Report on COVID-19 Data Analysis and Modeling

1. Objective

The goal of this analysis is to explore COVID-19 data, preprocess it for modeling, and use a machine learning approach (Decision Tree Classifier) to predict the severity of the outbreak based on deaths.

---

2. Data Overview

Dataset Source:

- File: `covid_19_clean_complete.csv`
- Rows: 49,068
- Columns: 10
- `Country/Region`: Name of the country.
- `Lat` and `Long`: Geographical coordinates.
- `Date`: Date of the observation.
- `Confirmed`: Number of confirmed COVID-19 cases.
- `Deaths`: Number of deaths.
- `Recovered`: Number of recovered cases.
- `Active`: Number of active cases.
- `WHO Region`: Regional classification by the WHO.

Initial Observations:

- `Province/State` column has 14,664 non-null entries and was dropped due to limited relevance.
- No missing values in other key columns after processing.

---

3. Data Cleaning and Preprocessing

- Removed duplicates and null values.
- Encoded the categorical column `WHO Region` using `LabelEncoder`.
- Scaled numerical features (`Confirmed`, `Deaths`, `Recovered`, `Active`) using `StandardScaler`.

New Features:

- **High_Severity**: A binary target column indicating high severity if deaths > 500.

---

4. Data Exploration

Descriptive Statistics:

- **Confirmed cases**: Mean = 16,884; Max = 4,290,259.
- **Deaths**: Mean = 884; Max = 148,011.
- **Recovered**: Mean = 7,916; Max = 1,846,641.
- **Active cases**: Mean = 8,085; Max = 2,816,444.

Key Observations:

- Most countries have relatively low numbers of deaths (<500).
- The data includes significant outliers (e.g., cases >4 million, deaths >148k).

Visual Analysis:

- Histograms revealed skewed distributions in `Confirmed`, `Deaths`, and `Recovered`.
- Scatter plots showed strong correlations between `Confirmed` and `Deaths` but weaker correlations with `Active` cases.
- Heatmap confirmed that `Confirmed`, `Deaths`, and `Recovered` are highly interrelated.

---

5. Modeling

Steps:

1. Splitting data into training and test sets (80% train, 20% test).
2. Training a Decision Tree Classifier with default parameters.
3. Evaluating the model using accuracy and a classification report.

Results:

- **Accuracy**: The model achieved an accuracy score of approximately 95% on the test set.
- **Classification Report**:
    - High precision and recall for predicting both classes.
    - Class imbalance due to a limited number of high-severity cases.

Model Limitations:

- The threshold for `High_Severity` (deaths > 500) is arbitrary.
- Decision Tree models may overfit; cross-validation and hyperparameter tuning were not applied.

---

6. Limitations and Challenges

- **Data Quality**: Extreme outliers in numerical columns and potential inconsistencies in reporting.
- **Feature Engineering**: A more robust approach could involve time-series analysis or country-specific factors (e.g., population).
- **Model Choice**: A more complex model like Random Forest or Gradient Boosting may yield better generalization.

---

7. Recommendations

- Explore hyperparameter tuning (e.g., tree depth, split criteria) for the Decision Tree Classifier.
- Use oversampling techniques (e.g., SMOTE) to handle class imbalance for `High_Severity`.
- Perform detailed E