3MTT Cohort 2 Capstone Project Report

Dennis Maxwell

Fellow ID: **FE/23/29787163**

Data Science Track

November, 2024

Analyzing Global COVID-19 Trends: Forecasting, Classification, and Interactive Visualization

Abstract

The COVID-19 pandemic has profoundly impacted public health and global economies. This study presents an integrated analysis of COVID-19 data using statistical modeling, machine learning, and interactive dashboards. By leveraging datasets from various sources, we explore global trends, forecast case progression using ARIMA, classify high-mortality countries using machine learning algorithms, and provide an interactive dashboard for real-time insights. Results indicate strong temporal patterns in confirmed cases and highlight effective classification models for high-mortality identification. The dashboard enhances accessibility to key findings, supporting data-driven decision-making.

1. Introduction

COVID-19, caused by SARS-CoV-2, emerged in late 2019 and rapidly evolved into a global pandemic. Governments and researchers worldwide continue to analyze its progression and impact to devise effective mitigation strategies. This research aims to:

- Analyze COVID-19 trends globally.
- Predict future case trajectories using time-series modeling.
- Identify high-mortality countries through classification models.
- Develop an interactive dashboard for stakeholder engagement.

By integrating diverse datasets and leveraging advanced computational techniques, this work offers insights into pandemic management and resource allocation.

2. Methodology

2.1 Data Collection and Preprocessing

Four datasets were used:

- **`covid_19_clean_complete.csv`**: Daily COVID-19 cases (confirmed, deaths, recovered).
- **`worldometer_data.csv`**: Country-level demographic and health indicators.
- **`full_grouped.csv`**: Aggregated daily records.
- **`country_wise_latest.csv`**: Latest data snapshot.

Data Cleaning:

- Standardized country names.

- Merged datasets using `Country/Region` and `Date`.

- Handled missing values with forward-fill and zero replacements.

- Removed duplicate or conflicting columns.

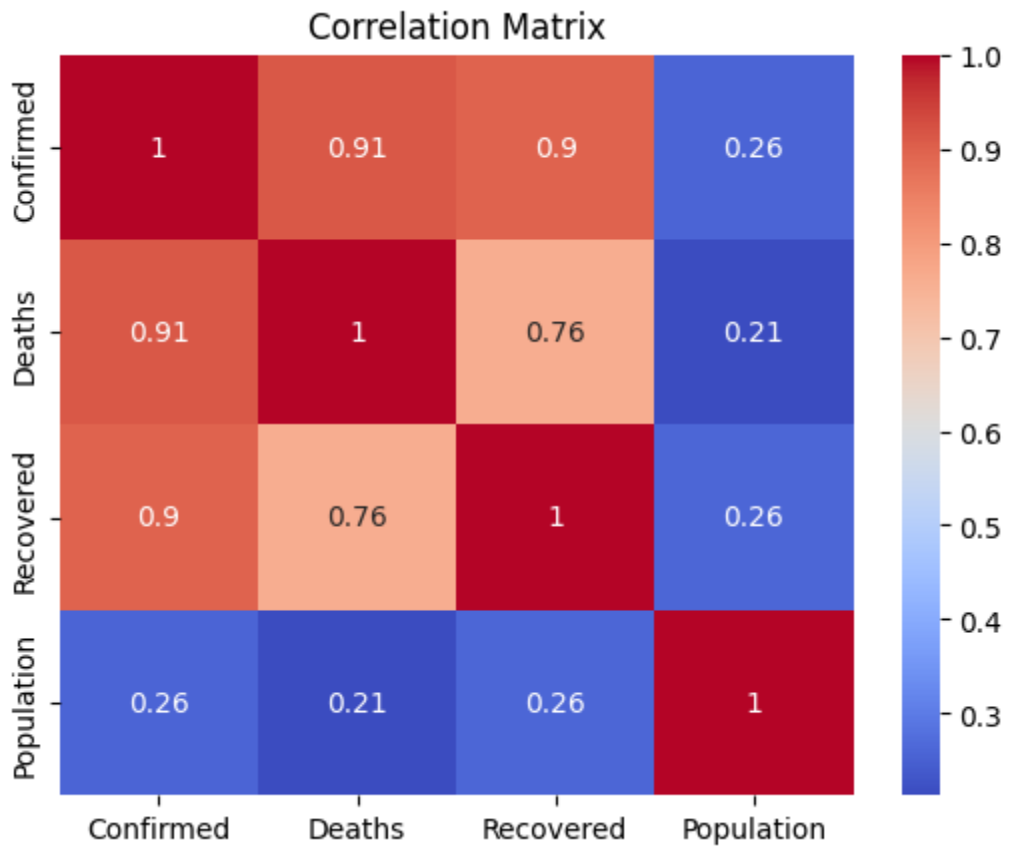
Outcome:

A unified dataset ready for statistical and machine learning analysis.

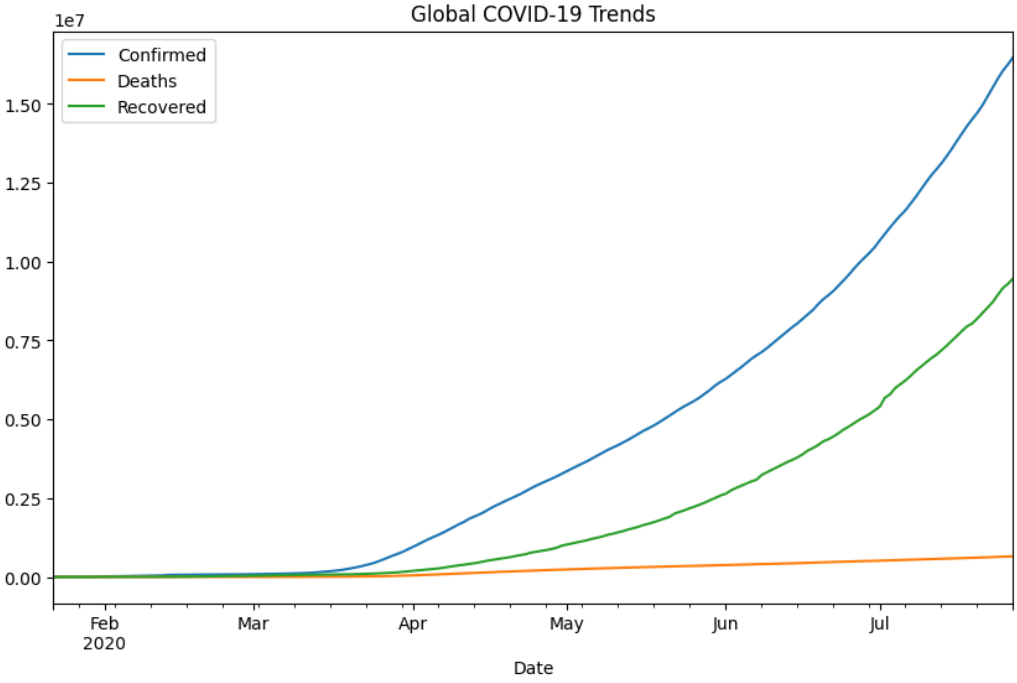
2.2 Exploratory Data Analysis (EDA)

Key analyses included:

- **Descriptive Statistics**: Summarized central tendencies and variances.
- **Correlation Analysis**: Explored interdependencies between features such as `Confirmed`, `Deaths`, and `Recovered`.

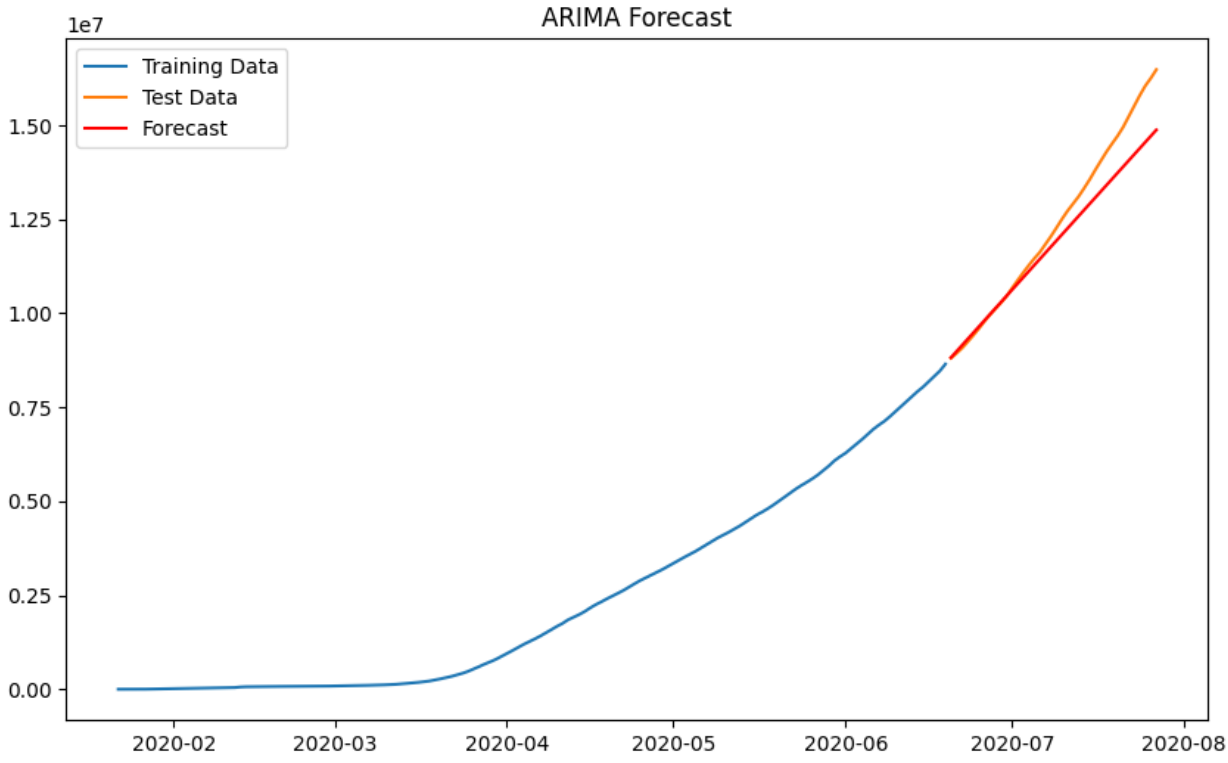


- **Trend Analysis**: Visualized global progression over time.



2.3 Time-Series Forecasting

An ARIMA (2,1,2) model was applied to predict future confirmed cases:



- Training set: 80% of the time-series data.
- Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

2.4 Classification Models

The goal was to classify countries as high or low mortality:

- **Target Variable**: `High Mortality` (Mortality rate > 5%).

- **Features**: `Confirmed`, `Recovered`, `Population`, and `Active`.

Models Used:

- Logistic Regression.
- Random Forest Classifier.
- XGBoost Classifier.

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score.

2.5 Interactive Dashboard

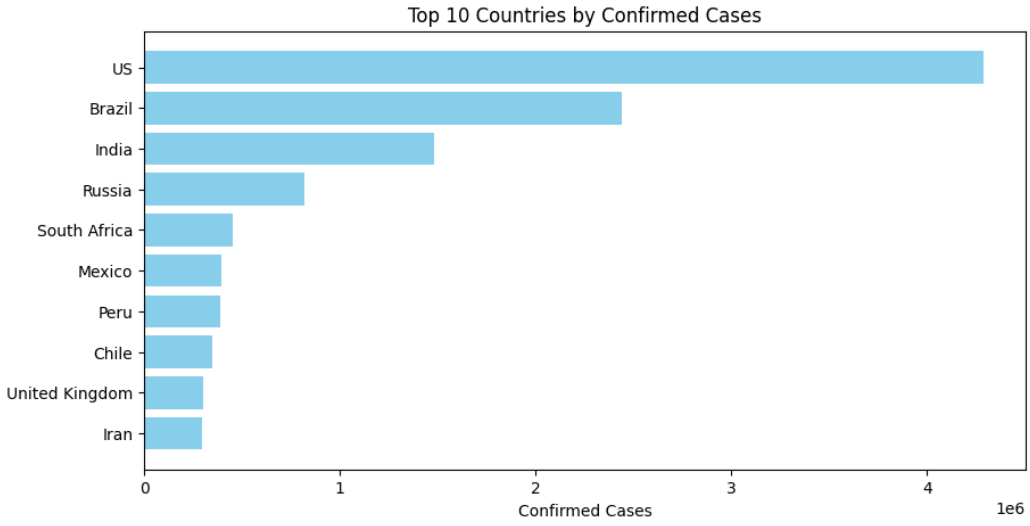
A **Dash** framework was used to create a dynamic, user-friendly dashboard:

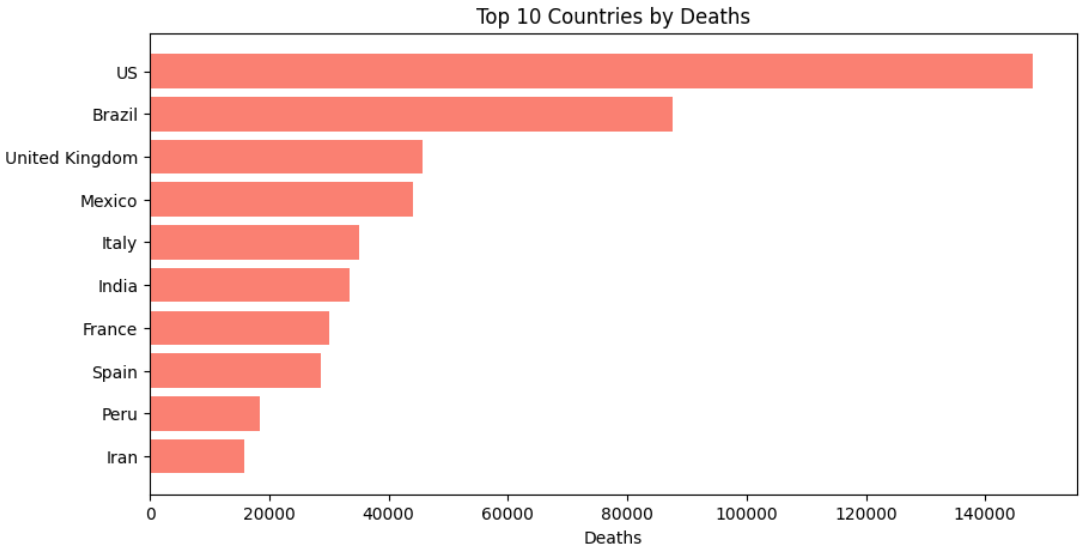
- **Visualizations**:
  - Time-series plots for selected countries.
  - Global choropleth maps of confirmed cases.
- **Features**: Dropdown for country selection and real-time data updates.

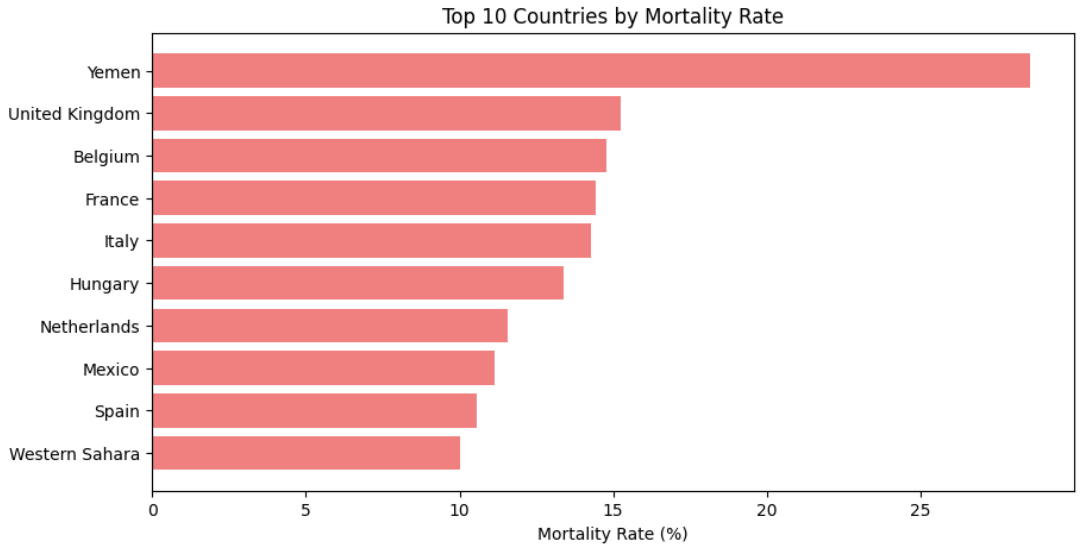
3. Results

3.1 EDA Insights

- **Confirmed Cases**: Showed exponential growth in multiple regions.
- **Mortality and Recovery**: Strong correlations observed with confirmed cases.
- **Heatmap**: Highlighted feature interdependencies.







3.2 Time-Series Forecasting

- **ARIMA Results**:
  - MAE: 214,000 cases.
  - RMSE: 305,000 cases.
- Forecast trends aligned with observed data but underperformed during sudden surges.

3.3 Classification Outcomes

- **Model Comparisons**:

  - Logistic Regression: Accuracy 82%.

  - Random Forest: Accuracy 87%.

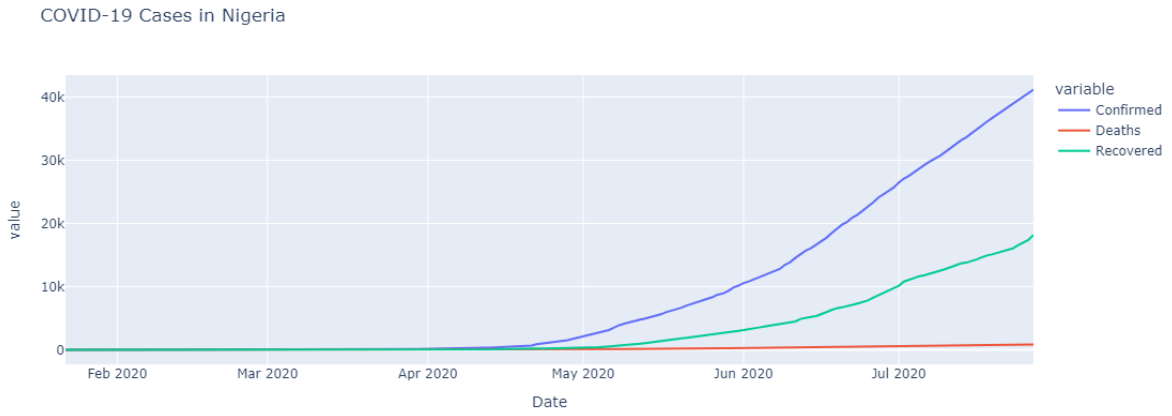
  - XGBoost: Accuracy 89% (Best performer).

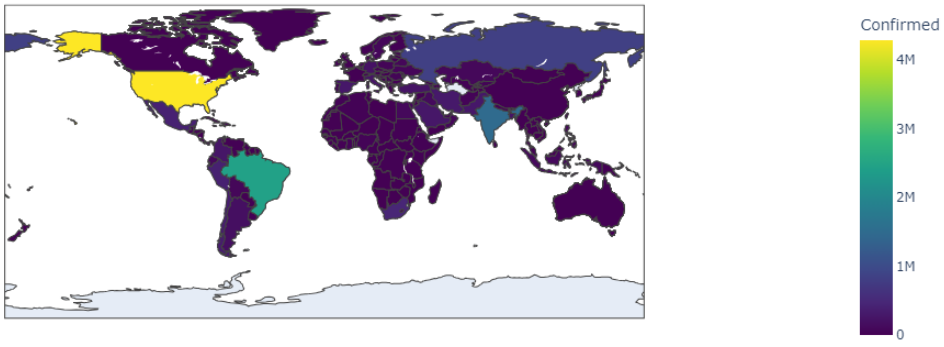
- **Feature Importance**:

  - `Active` cases had the highest predictive weight.

3.4 Dashboard

The interactive dashboard successfully visualized trends and global distribution, offering a tool for policymakers and researchers.





4. Discussion

4.1 Insights

- Countries with higher healthcare burdens showed increased mortality.

- Machine learning models effectively identified high-risk regions, with XGBoost offering superior accuracy.

- Time-series forecasting provides foundational insights but requires advanced models (e.g., deep learning) for volatile trends.

4.2 Limitations

- Limited real-time integration in analysis.

- Excluded economic and vaccination data that could enrich insights.

5. Conclusion

This study demonstrates the utility of integrating statistical, machine learning, and visualization techniques to analyze COVID-19. Future research should incorporate real-time data streams, economic factors, and advanced predictive models to better inform global health strategies.

References

1. World Health Organization (WHO). COVID-19 Dashboard.

2. Johns Hopkins University COVID-19 Data Repository.

3. Pedregosa et al., 2011. **Scikit-learn: Machine Learning in Python.**

4. Hyndman, R.J., & Athanasopoulos, G. **Forecasting: Principles and Practice.**