COVID-19: Applying Data-Driven Methods to Analyze, Forecast Spread and Resource Requirements

Jethru Varshik
21110089
Sujith
21110100
Rudhreshwar
21110172

Abstract

The world has been profoundly affected by the COVID-19 pandemic in ways never seen before. To comprehend and lessen its impact, data analysts globally have been actively working with pandemic-related data. This situation has underscored the significance of data analysis and forecasting in pinpointing potential hotspots and determining resource requirements. Despite the abundance of data generated from diverse sources during the COVID-19 pandemic, many vital questions about its transmission, potential hotspots, and effects are still unresolved. This report intends to delve into various COVID-19 data sources, including Kaggle, CoViD19-India, WHO, AnalyticsIndiaMag, the National Institute of Health, and Johns Hopkins University. It will explore potential inquiries that data analysts might pose about the COVID-19 data, such as pinpointing specific irregularities in the reported data, discovering relationships between case growth and other health indicators at the district or block level, forecasting the epidemic's expansion, and identifying regions that need to increase their hospital bed capacity imminently. In this project, our goal is to scrutinise COVID-19 data from various sources and employ data-driven methodologies to discern patterns, correlations, and potential determinants that influence the virus's spread and the necessity for resources.

Acknowledgement

We want to thank Prof. Anirban Dasgupta, Computer Science and Engineering, IIT Gandhinagar for his help and support throughout the project.

1. Introduction

The global crisis of COVID-19 has left an indelible mark, with a staggering tally of over 150 million cases and 3 million fatalities by April 2023. A wealth of COVID-19 information is accessible from diverse platforms such as Kaggle, CoViD19-India, WHO, AnalyticsIndiaMag, the National Institute of Health, and Johns Hopkins University. This data trove invites data analysts to delve into various aspects of the pandemic. They can scrutinise anomalies in the data, like the variances in case counts across different platforms or death report inconsistencies. Further, they can explore how case surges correlate with health metrics at local levels, including hospital bed availability per capita, comorbidity rates, or healthcare resource accessibility.

To encapsulate, the COVID-19 pandemic has carved out a critical role for data analysts to deploy their acumen on a pressing global challenge. Utilising the myriad data sources and employing advanced analytics, they can unearth patterns and trends that could shape policy-making and public health strategies.

Anticipating future COVID-19 waves, it's imperative to dissect historical data to pinpoint factors influencing viral transmission in specific locales, encompassing demographic analysis. Tackling these queries is pivotal for a thorough pandemic understanding, which is essential for crafting informed policies to curb and control its impact. Our endeavour aims to augment pandemic comprehension and guide policymakers on strategic measures to curb viral dissemination.

Our initiative zeroes in on pivotal tasks to dissect COVID-19 data, aiming to extract meaningful insights from the pandemic narrative.

Initially, we employed the **SIR (Susceptible-Infected-Removed)** model to forecast COVID-19's trajectory across various regions. This model was instrumental in deciphering the pandemic's progression and anticipating case trends.

Next, we utilised **Autoregression** models to analyse confirmed case trends and project future patterns. These models were key in revealing data trends and facilitating educated forecasts regarding the virus's proliferation.

Subsequently, we applied the **Z-score** method to detect anomalies within the reported figures. This step was crucial in pinpointing data deviations that might signal significant shifts in the pandemic's pattern.

In sum, our analysis underscored the value of data-centric methods in shedding light on the COVID-19 pandemic and shaping policy choices. Through diverse analytical tools, we pinpointed factors potentially affecting the virus's spread, projected forthcoming trends, and spotted irregularities in the data.

Covid-19 data analysis

Problem Statement

Given diverse COVID-19 datasets from multiple sources, employ data-driven techniques to uncover patterns, correlations, and influential factors affecting the virus's transmission and resource requirements.

Methodology

Overview:

The dataset primarily comprises COVID-19 data collected from Kaggle and the WHO website.

It covers various countries and includes the following characteristics:

- Confirmed cases
- Deaths
- Recovered individuals

Data is available for nearly 100 countries and their respective states/provinces.

India-Specific Data:

The dataset also contains state-wise information for India, including:

- Negative tested people
- Positive tested people
- Total samples

Data Analysis Approach:

1. **The dataset is split into two parts:**
   - Trained dataset: Used for model training
   - Testing dataset: Used for validation

   Some columns are dropped during prediction, focusing on relevant rows.
2. **The dataset related to Indian COVID-19 cases is used for:**
   - Autoregressive models (for prediction)
   - Z-score analysis (to identify outliers)
3. **Key features in this dataset include:**
   - Date
   - Time
   - State/Union Territory
   - Confirmed cases
   - Cured cases
   - Deaths
4. **Data Cleaning:**
   - Irrelevant columns have been removed to streamline the analysis.

2. Prediction of Covid 19 Cases using Autoregression Models and Finding Outliers

2.1 Dataset Preprocessing

1. **Managing Missing Values:**

   When encountering columns with missing values, we opted to either discard them to prevent potential distortion of our models or fill them with zeros to ensure seamless code execution.

2. **Data Type Conversion Post CSV Import:**

Post-import from CSV files, the majority of data appeared as strings. We converted these to appropriate data types, transforming date-related columns into Date-Time formats and numerical columns into their respective numeric data types.

3. **Synchronising Columns with Overlapping Information:**

   In the dataset pertaining to India, we noted that the data initially tracked 'Individuals Vaccinated' and later transitioned to recording 'Doses Administered'. Although different, these metrics were numerically similar during the early phase when total doses weren't tallied. This similarity was observed across various demographic segments such as gender and age groups (18-44, 45-60, and 60+ years). Consequently, we populated the 'Doses Administered' fields with the corresponding 'Individuals Vaccinated' data for consistency.

2.2 Prediction of Covid 19 Cases using Autoregression Models

2.2.1 Analytical Approach

Our approach involved utilising autoregressive models with a 5-day lag to project the trend of confirmed COVID-19 cases over the last 30, 90, and 150 days. We divided the dataset into a training set and a testing set, the latter representing the data for the duration we aimed to forecast. We also adjusted various hyperparameters within the models to enhance their predictive performance. The models we applied include:

2.2.2 ARIMA Model

The ARIMA model is a forecasting tool that combines Autoregression (AR), Integration (I), and Moving Average (MA) techniques. The AR component predicts future values based on past data points, while the MA component uses past forecast errors for prediction. The integration aspect involves differencing the data to achieve stationarity. ARIMA is suitable for both stationary and non-stationary datasets and is relatively straightforward, requiring estimation of only a few parameters. Despite its assumption of linearity and stationarity in data, ARIMA's utility is broadened through extensions like SARIMAX, which incorporate seasonal patterns and external variables.

2.2.3 SARIMAX Model

SARIMAX stands for Seasonal Autoregressive Integrated Moving Average with exogenous factors, an advanced version of ARIMA that accounts for seasonality and external influences. This model is adept at capturing both time-related and seasonal fluctuations, making it a versatile choice for forecasting across sectors like finance and public health. It includes parameters for both non-seasonal and seasonal components, and the inclusion of exogenous factors like economic or weather data can significantly refine forecasting precision. SARIMAX's adaptability and comprehensive pattern recognition make it a preferred model for complex forecasting tasks.

2.2.4 Auto Reg Model

The Auto Reg model is a forecasting method that relies on historical data points to predict future values. It regresses a variable's current value against its previous values up to a specified lag order (p). Auto Reg models are favoured for their ability to discern the data's autocorrelation and are effective for predicting time series data. However, they presuppose data stationarity and may falter with data exhibiting trends or seasonal shifts. For such datasets, ARIMA or SARIMAX models are recommended alternatives.

2.2.5 Manual Autoregressive Model with Linear Regression

In the manual Autoregressive (AR) model using linear regression, we treat lagged values of the dependent variable as distinct predictor variables within a multiple linear regression framework. This approach proves valuable when the order of the AR model remains relatively small. The process involves the following steps:

1. Order Selection: Choose the appropriate order for the AR model.
2. Feature Engineering: Create new columns for each lagged value up to the chosen order.

3. Data Splitting: Divide the dataset into training and test sets.
4. Model Fitting: Train a multiple linear regression model on the training set.
5. Prediction and Evaluation: Make predictions on the test set and assess model performance.

However, this approach becomes impractical for larger orders due to the proliferation of predictor variables. In such cases, specialised time series modelling techniques like ARIMA or SARIMAX may offer more efficiency.

2.2.6 Results and Conclusion

We summarise the results of various autoregressive models applied to the India COVID-19 dataset for predicting confirmed cases in Table 2. Our comparison relies on two metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Here are our key observations:

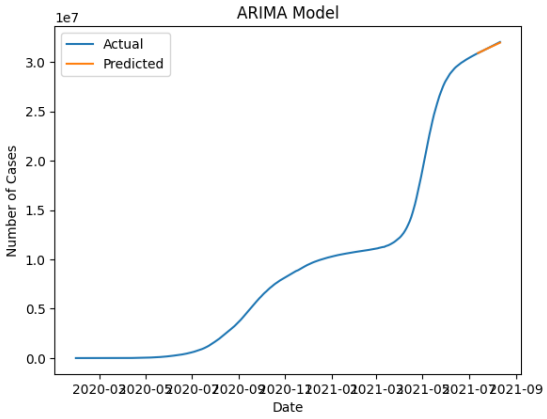
1. Auto Reg Model Performance: The Auto Reg model performed poorly. Its assumption of data stationarity (where mean and variance remain constant) led to suboptimal projections.
2. Auto Regression (AR) Model with Linear Regression: The AR model with linear regression excelled. Simplicity and straightforwardness likely contributed to its success, allowing it to capture internal correlations within the data structure. Unlike more complex models (such as SARIMAX), the Auto Reg with Linear Regression model required minimal data preprocessing and parameter tuning. Exogenous variables' impact may not be significant in this specific COVID-19 forecasting context, rendering SARIMAX less suitable.
3. ARIMA and SARIMAX Performance: ARIMA and SARIMAX performed well for 30 days but showed degradation over longer horizons. Suboptimal hyperparameters might have hindered their ability to capture data trends and underlying patterns.

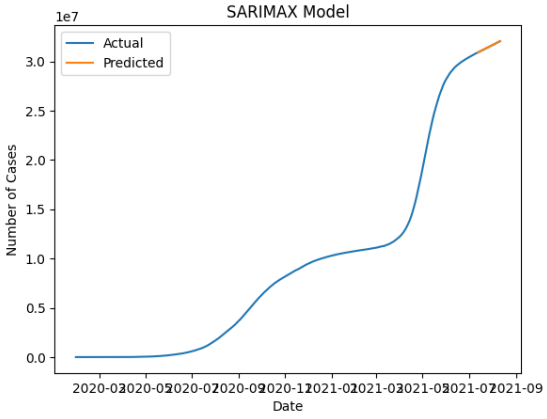
Overall, if we refine the model to allow it to capture the underlying patterns, the SARIMAX and ARIMA models may improve. However, given its relative simplicity and ability to capture the finer characteristics of the data, it appears that the manual Auto Reg model with Linear Regression is now the most appropriate for forecasting the COVID-19 situations. Additionally, as the mean and variance of the data fluctuate a lot, the data may contain a lot of outliers. As a result, several of the models failed because the real-world event was far more unpredictable than predicted.

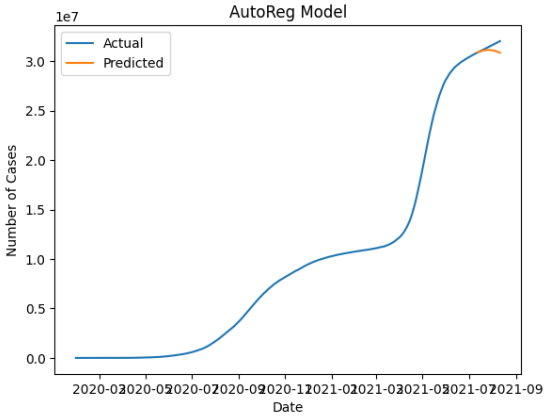
Model	No of Last Days Predicted	RMSE	MAE
Auto Regression (Lag=5) with Linear Regression	30	6206.154816153456	4831.723224785427
	90	8085.183857921982	6369.83252155962
	150	34855.125544396177	26868.801439922 8
ARIMA(Lag=5)	30	38100.387243781435	27906.30220860454
	90	7973045.223877883	6729851.00304631
	150	5860972.151271556	5149824.539301599
SARIMAX (Lag=5, Seasonal Lag=3 for No of periods=7)	30	8631.26648503542	5571.476575092723
	90	4944143.126863601	4212161.443631974
	150	7058622.049218587	6117415.422327464
Auto Reg (Lag=5)	30	562998.1663431099	427453.90194417787
	90	24848716.29951145	19168409.521238685
	150	13299784.971162418	11254562.003256382

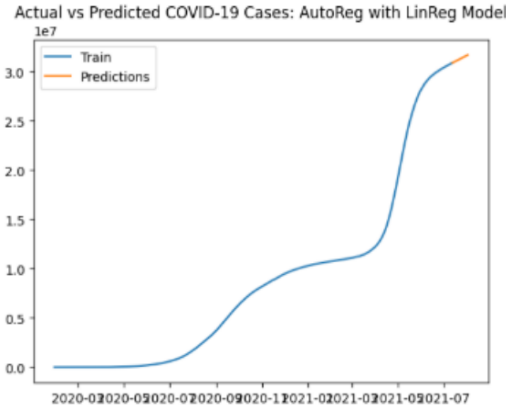
Table 1 Comparison of different models

No of Last days being predicted = 30

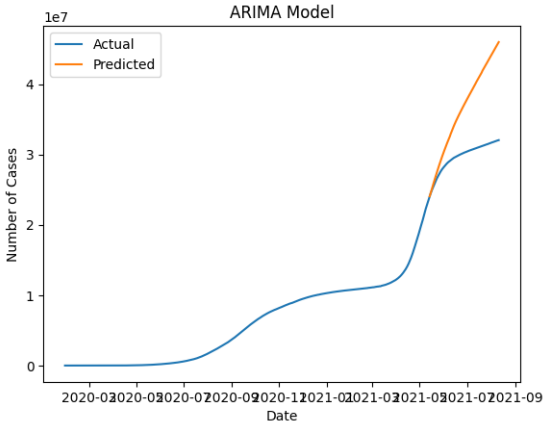


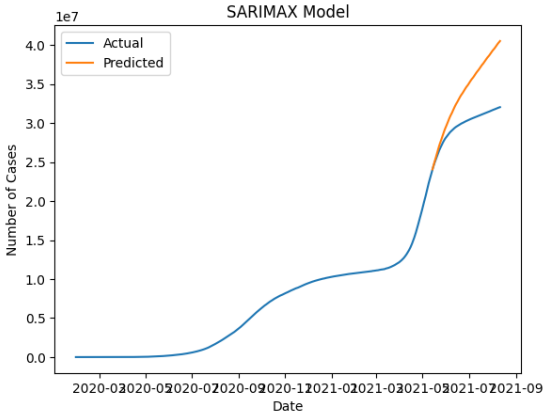


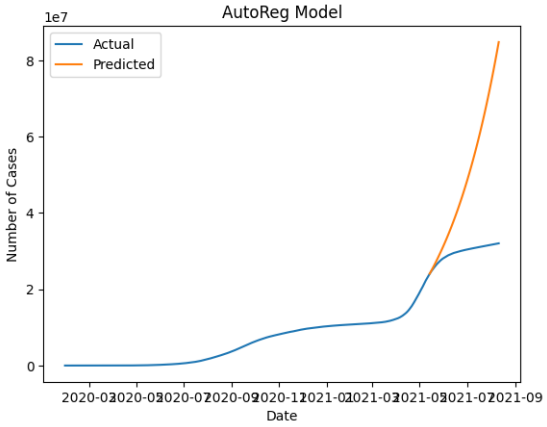


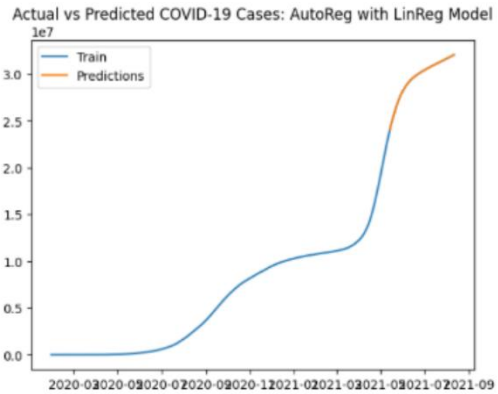


No of Last days being predicted = 90

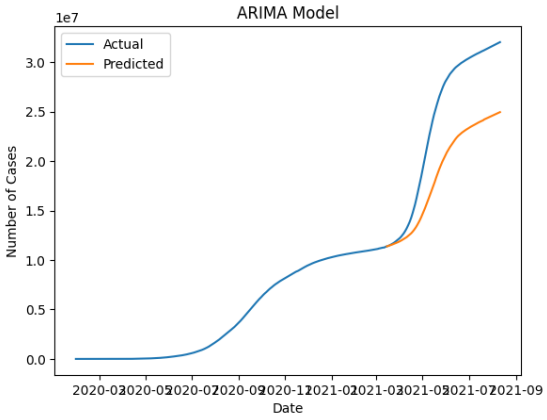


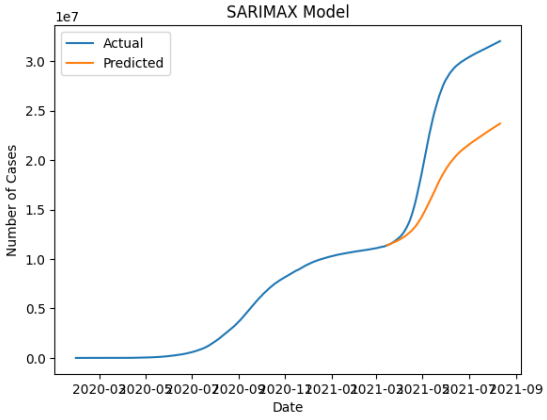


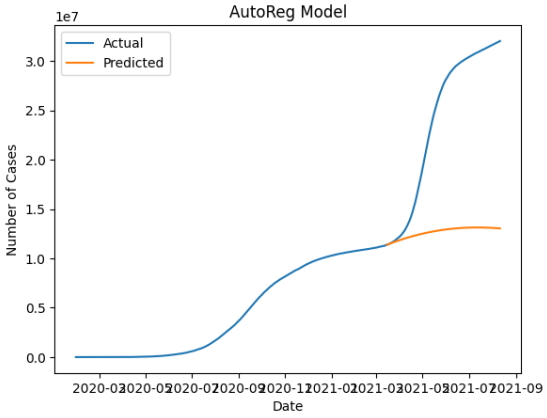


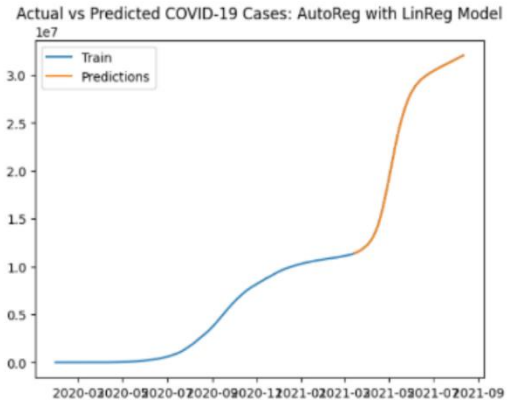


No of Last days being predicted = 150









2.3 Finding Outliers/Anomalies in the Covid Dataset

2.3.1 Methodology

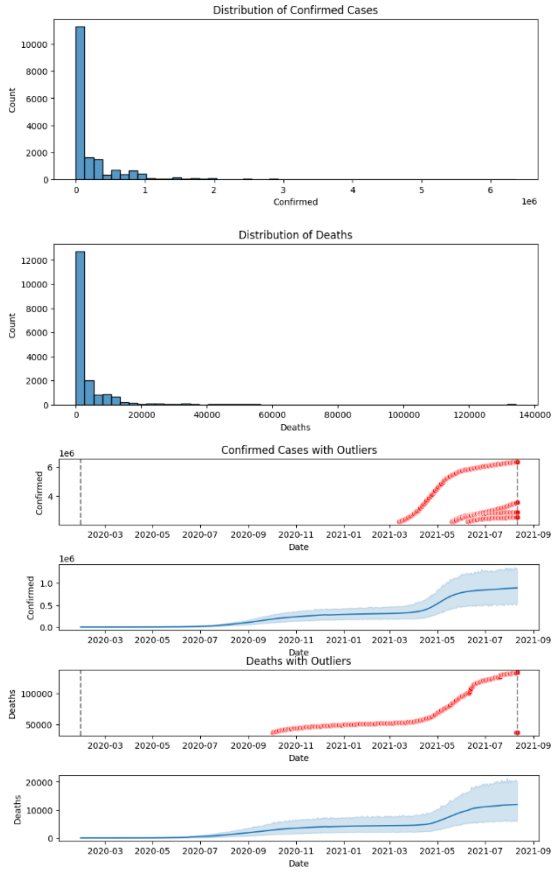
To analyse the confirmed cases and the deaths, we employed a few models and approaches. In order to improve the models' performance for a given set of parameters, we attempted to adjust a few of the hyperparameters that underlie some of the models. The following models were employed:

2.3.2 Z-Score

A statistical metric called the Z-score is frequently used to locate outliers in a dataset. It entails determining a data point's standard deviation from the dataset mean and then normalising it by dividing the difference by the standard deviation. This yields a standard score, or Z-score, that represents the number of standard deviations a data point deviates from the mean.

Since they deviate more than three standard deviations from the mean, data points with a Z-score of larger than three or fewer than three are typically regarded as outliers. However, depending on the features of the dataset or the demands of the particular challenge, the threshold for recognizing outliers can be changed. Z-score is a popular and easy-to-use technique for identifying outliers.
However, it may not be suitable for datasets with non-normal distributions or for datasets with a large number of dimensions, as it assumes that the data is normally distributed and independent.

2.3.3 Results and Conclusion



Z-Score Plots with Outliers

It seems that Z-score was able to find outliers in the later portion of the trend in the instance of COVID-19 confirmed cases and deaths. These anomalies may be a symptom of major shifts in the underlying patterns of the data, like abrupt increases or decreases in the number of cases or fatalities. Another possibility is that, in the later stages of the trend, the mean and variance have grown to the point where the data values relative to them are extremely large or small, creating outliers and signalling that the trend is becoming unpredictable.

The Z-score is a popular and easy-to-use technique for finding outliers in a dataset. It is especially helpful in spotting notable alterations to the data's underlying patterns. It's crucial to remember that the approach relies on the assumption that the data have a normal distribution, which may not always hold true in practical situations.

All things considered, the Z-score-identified outliers in the later stages of the COVID-19 confirmed and mortality trends may be helpful for additional research and analysis, and they might also provide insight for public health policies and initiatives. To obtain a more thorough comprehension of the data, it is crucial to take into account the method's limits and combine it with additional analytical techniques.

3. The SIR Model for Spread of Disease - The Differential Equation Model

3.1 Introduction

We identify the independent and dependent variables as the first step in the modelling process. The independent variable is time t, measured in days. We are considering two sets of dependent variables related to each other.

The first set of dependent variables counts people in each of the groups, each as a function of time:

S = S(t) is the number of susceptible individuals,
I = I(t) is the number of infected individuals, and
R = R(t) is the number of recovered individuals.

The fraction of the total population is represented by the second set of dependent variables in each of the three categories. So, if N is the total population, we have:

s(t) = S(t)/N the susceptible fraction of the population,
i(t) = I(t)/N the infected fraction of the population, and
r(t) = R(t)/N the recovered fraction of the population

The governing differential equations are as follows:
The SIR model is described by a system of three ordinary differential equations that govern the rate of change of each category (S, I, R) over time. The equations are:

dS/dt = -βSI
dI/dt = βSI - γI

$dR/dt = \gamma I$

Where:

S denotes the count of individuals who are susceptible to the infection.

I signifies the count of individuals who are currently infected.

R represents the count of individuals who have recovered from the infection.

β is the transmission rate, indicating how rapidly the disease spreads from infected to susceptible individuals.

γ is the recovery rate, indicating how rapidly infected individuals recover and become immune to the disease.

- The initial equation outlines the rate of change of the susceptible population, which diminishes as individuals become infected. The term -βSI signifies the rate at which susceptible individuals become infected.

- The subsequent equation outlines the rate of change of the infected population, which escalates as individuals become infected and diminishes as they recover. The term βSI signifies the rate at which susceptible individuals become infected, while the term γI signifies the rate at which infected individuals recover.

- The final equation outlines the rate of change of the recovered population, which escalates as individuals recover from the disease. The term γI signifies the rate at which infected individuals recover and become immune to the disease.

- The SIR model is a straightforward yet effective method to model the spread of infectious diseases in a population and has been utilised to inform public health policy and decision-making during outbreaks.

3.2 Assumptions

1. The population is constant - There is no migration, birth, or death of individuals in the population during the duration of the epidemic.

2. Homogeneous mixing - Every individual has an equal chance of coming into contact with every other individual in the population, regardless of location or behaviour.

3. Infectivity is constant - Every infected individual has the same level of infectiousness throughout the duration of the disease.

4. Recovery is permanent - Individuals who have recovered from the disease are immune and cannot be infected again.

5. No latent period - Individuals become infectious immediately after being infected.

6. No vaccination or intervention measures - The model assumes that there is no intervention, such as vaccination or treatment, that could modify the transmission dynamics of the disease.

3.3 Data Analysis Procedure:

1. Importing necessary libraries.

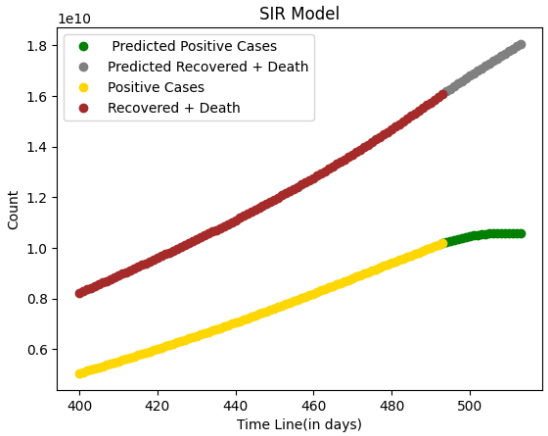
2. Preprocessing:

- Data is cleaned by dropping redundant columns like Province/State, Country/Region, Last Update, etc.
- Data is grouped by date and aggregated to sum the number of confirmed cases, deaths, and recoveries .
- The cumulative sum of confirmed cases, deaths, and recoveries is calculated.

3. Training and further working:

The SIR model is a compartmental framework that segregates the population into three categories: susceptible, infected, and recovered. It postulates that individuals can transition between these categories based on specific rates. We initially extract the confirmed, recovered, and death cases from the input data and compute the count of susceptible individuals. Subsequently, we calculate the transmission rate (beta) and the recovery rate (gamma) from the data. The reproduction number (R0) is computed as the quotient of beta and gamma.

The subsequent step involves employing a Ridge regression model to forecast the future transmission and recovery rates. We train two Ridge regression models on the transmission and recovery rates of the past 10 days and predict the rates for the upcoming 10 days. These predicted rates are then utilised to forecast the future count of susceptible, infected, and recovered cases using the SIR model.



Plot : the actual and predicted number of infected and recovered cases over time

The graph encompasses the actual data up to the present day and the forecasted data for the forthcoming 20 days. The forecasted data is derived using the anticipated transmission and recovery rates. The graph also illustrates the projected number of confirmed cases, which is the aggregate of predicted infected and recovered cases. We observe that the death count is rising approximately linearly and the confirmed cases are reaching a plateau. This could be in line with the SIR model, which presumes the population's homogeneity and its constant size. We can infer that as the number of infected individuals increases, the probability of others getting infected decreases, and those who were infected earlier are succumbing to the virus.

4. Conclusion

To summarise, our project's objective was to scrutinise COVID-19 data from a variety of sources and employ data-centric methodologies to discern patterns, correlations, and potential determinants that impact the propagation and resource requirements of the pandemic. We utilised a range of machine learning methods such as the SIR model, autoregression model, and Z-score to examine and forecast COVID-19 cases and trends that influence the virus's spread.

Our research findings can enhance the comprehension of the pandemic and provide valuable insights to policymakers regarding potential strategies to curb the virus's spread. We pinpointed potential determinants that impact the virus's propagation, including population density and demographic factors. Additionally, we employed machine learning methods to forecast future trends in COVID-19 cases and detect irregularities in the reported data.

The Auto Regression (AR) model coupled with linear regression exhibited excellent performance. ARIMA and SARIMAX models performed reasonably well for a duration of 30 days but their performance deteriorated subsequently.

In essence, our project underscores the potential of data-centric methodologies to guide policy decisions and manage the repercussions of the COVID-19 pandemic. As the pandemic continues to impact millions of individuals worldwide, our research can contribute to a more profound understanding of the virus and guide strategies to control its propagation and alleviate its effects.

Our git-hub repository link:
https://github.com/Codehack09876/COVID-19-ANALYSIS

References

https://github.com/AaronWard/covidify
https://github.com/tirthajyoti/Covid-19-analysis
https://github.com/CityOfLosAngeles/covid19-indicators
https://github.com/Yu-Group/covid19-severity-prediction
https://github.com/paulvangentcom/python_corona_simulation
https://github.com/rv20197/COVID-19-Analysis-and-Prediction-Using-AI
https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge