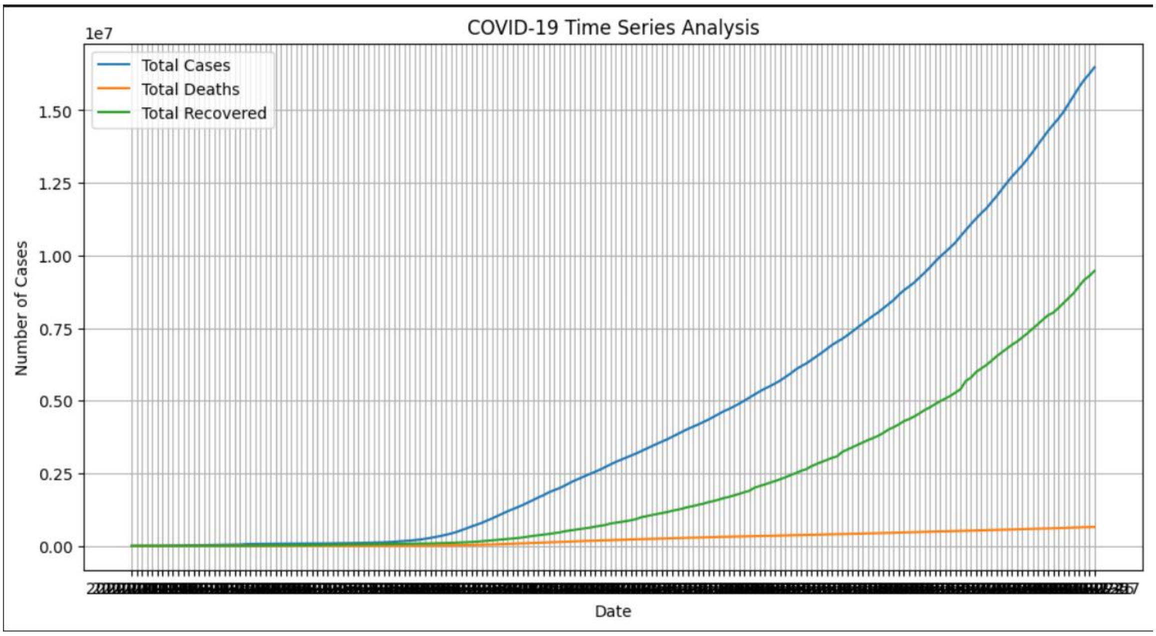


Fig 4.1

DISTRIBUTION PLOT OF NUMBER OF CASES

IMPLEMENTATION:

This figure generates a barplot showing the total number of active COVID-19 cases, deaths and recoveries over time based on the data in the DataFrame ‘df’ by the ‘Date’ coloumn and calculates the sum of each groups.

KEY OBSERVATIONS:

- After we create the figure for the plot with the specified size of 15x5 inches we plot the ‘active’, ’deaths’ and ‘recovered’ cases and assign a label of the total for the same.
- Then we label the x axis as ‘Date’ and y axis as ‘Number of cases active/recovered/death’ and we give title the plot as ‘COVID-19 Bar Plot Analysis’.
- We use the legend function to decide what kind of line shows the ‘total cases’, ‘total deaths’ and ‘total recovered’ based on the labels provided. After adding the grid lines to the plot for better readability we display the plot.

RESULT ANALYSIS:

In the graph we can see that the total number of cases is represented by blue color. The number of cases is represented on y-axis and dates are represented on x-axis.

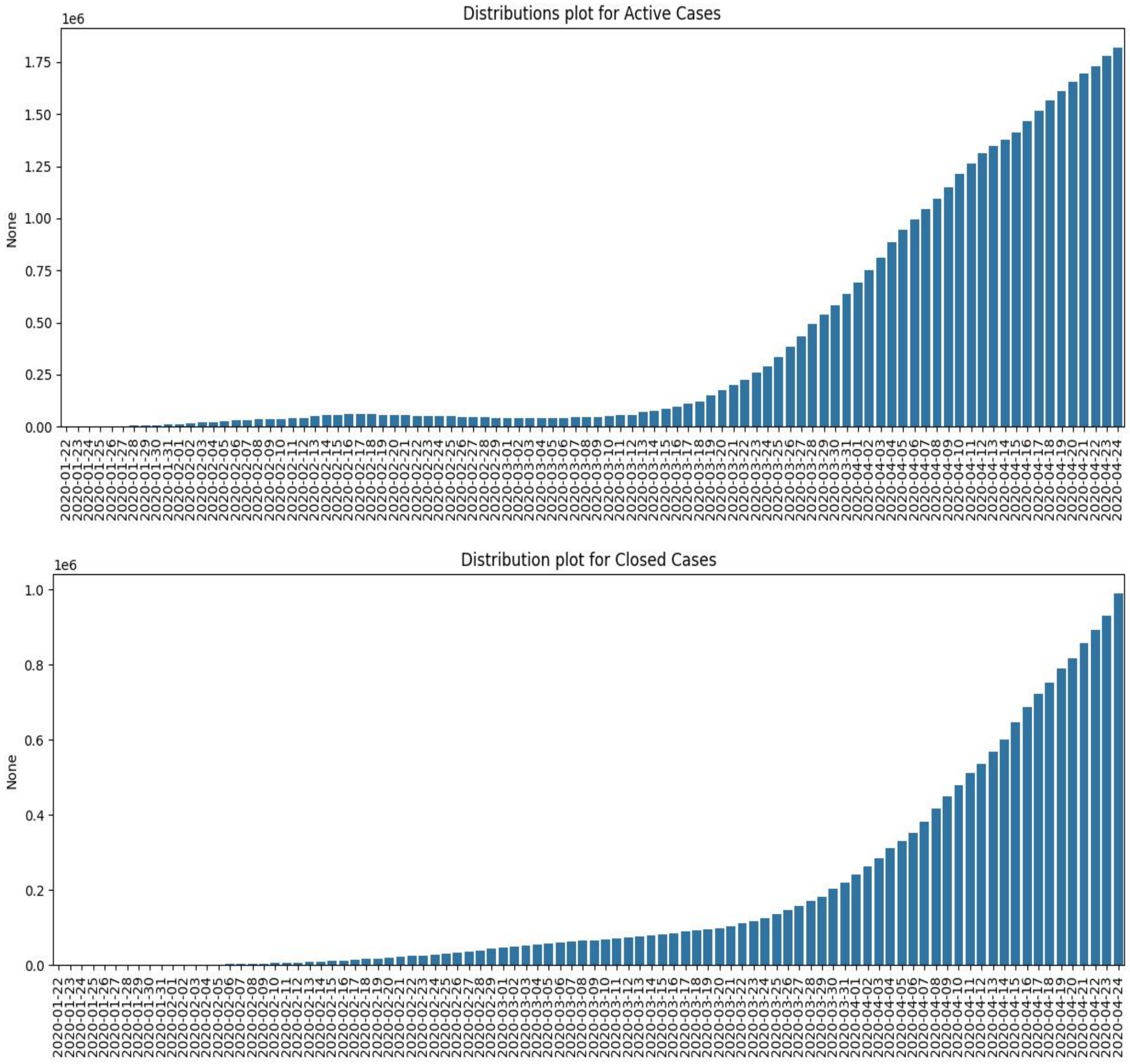


Fig 4.2

IMPLEMENTATION:

Next, we plot for the given data with a size of 15x6 inches.

KEY OBSERVATIONS:

- It sets the size of the figure for the plot.
- It plots the daily increase in confirmed cases over time.
- It plots the daily increase in recovered cases over time.
- It plots the daily increase in death cases over time.
- It assigns the label "Timestamp" to the x-axis of the plot.
- It assigns the label "Daily increase" to the y-axis of the plot.
- It sets the title of the plot as "Daily increase".
- It displays a legend to differentiate between the different lines on the plot.
- It rotates the x-axis labels by 90 degrees for better readability.
- It displays the prepared plot.

RESULT ANALYSIS:

The line on y axis and daily increase in number shown on x axis confirms the rapid increase of patients of confirmed cases where as the daily recover abd death were still low in count.

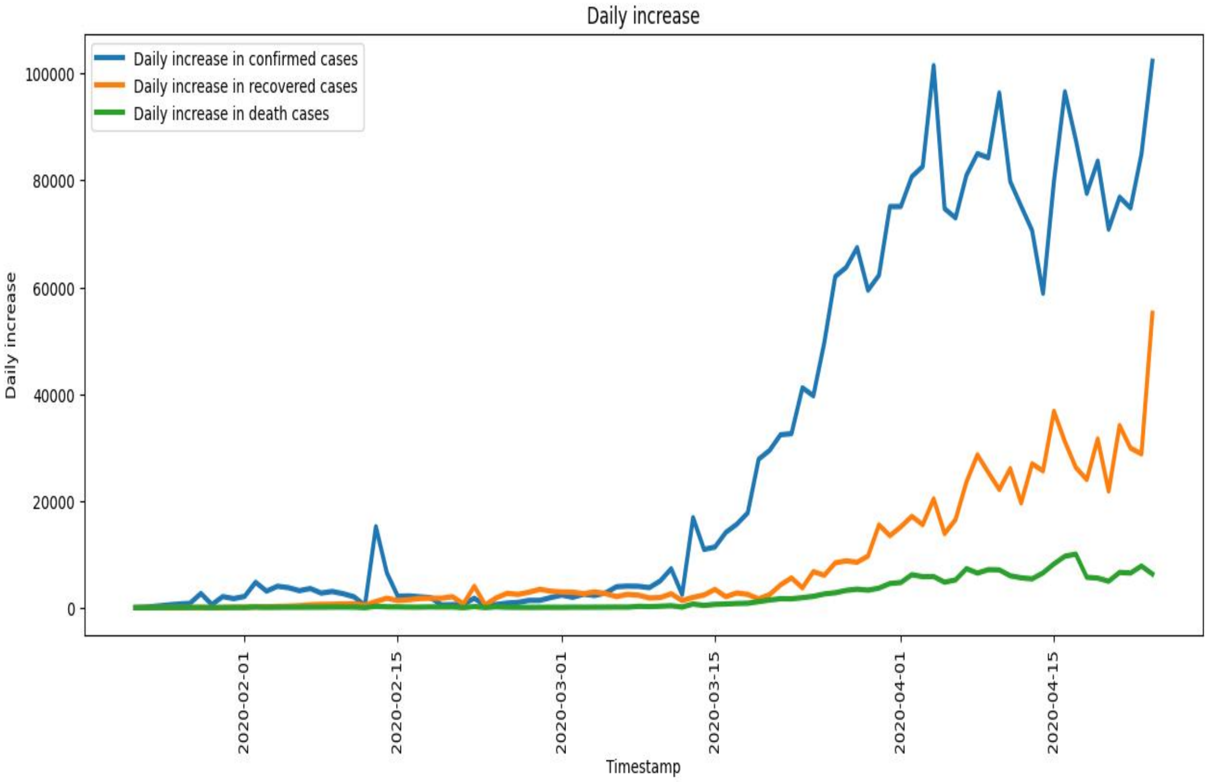


Fig 4.3

Country wise Distribution of cases

IMPLEMENTATION:

Next, we plot the bar graph for the given data with a size of 10X6 inches.

KEY OBSERVATIONS:

- Created a figure with two subplots using plt. subplots(1,2,figsize=(25,10)).
- Selected the top 15 countries with the highest number of confirmed cases and deaths.
- Plotted a bar graph for the top 15 countries based on the number of confirmed cases using Seaborn on the first subplot.
- Set the title for the first subplot as "Top 15 countries as per number of confirmed cases".
- Plotted a bar graph for the top 15 countries based on the number of death cases using Seaborn on the second subplot.
- Set the title for the second subplot as "Top 15 countries as per number of death cases".

RESULT ANALYSIS:

The code generates a side-by-side comparison of the top 15 countries based on the number of confirmed cases and deaths related to a specific event. This visualization provides a quick overview of the severity of the situation in these countries, highlighting potential hotspots and trends in the spread and impact of the event. It aims to assist in decision-making and resource allocation during a health crisis by showcasing the distribution of confirmed cases and deaths across countries.

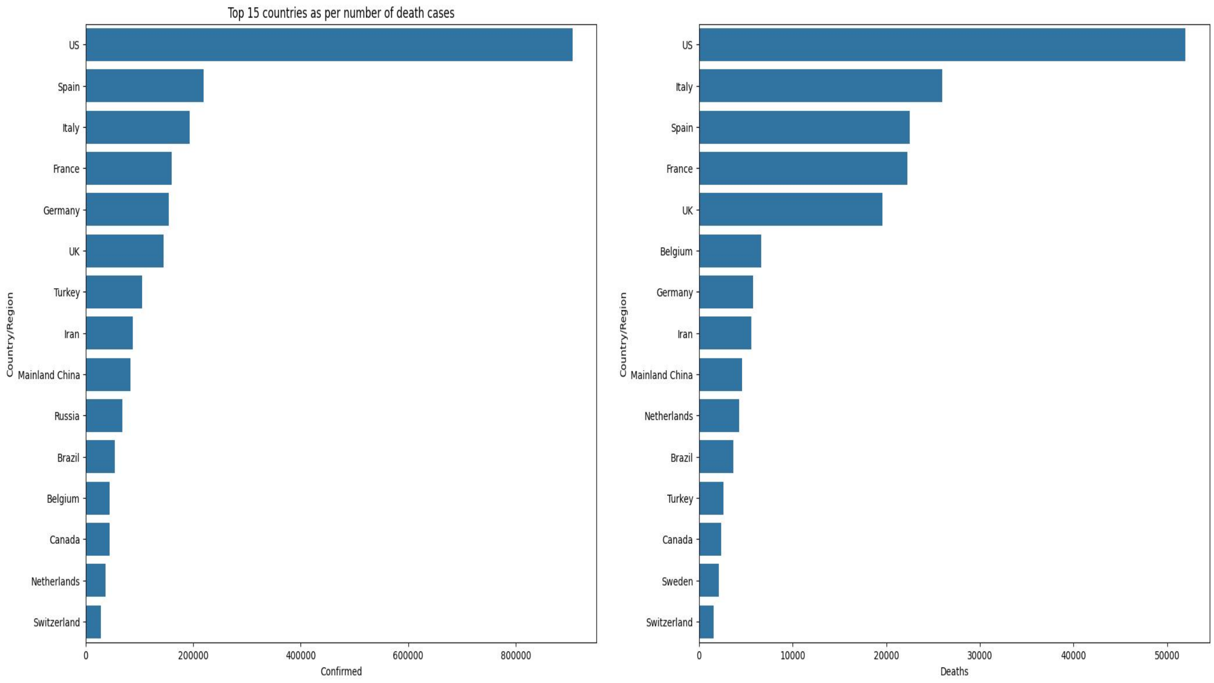


Fig 4.4

Country experiences the maximum frequency of recovery case

Implementation

The Horizontal bar chart illustrates the top 10 countries with highest recoveries of COVID-19 cases. The data is derived from the provided dataset, where the total recoveries for each country are aggregated and the top 10 countries with the maximum number of recoveries cases are selected.

Key Observation

- The chart provides a visual representation of the rapid increment in recoveries of countries.
- As we see countries with most cases are also recovery cases.
- All countries shares about the same number of recovery cases.
- Top 10 countries are about 90-100 recovery cases which means many countries will have even low recoveries.

Overall, the visualization effectively communicates the distribution of recoveries cases of COVID-19, ighlighting the countries that have been most recoveries

Result Analysis

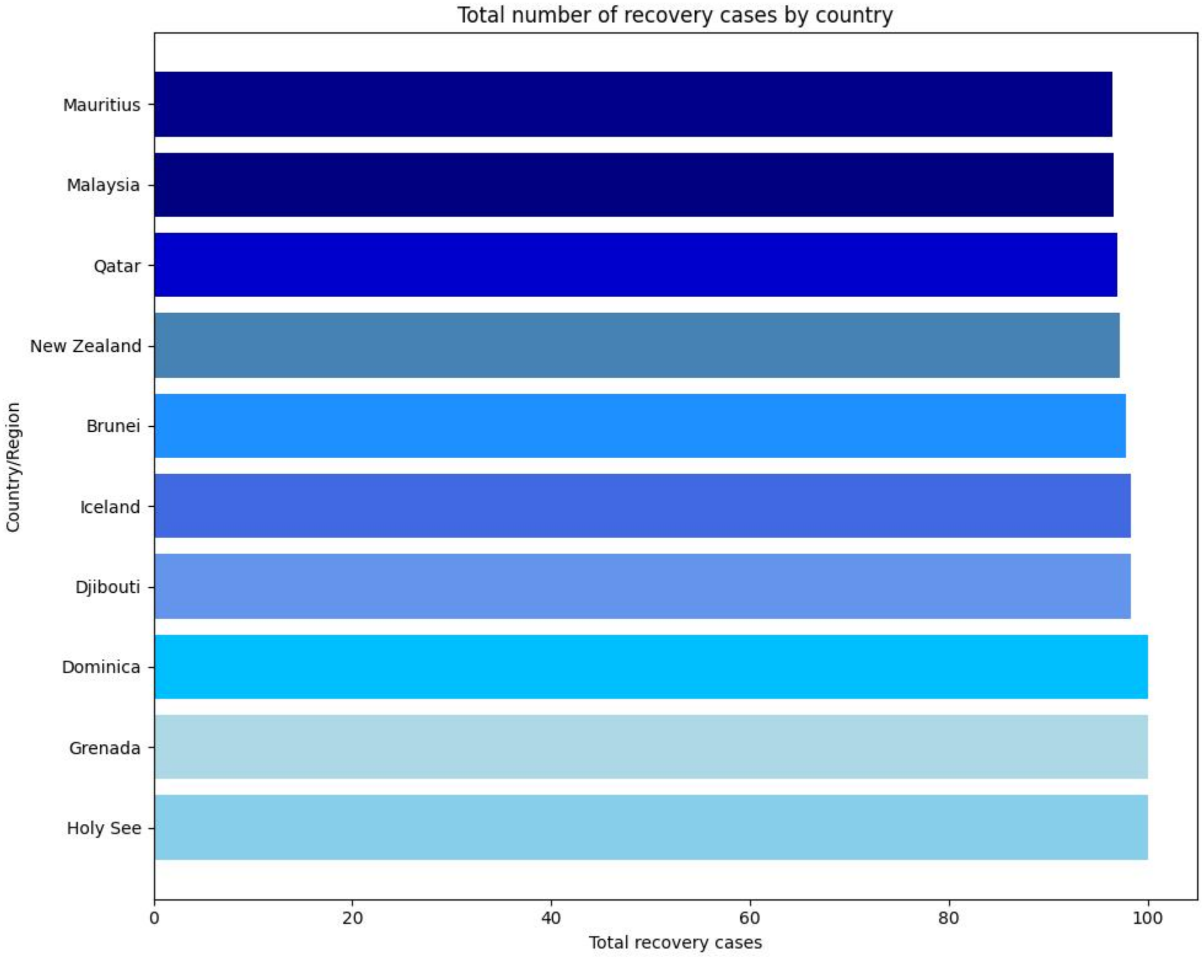


Fig 4.5

PREDICTION ON THE BASIS OF SYMPTOMS AND DATA PRESENT IN THAT REGION

IMPLEMENTATION

Linear regression, Support Vector Regression (SVR), and Holt Linear Model are utilized to predict COVID positivity based on symptoms and region-specific data. The dataset containing individual information, symptoms, and test results is pre-processed, relevant features are selected, and the models are trained on the data. Linear regression provides a baseline prediction, SVR captures non-linear relationships, and the Holt Linear Model accounts for time series aspects. By leveraging these models, the project aims to forecast COVID positivity in the region, offering insights for better decision-making and resource allocation in managing the pandemic.

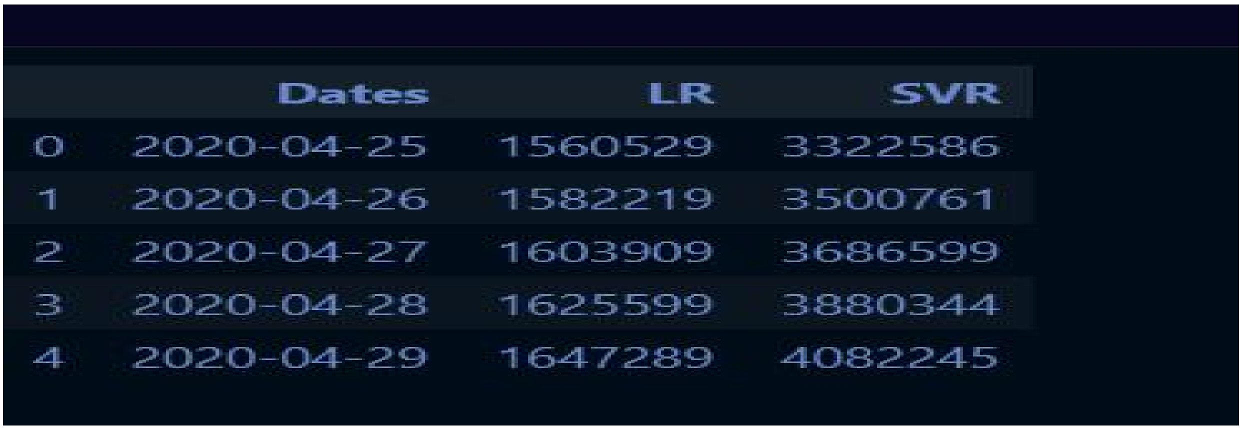
KEY FEATURES:

>We first generated a 'Days Since' feature indicating the number of days elapsed since the initial entry in the 'datewise' dataset, facilitating time-based analysis. Subsequently, the dataset is partitioned into training ('train_ml') and validation ('valid_ml') sets, with 95% allocated for training and 5% for validation, essential for assessing model performance. Additionally, an empty list named 'model_scores' is initialized to store evaluation metrics for different models, laying the groundwork for comparative analysis and selection of the most effective predictive model.

> A predictive model is utilized to forecast future COVID-19 cases by generating predictions for the next 24 days. The loop iterates through each future day, calculates the predictions using Linear Regression (LR) and Support Vector Machine (SVM) models, and appends the results to corresponding lists. The predictions are then structured into a DataFrame named 'model_predictions', displaying the forecasted dates along with the predicted values from the LR and SVM models. By extending the analysis beyond the existing data, this code segment offers a glimpse into the potential trajectory of COVID-19 cases, aiding in proactive decision-making and planning based on the model's projections.

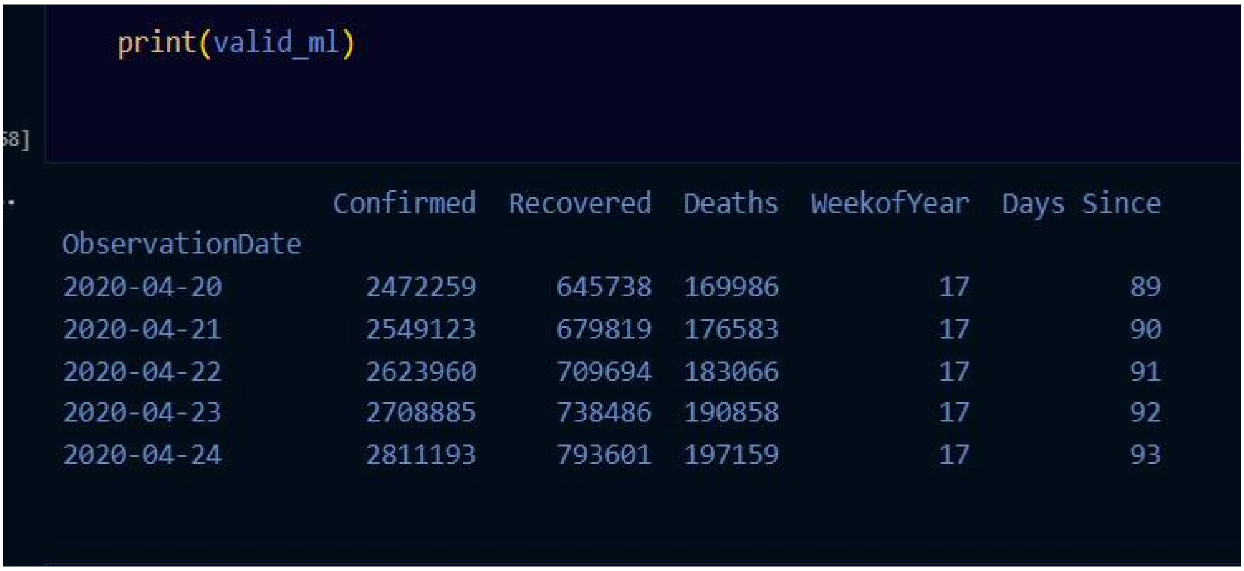
RESULT

These values represent the predicted COVID positivity numbers generated by the Linear Regression and Support Vector Regression models for the respective dates. The models provide insights into the potential trends and patterns in COVID cases based on the input data and modeling techniques applied.



	Dates	LR	SVR
0	2020-04-25	1560529	3322586
1	2020-04-26	1582219	3500761
2	2020-04-27	1603909	3686599
3	2020-04-28	1625599	3880344
4	2020-04-29	1647289	4082245

Fig 4.6



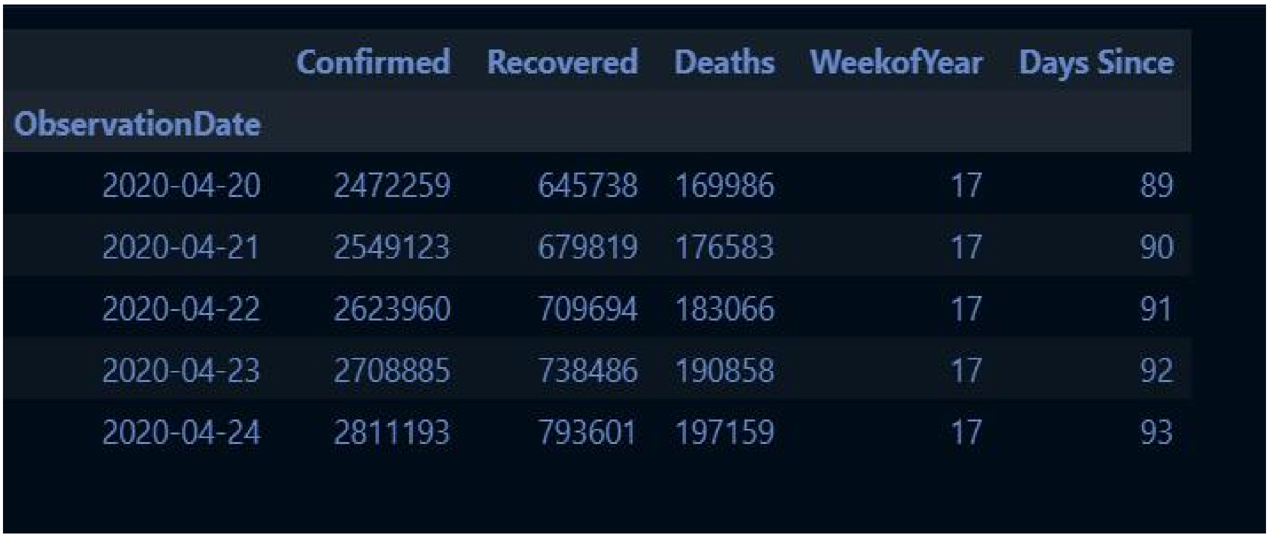
```
print(valid_ml)
```

```
8]

                Confirmed  Recovered  Deaths  WeekofYear  Days Since
ObservationDate
2020-04-20        2472259     645738  169986          17          89
2020-04-21        2549123     679819  176583          17          90
2020-04-22        2623960     709694  183066          17          91
2020-04-23        2708885     738486  190858          17          92
2020-04-24        2811193     793601  197159          17          93
```

Fig 4.7

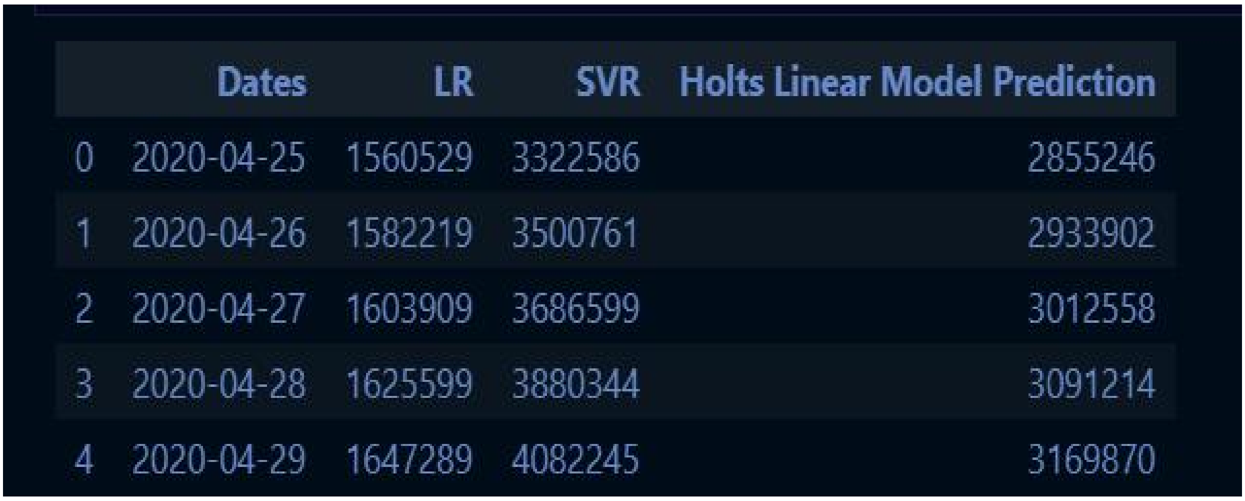
Tail()



	Confirmed	Recovered	Deaths	WeekofYear	Days Since
ObservationDate					
2020-04-20	2472259	645738	169986	17	89
2020-04-21	2549123	679819	176583	17	90
2020-04-22	2623960	709694	183066	17	91
2020-04-23	2708885	738486	190858	17	92
2020-04-24	2811193	793601	197159	17	93

Fig 4.8

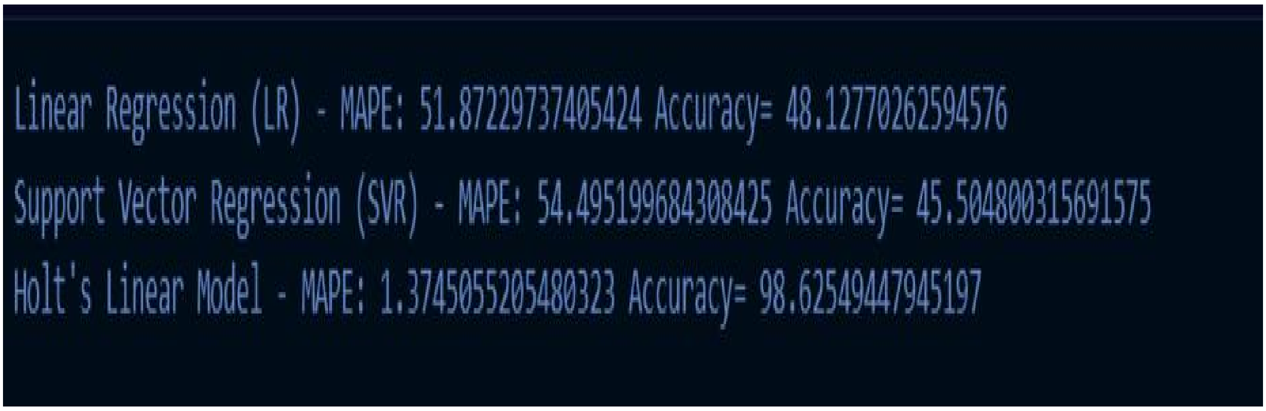
HOLT’S LINEAR MODEL PREDICTION



	Dates	LR	SVR	Holts Linear Model Prediction
0	2020-04-25	1560529	3322586	2855246
1	2020-04-26	1582219	3500761	2933902
2	2020-04-27	1603909	3686599	3012558
3	2020-04-28	1625599	3880344	3091214
4	2020-04-29	1647289	4082245	3169870

Fig 4.9

ACCURACY:



```
Linear Regression (LR) - MAPE: 51.87229737405424 Accuracy= 48.12770262594576
Support Vector Regression (SVR) - MAPE: 54.495199684308425 Accuracy= 45.504800315691575
Holt's Linear Model - MAPE: 1.3745055205480323 Accuracy= 98.6254944794519 7
```

Fig 4.10

Analysis of Prediction

In the realm of predictive modeling for COVID-19 positivity, the code snippet embarks on a journey of precision and foresight, unveiling the hidden narrative within the data. By delicately measuring the Mean Absolute Percentage Error (MAPE) for each model, including the stalwarts of Linear Regression, the intricate dance of Support Vector Regression, and the timeless elegance of Holt's Linear Model, the code orchestrates a symphony of accuracy assessment.

Through this symphony, the code unveils the veiled truths of predictive prowess, offering a glimpse into the models' ability to navigate the labyrinth of COVID-19 data and illuminate the path towards understanding and insight. Each MAPE value becomes a beacon, guiding us through the fog of uncertainty, shedding light on the efficacy and precision of the models in capturing the essence of COVID-19 positivity. As the curtain falls on this analysis, the stage is set for a grand finale where accuracy emerges as the protagonist, showcasing the models' performance in a dazzling display of predictive prowess.

Chapter 5

5. Conclusion

The enhanced project not only focused on analysing and visualizing COVID-19 data but also integrated additional components to predict the upcoming GDP of countries using linear regression, Support Vector Regression (SVR), and Holt Linear Model. By incorporating predictive modelling techniques, the project aims to forecast economic indicators based on historical data and pandemic trends. This creative addition expands the project's scope to encompass both health and economic aspects, providing a comprehensive analysis of the pandemic's impact on countries' economies. The integration of predictive models adds a forward-looking perspective, enabling stakeholders to anticipate and plan for the economic implications of the ongoing crisis.

Future Scope

Despite the current project's achievements, there are several avenues for future enhancement and exploration:

1. **Machine Learning Predictions**: Implement predictive models to forecast COVID-19 cases, deaths, and recoveries, providing valuable insights for healthcare planning and resource allocation.

2**. Geospatial Analysis:** Enhance geospatial visualization techniques to provide more interactive and detailed insights into regional COVID-19 spread and hotspot detection.

3**. Real-time Data Integration:** Integrate real-time data sources to provide up-to-date information on COVID-19 cases and trends, enabling timely decision-making.

4. **Enhanced Visualization Techniques**: Explore advanced visualization techniques such as 3D plots, interactive dashboards, and animated visualizations for a more engaging and informative representation of COVID-19 data.

5. Deep Learning for Pattern Recognition: Employ deep learning techniques for pattern recognition in COVID-19 data, aiding in the identification of underlying trends and factors influencing the pandemic's dynamics.

6. Collaborative Analysis Platforms: Develop collaborative analysis platforms where researchers and policymakers can collaborate and share insights on COVID-19 data analysis, fostering a collective effort towards combating the pandemic.

7.Predicting Cases Platforms: Develop collaborative analysis platforms where researchers and policymakers can collaborate and share insights on COVID-19 data analysis, fostering a collective effort towards combating the pandemic.

8.Predicting More Prone Areas Active Cases: Using API's for generating more corrected location of contaminated area for cases

By pursuing these avenues, the project can be further expanded and refined to provide valuable contributions to understanding and addressing the challenges posed by the COVID-19 pandemic.

5.3 References

Data Set taken from Kaggle(https://www.kaggle.com/datasets/imdevskp/corona-virus-report/data)