### **Report on the Development of the Notebook**

#### **1. Objective:**

The primary goal of this notebook was to analyze COVID-19 data, perform exploratory data analysis (EDA), and build predictive models for confirmed cases. The focus was on understanding the data's trends and developing models to predict future confirmed cases.

### **2. Process and Workflow:**

#### **2.1 Data Collection and Cleaning:**

* The COVID-19 dataset was loaded from Kaggle. Additional data, including country populations, was incorporated for contextual analysis.
* **Date Conversion**: The Date column was converted to a DateTime format for time-series analysis.
* **Missing Values**: Missing data in columns such as Mortality Ratio, Daily Spread Rate, and Cases Per Million were filled or replaced with zeros.
* **Feature Engineering**:
  + Calculated new features such as Mortality Ratio, Daily Spread Rate, and Cases Per Million to add more dimensions to the analysis.
  + Added rolling averages and shifted versions of features like Confirmed, Deaths, Recovered, and Active to capture trends over time.

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

* **Visualizations**:
  + Created heatmaps to analyze correlations among numerical columns.
  + Generated global scatter plots of confirmed cases, deaths, and recoveries.
  + Trend analysis over time using line plots for confirmed, deaths, active, and recovered cases.
  + Regional comparisons of confirmed cases using bar plots and density plots.
* **Insights**:
  + Data showed skewed distributions for confirmed cases, recoveries, and deaths, with notable outliers.
  + Mortality rates varied significantly by region, and the spread rate fluctuated over time.

#### **2.3 Feature Selection and Preprocessing:**

* Redundant columns (e.g., Lat, Long, WHO Region) were dropped.
* Features were scaled using StandardScaler to normalize ranges and mitigate outliers.
* The dataset was split into training, validation, and testing subsets using TimeSeriesSplit and DARTS's built-in methods.

#### **2.4 Model Training and Evaluation:**

Two models were developed:

1. **Linear Regression**:  
   * Target Variable: Confirmed.
   * Features: Engineered and scaled features such as rolling averages, shifts, and population-based metrics.
   * Performance:
     + Root Mean Squared Error (RMSE): High.
     + Coefficient of Determination (R²): Indicates poor explanatory power.
2. **Naive Moving Average (NMA)**:  
   * Used as a baseline time-series forecasting model.
   * Performance:
     + Validation MAE: **7154.64**.
     + Test MAE: **15963.29**.
   * The model struggled to generalize, as evidenced by the high error on the test set.

### **3. Areas of Improvement:**

1. **Data Quality**:  
   * Confirmed cases, deaths, and recovery columns contained missing values, which may have affected model performance.
   * Address data imbalance and outliers through techniques like transformation or trimming.
2. **Feature Engineering**:  
   * Add lag features for confirmed cases (e.g., Confirmed (t-1), Confirmed (t-2)).
   * Enrich the data set by including exogenous variables, such as lockdown policies, vaccination rates, and demographic factors, to enhance model input.
3. **Model Selection**:  
   * Replace or complement simple models with advanced models:
     + **ARIMA/Seasonal ARIMA**: Suitable for capturing trends and seasonality.
     + **Gradient Boosting Models** (e.g., XGBoost, LightGBM): For more complex feature interactions.
     + **Deep Learning**: Long Short-Term Memory (LSTM) networks for sequential modeling.
4. **Evaluation Metrics**:  
   * Use additional metrics like RMSE and Mean Absolute Percentage Error (MAPE) for better context on model accuracy.
5. **Data Splitting**:  
   * Train-test splitting may have temporal leakage due to random assignment. Ensure strict temporal separation.

### **4. Summary:**

#### **EDA Insights:**

* COVID-19 data exhibited significant skewness, with high variability across regions.
* Mortality ratios and case-per-million metrics highlighted disparities in impact and response.
* Trends showed confirmed cases followed a rapid growth pattern during early 2020, with varying recovery rates by region.

#### **Model Performance:**

1. **Linear Regression**:
   * Poor performance is likely due to the linear nature of the model, which fails to capture the time-series dependency and complex relationships in the data.
2. **Naive Moving Average**:
   * As a baseline model, it provided a simple forecast but needed more sophistication to handle complex time-series patterns.
   * The large gap between validation and test MAE indicates poor generalization.

#### **Conclusion:**

Both models need to perform better, likely due to the dataset's complexity and limited modeling techniques. Enhancing feature engineering, incorporating exogenous variables, and using more robust models are critical next steps for improving predictive accuracy.