1. **Introduction:**

The COVID-19 Data Repository by Johns Hopkins University (JHU) provides a comprehensive collection of COVID-19 data, updated daily at 6 AM UTC. It includes confirmed cases and deaths at the country level and US county level, along with various metadata. The data is available in two versions: the RAW version, which mirrors the original dataset, and the CONVENIENT version, which is organized for easier analysis. The CONVENIENT version includes daily changes in cases and deaths (instead of cumulative totals) and separates metadata into distinct files. The objective is to track the global and regional progression of COVID-19, monitor trends in cases and deaths, and predict future changes based on historical data. This dataset supports analyses like understanding virus spread, identifying high-risk areas, and forecasting future cases and mortality. It can be used to develop predictive models, assess the impact of interventions, and prepare for future healthcare needs. It aids in forecasting and planning for COVID-19’s ongoing impact.

**2. Dataset Overview:**

1. The dataset includes 5 explanatory variables (Province/State, Country/Region, Lat, Long, and the Date Columns) describing different aspects of COVID-19 trends.
2. Target: The objective is to track the global and regional progression of COVID-19, monitor trends in cases and deaths, and predict future changes based on historical data.

**Attribute Information:**

* Province/State: This refers to the administrative division or region within a country, like a state in the U.S. or a province in Canada, that indicates a geographical or political area.
* Country/Region: This indicates the country or broader geographical area (such as a continent or region) where a location is situated. For example, it could be "India," "United States," or "Asia."
* Lat: Latitude is a geographical coordinate that specifies the north-south position on Earth. It is measured in degrees, with values ranging from -90° (South Pole) to +90° (North Pole).
* Long: Longitude is a geographical coordinate that specifies the east-west position on Earth. It is measured in degrees, ranging from -180° to +180°.
* Date Columns: These refer to columns in a dataset that contain date-related information, such as the specific date of an event, transaction, or measurement. This feature is often used for time series analysis or tracking changes over time.

**3. Data** **Summary**:

Number of records: 266

Number of features: 248

**Data Types:**

1. Categorical Features - 2
2. Numerical Features- 246

**Null Values:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Null Values** | **Type** |
| Province/state | 185 | Categorical |

**4. Handling Missing Values:**

1. Identified missing values in column such as Province/State.
2. Strategies used: The missing values were not replaced because the column in question is not needed or relevant to the analysis, so it was considered unnecessary to fill those missing values.

**5. Data Cleaning and Preprocessing:**

* 1. Unpivot data and format the date
  2. Join the confirmed cases with the deaths statistics and recovered
  3. Save countries with covid cases
  4. Group by country

**6. Exploratory Data Analysis:**

**Correlation Analysis**

* Identified strong correlations between Confirmed and features like Deaths and Recovered.

**Data Visualization**

* COVID-19 Confirmed Cases in Germany
* Daily Confirmed COVID-19 Cases in Germany
* COVID-19 Time Series for Germany
* ARIMA Forecast for Deaths
* 7-day Rolling Average of Confirmed Cases in Germany

**7. Feature Engineering**

* Extract features like day of the week, month, or year from date columns to capture seasonal patterns and trends for better time series forecasting.
* Utilize latitude and longitude to create features identifying regional patterns and proximity to hotspots, improving risk assessments.
* Combine related columns like Confirmed cases, Deaths, and Recovered to create features like mortality or recovery rates, offering better insights into the pandemic’s progression.
* Selected important features using feature selection techniques.

**8. Modeling & Prediction**

**Model Selection**

* Implemented various models:
  + **ARIMA MODEL**

P= Past Values

I = differencing

Q= Past Error

**Model Evaluation**

* Performance measured using:
  + Mean Squared Error (MSE)
  + R-squared

**9. Model Comparison Report**

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R-squared** |
| Arima Model | 61143869.321031004 | 0.8185140269955934 |

**10. Challenges Faced and Solutions**

**Data Quality Issues**

* Missing Values in Province/State Column.
* Handling Large Dataset with Numerous Features.
* Unpivoting the Data.
* Time Series Forecasting with ARIMA Model.
* Performance Evaluation of Predictive Models.

**Model Performance**

* Feature selection techniques were used to reduce the dimensionality of the data.
* The data was unpivoted to transform it into a long format, which is better suited for time series analysis.
* The ARIMA model was selected, and careful parameter tuning was performed to ensure that the best fit for the data was achieved.
* Feature selection was used to avoid overfitting.

**11. Deployment**

Flask Link - <http://127.0.0.1:5000>

**12. Suggestions to the Government:**

* **Resource Allocation:** Use ARIMA model forecasts to allocate healthcare resources (hospital beds, ventilators) in regions predicted to experience surges in cases.
* **Early Warning Systems:** Implement systems that track potential spikes and enable timely interventions, like increasing hospital capacity or activating emergency protocols.
* **Public Health Campaigns:** Focus on regions with predicted surges, promoting vaccination, social distancing, and mask-wearing to reduce future infections.
* **Surveillance & Data Collection:** Enhance surveillance for new variants and update predictive models with real-time data for accuracy.
* **Vaccine Distribution:** Prioritize vaccine rollouts in high-risk areas to reduce severe cases.
* **Healthcare Workforce Support:** Strengthen the healthcare workforce with additional staff, training, and protective equipment.
* **Testing & Contact Tracing:** Increase testing and contact tracing in high-transmission regions to identify and isolate cases quickly.
* **Emergency Plans:** Refine preparedness plans for scalable healthcare and logistical support during future surges.
* **Data-Driven Policies:** Adjust policies regularly based on predictive trends for better pandemic management.

**13. Conclusion**

* The analysis of the Johns Hopkins University COVID-19 dataset highlighted global trends in cases, deaths, and recoveries. By cleaning and transforming the data, we applied the ARIMA model for time series forecasting, achieving good accuracy with an R-squared value of 0.82.
* Challenges like missing values, large feature sets, and data unpivoting were addressed through feature selection, log transformations, and focusing on relevant data.
* The findings, including daily trends and forecasts, provide valuable insights for managing the pandemic. The analysis emphasizes the role of predictive models in planning for future healthcare needs and responding to ongoing COVID-19 developments.