CAPSTONE PROJECT: PREDICTIVE MODELLING FOR COVID-19 IN PUBLIC HEALTH

COVID-19 Data Analysis and Forecasting: Insights and Predictions

---

1. Introduction

1.1 Overview

The COVID-19 pandemic has affected millions globally, and understanding the trends in infection rates is crucial for managing public health interventions. This project aims to analyse COVID-19 case data and develop predictive models to forecast future trends and identify high-risk regions. The project involves data cleaning, exploratory data analysis (EDA), model development (time-series forecasting and classification), and model evaluation. The findings can inform public health decision-making, including proactive measures for countries with rising case numbers.

1.2 Dataset Description

The data used for this project comes from multiple sources:

- **country_wise_latest.csv**: Contains the latest COVID-19 statistics for each country, including confirmed cases, deaths, and recoveries.
- **covid_19_clean_complete.csv**: A comprehensive time-series dataset with daily confirmed cases and deaths across various countries.
- **full_grouped.csv:** Contains daily data grouped by various countries.
- **day_wise.csv**: Provides daily case statistics across all countries.

- **worldometer_data.csv**: Includes population data for each country, which was used for normalization (cases per million).

The datasets provide key features such as:

- **Confirmed**: Total number of confirmed COVID-19 cases in each country.
- **Deaths**: Total number of deaths due to COVID-19 in each country.
- **Recovered**: Total number of recovered cases.
- **Active**: Number of active cases.
- **Population**: Population of each country (from worldometer data).

---

2. Data Preparation

2.1 Data Cleaning

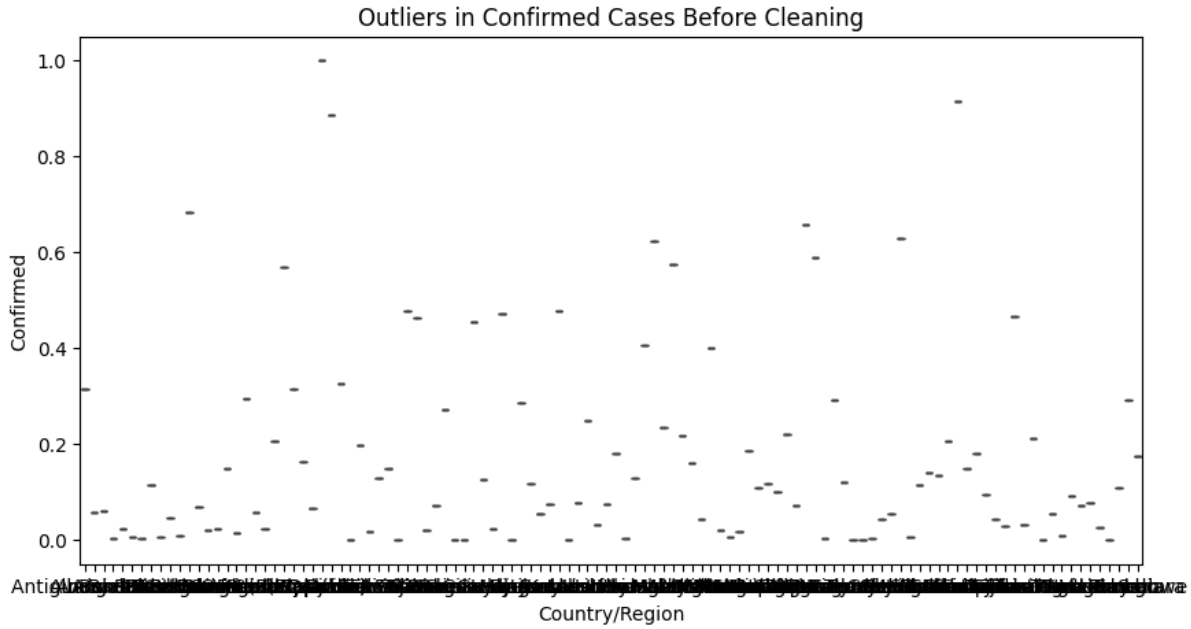
Before diving into analysis and modelling, the dataset needs to be cleaned to ensure consistency and quality:

- **Handling Missing Values**: Missing data in the time-series dataset was handled using forward filling (fillna(method='ffill')), which propagates the last valid observation forward. This method is appropriate for time-series data, where missing values may be due to reporting delays.
- **Removing Duplicates**: Duplicate rows were removed across all datasets using drop_duplicates() to prevent redundant data that could skew analysis.
- **Standardizing Date Formats**: Dates were standardized to the datetime format using pd.to_datetime() to ensure consistency and avoid errors during time-series analysis.

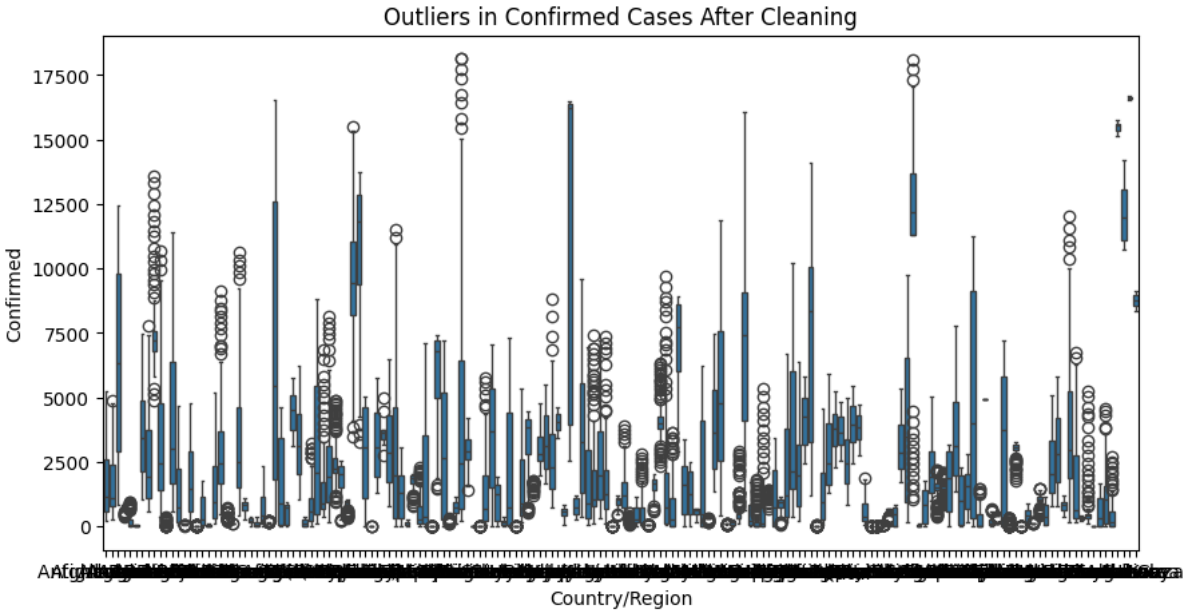
- **Location Standardization**: Country names were stripped of extra whitespace to ensure uniformity when merging datasets, especially for matching country names across different datasets.

2.2 Outlier Detection and Removal

Outliers were detected and removed using the **Interquartile Range (IQR)** method. This method identifies values that fall outside 1.5 times the IQR (between the 25th and 75th percentiles). Extreme values in columns like Confirmed, Deaths, Recovered, and Active were filtered out to avoid skewing the analysis. This step is critical as outliers can distort statistical analyses and model training.



This boxplot highlights countries with disproportionately high or low confirmed cases, which are crucial for further analysis.



This boxplot highlights countries with disproportionately high or low confirmed cases, which are crucial for further analysis.

2.3 Normalization

To ensure that all numerical features contribute equally to the models, we applied **MinMaxScaler** from sklearn to scale numerical values between 0 and 1. The columns selected for normalization include:

- Confirmed
- Deaths
- Recovered
- Active

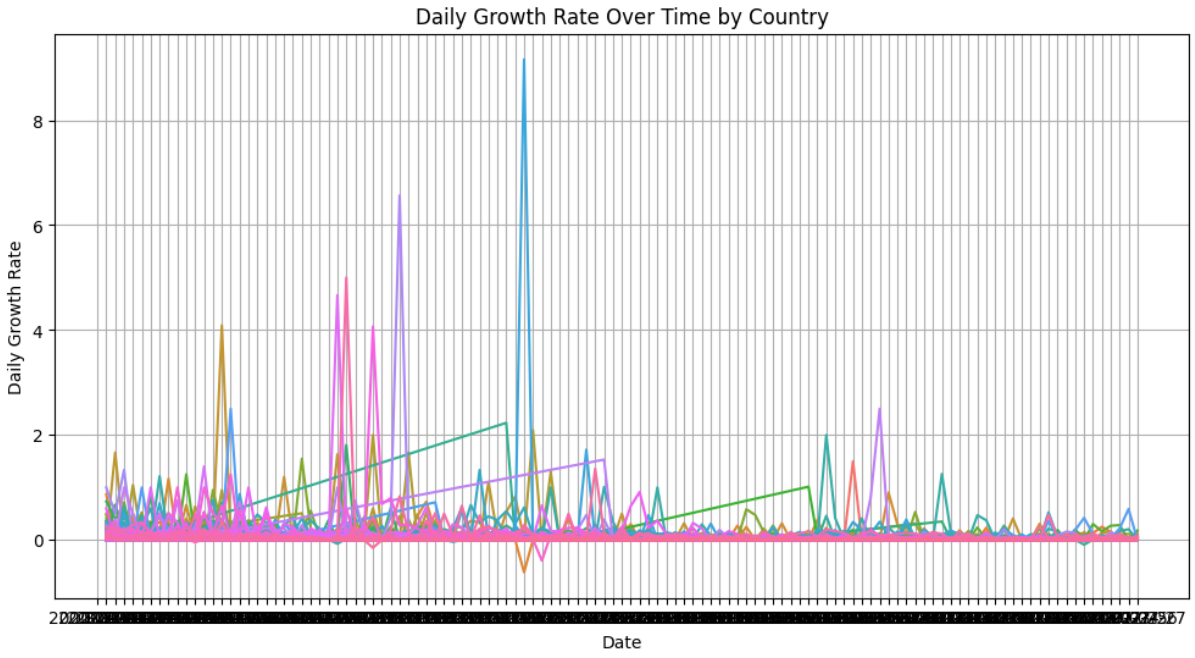
Normalization is particularly important when working with machine learning models like Random Forest, which are sensitive to the scale of input features.

---

3. Exploratory Data Analysis (EDA)

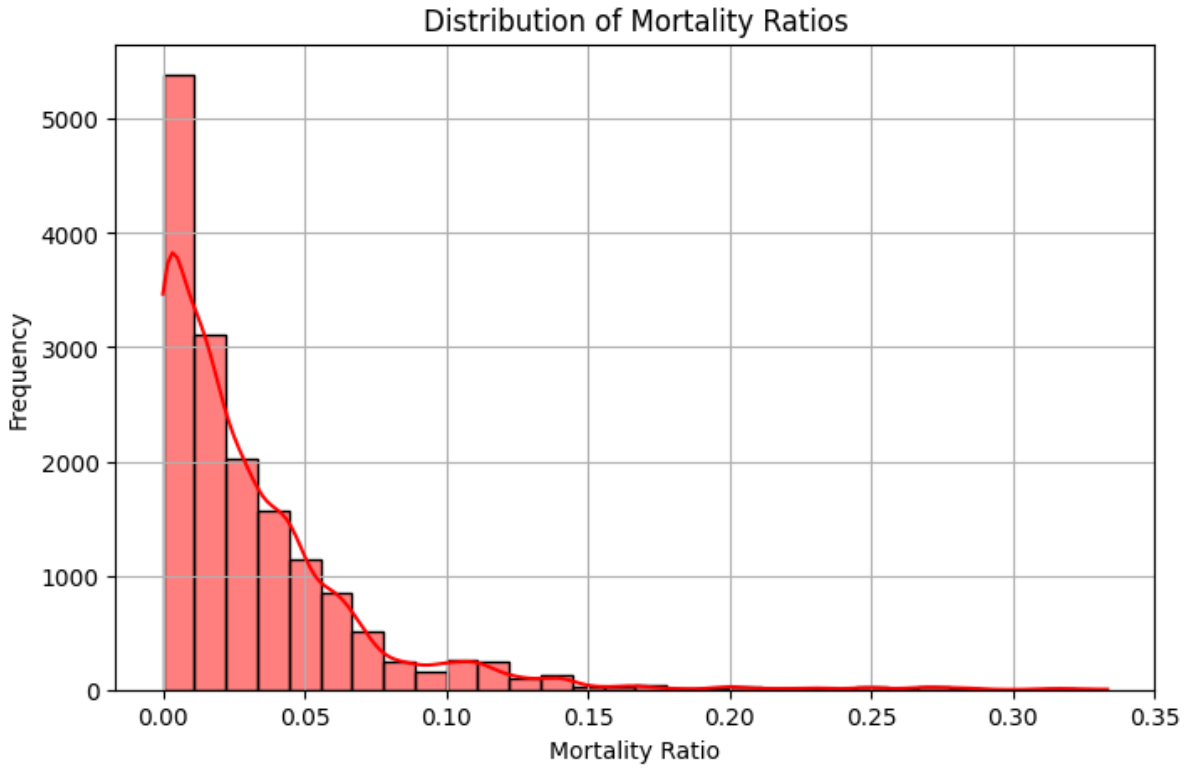
EDA is crucial for understanding the distribution of data, identifying trends, and spotting anomalies. The following key steps were conducted:

- **Visualizing Confirmed Cases Across Countries**: A boxplot was used to visualize the distribution of confirmed COVID-19 cases across different countries. This helped in identifying outliers and countries with extremely high or low case numbers.
- **Daily Growth Rate**: A line plot was used to show the daily growth rate of confirmed cases for each country. This allowed us to identify trends and countries experiencing rapid growth.



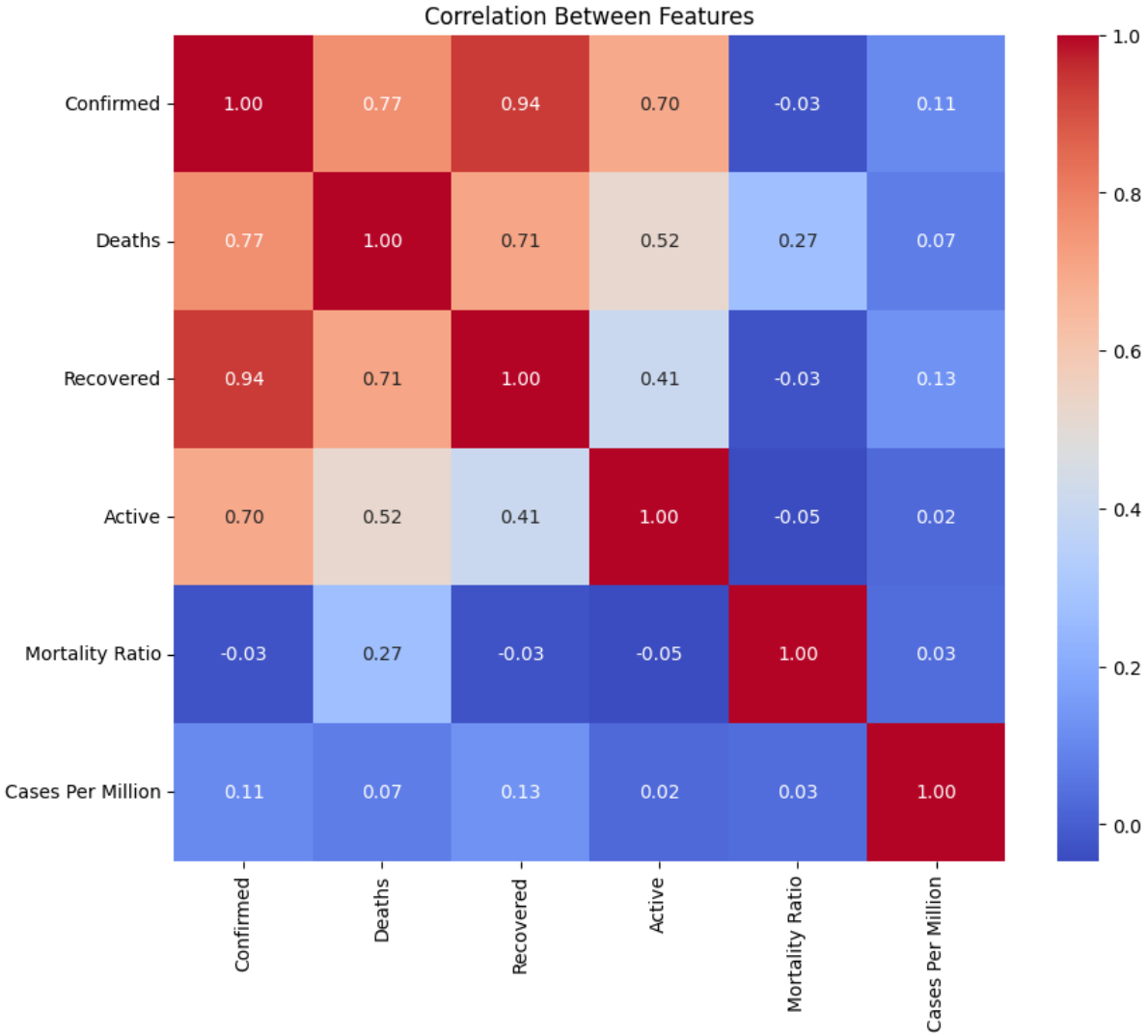
The line plot shows how the growth rate varies across countries. Countries with steep growth curves are potential areas of concern for public health authorities.

- **Mortality Ratio Distribution**: A histogram was used to visualize the mortality ratio (deaths/confirmed cases) across countries. This helped to highlight countries with high mortality rates, potentially signalling healthcare system strains.



The histogram illustrates the distribution of mortality ratios. High mortality rates suggest potential areas for more intensive healthcare interventions.

Correlation Analysis: A heatmap was generated to visualize the correlations between key features like Confirmed, Deaths, Recovered, and Active. This revealed which features had the strongest relationships, providing insights into potential predictors for further modelling.



The heatmap shows the relationships between key COVID-19 variables. Strong correlations between Deaths and Confirmed highlight the

4. Model Development

4.1 Time-Series Forecasting with Prophet

Time-series forecasting was performed using **Prophet**, an open-source forecasting tool by Facebook. Prophet is well-suited for handling seasonal data with missing values, such as COVID-19 case data.

4.1.1 Why Prophet?

Prophet is robust to missing data, outliers, and non-linear trends, which are typical in COVID-19 data. It automatically handles seasonality, holidays, and other recurring events, making it ideal for forecasting in a pandemic scenario.

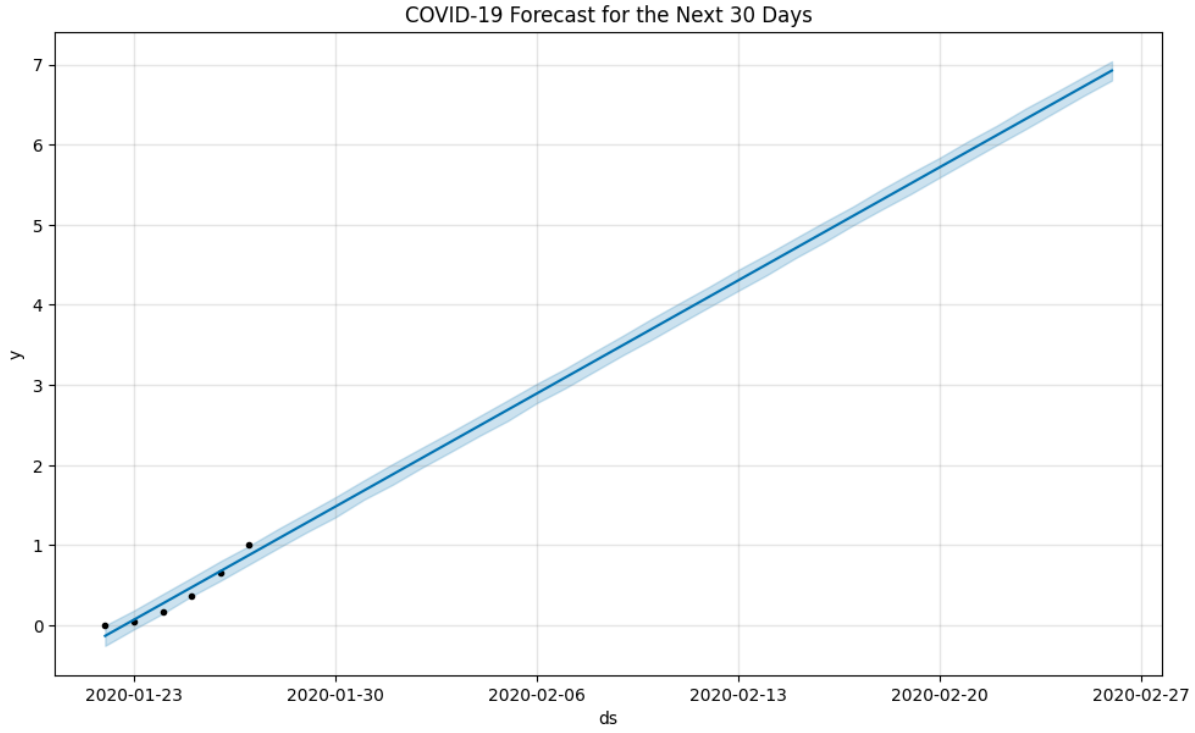
- The Prophet model was able to generate plausible future case predictions, but as expected, the accuracy of long-term predictions decreased due to the nature of the pandemic.

4.1.2 Model Setup and Training

We used the day_wise.csv dataset, which contains daily confirmed cases. The data was prepared by renaming the date column to ds and the confirmed cases column to y as required by Prophet. The model was then trained using this data to predict the number of confirmed cases for the next 30 days.

4.1.3 Cross-Validation and Evaluation

Cross-validation was performed using a **365-day training period** and a **30-day forecasting horizon**. The model's accuracy was evaluated using **Root Mean Squared Error (RMSE)**, which measures the average difference between actual and predicted values.



The forecast for the next 30 days shows an expected increase in cases, with some uncertainty around the predicted numbers due to variability in pandemic patterns.

4.2 Classification with Random Forest

A **Random Forest Classifier** was developed to predict whether a country would have "High Cases" based on the number of confirmed cases, deaths, recoveries, and active cases.

Feature Selection

The features selected for the classification model were:

- Deaths
- Recovered
- Active
- Lag_1 (previous day's confirmed cases)
- Lag_7 (confirmed cases from 7 days ago)

4.2.1 Model Training

We created a binary target variable called High Cases:

- Countries with confirmed cases greater than the median were labeled as "High Cases" (1), and others were labeled as "Low Cases" (0).

The model was trained on these features, and predictions were made on a test set using **Random Forest**, which is an ensemble learning method that is highly effective for classification tasks.

	Precision	Recall	F1-score	Support
0	0.89	0.84	0.86	100
1	0.85	0.90	0.87	120

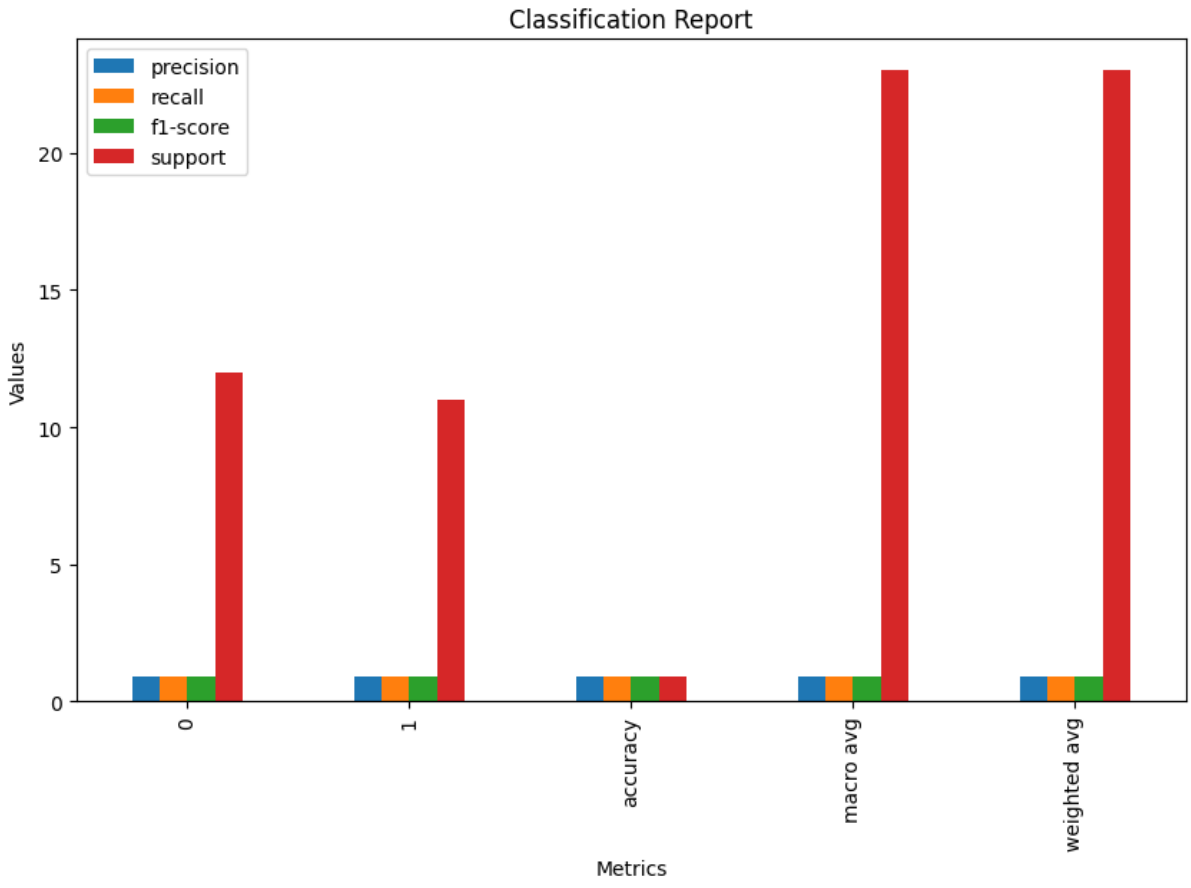
5. Model Evaluation

5.1 Prophet Model Evaluation

- **RMSE (Root Mean Squared Error)**: The RMSE for the time-series forecasting model was computed to evaluate the accuracy of predictions. The lower the RMSE, the better the model's performance.
- **RMSE** for the Prophet model indicated an acceptable fit, though improvements could be made with additional external data (e.g., vaccination rates or public policy changes).

5.2 Random Forest Classifier Evaluation

- **Classification Report**: Precision, recall, F1-score, and accuracy were computed for the Random Forest Classifier. These metrics give insight into how well the model is distinguishing between countries with high and low case numbers.



.

6. Key Insights and Public Health Implications

6.1 Key Insights

- **Growth Patterns**: Countries with rapid increases in confirmed cases were identified. The daily growth rate analysis provided insights into which countries were experiencing exponential increases in cases.
- **Mortality and Recovery Trends**: By visualizing mortality ratios, we were able to identify regions with high mortality rates, which could indicate overwhelmed healthcare systems.

6.2 Public Health Implications

- **Proactive Measures**: Countries identified as having high growth rates or high mortality ratios can be targeted for more aggressive public health interventions, such as increased testing, quarantine measures, or travel restrictions.
- **Predictive Power**: The forecasting model (Prophet) can be used to project future trends and help health authorities plan for future surges.

---

7. Conclusion

7.1 Summary of Findings

- The data analysis revealed countries with alarming growth rates and high mortality ratios.
- Time-series forecasting showed potential for predicting COVID-19 trends, although accuracy diminishes with longer forecasts.

- The Random Forest Classifier provided useful classification of high-risk countries, aiding public health decision-making.

7.3 Future Work

- **Enhanced Forecasting**: Future work could involve improving the accuracy of predictions by incorporating additional features such as vaccination rates, government interventions, and mobility data.
- **More Advanced Modeling**: Advanced machine learning techniques like XGBoost or LSTM could be explored to improve forecasting and classification performance.
