# **Capstone Project “Exploring Global COVID-19 Trends and Predicting New Cases in the USA” Report**

**1. Introduction**

The COVID-19 pandemic has profoundly impacted public health and the economy, necessitating forecasting and risk prediction models to help decision-making. Accurate forecasting of COVID-19 cases can assist policymakers, healthcare professionals, and researchers in optimizing resource allocation, implementing timely interventions, and mitigating the spread of the virus.

**Goal:** The primary goal of this project is to explore the COVID-19 trends around the globe and to develop a machine learning model to predict the daily new cases in the United States. This will involve:

* Time Series Forecasting: predicting the daily number of new cases over the next few months.

**2. Data Overview**

* Dataset: WHO COVID-19 dataset
* Source: [Kaggle](https://www.kaggle.com/datasets/abdoomoh/daily-covid-19-data-2020-2024)
* Time Period: Covers cases from early 2020 to 2023
* Target Variable: New COVID-19 cases reported daily
* Data Size: It consists of 57,840 entries and 8 columns.
* Data Preprocessing: Handled missing values, ensured stationarity and applied log transformation for stabilization

**3. Methodology**

### **Exploratory Data Analysis (EDA)**

We used EDA to understand the distribution and trends of COVID-19 cases and deaths across different regions and countries.

**Forecasting Models**

We used two-time series models to predict the future COVID-19 cases in the USA.

* **ARIMA (AutoRegressive Integrated Moving Average):** A widely used model for non-seasonal time series forecasting.
* **SARIMA (Seasonal ARIMA):** An extension of ARIMA that accounts for seasonality in the data.

**4. Model Evaluation**

The following evaluation metrics were used:

| **Metric** | **ARIMA** | **SARIMA** |
| --- | --- | --- |
| **MAE** (Mean Absolute Error) | 134,360 | 97,100 |
| **MSE** (Mean Squared Error) | 24.81 billion | 16.38 billion |
| **RMSE** (Root Mean Squared Error) | 157,527 | 128,000 |
| **AIC** (Akaike Information Criterion) | 3380.63 | 3462.81 |
| **BIC** (Bayesian Information Criterion) | 3397.55 | 3474.15 |

**5. Results and Analysis**

**EDA provides the following insights:**

* The EMRO and EURO regions show a higher frequency of case counts, whereas other regions exhibit a wider range of variation.
* The distribution of new deaths across regions is right-skewed, indicating many days with low deaths but occasional extreme spikes, suggesting outbreaks.
* The data reveals multiple peaks, likely corresponding to different COVID-19 waves, such as the Delta and Omicron variants.
* The EURO region had the highest number of cases at the beginning of 2022, while WPRO saw the highest cases at the end of 2022.
* The USA recorded the highest number of new cases at the start of 2022, and China had the highest cases by the end of the year.
* The pie chart of new cases showed the USA leading in infections among the top eight countries, with Japan having the lowest share.
* Analysis of death surges indicated about six major waves, with the highest peaks occurring in early 2021 and early 2022 due to insufficient intervention measures.
* The highest death toll was observed in the US, particularly during early 2021 and early 2022, whereas India experienced a massive surge in mid-2021 due to the Delta variant.

**Modeling provides the following insights:**

SARIMA **outperforms ARIMA** in terms of accuracy (lower MAE, MSE, and RMSE), suggesting that seasonality plays a significant role in COVID-19 case trends.

* ARIMA has a **lower AIC and BIC**, indicating a slightly simpler model, but at the cost of higher prediction errors.
* The **forecasted trends** from SARIMA align more closely with the actual future data, validating its suitability for COVID-19 forecasting.

**6. Visual Insights**

* The forecast plots indicate that SARIMA provides a more accurate representation of future trends, capturing both periodic fluctuations and general trends.
* ARIMA's forecasts deviate more significantly from actual data, particularly in high-variance periods.

**7. Recommendations**

* **SARIMA should be the preferred model** for deployment due to its improved accuracy.
* **Further improvements:** Adding external variables (e.g., vaccination rates, mobility data) using an SARIMAX model could enhance performance.
* **Deployment Strategy:** Automate model updates with real-time data to refine forecasts continuously

**8. Conclusion**

The SARIMA model is the best choice for forecasting COVID-19 cases due to its superior performance in handling seasonality. This model will help public health officials and policymakers make informed decisions regarding pandemic response and resource allocation.

**Link to access the Jupyter notebook**

[**https://github.com/preetikumar20/Capstone\_Covid19/blob/ac8c34bb6a10d9860434f838e5dfcb858895507f/prompt\_final.ipynb**](https://github.com/preetikumar20/Capstone_Covid19/blob/ac8c34bb6a10d9860434f838e5dfcb858895507f/prompt_final.ipynb)