Technical Report: Predictive Modeling for COVID-19 in Public Health

1. Executive Summary

This report presents a comprehensive predictive modeling system for understanding and mitigating the impact of COVID-19. Using historical data, the following was performed:

- performed data cleaning
- exploratory data analysis (EDA)
- predictive modeling.

The key deliverables include actionable insights to inform public health policies, forecast COVID-19 trends, and improve resource allocation.

2. Data Preparation

Data Source: below is the source of the dataset used:

- Dataset: COVID-19 Clean Complete Dataset
- Source: Kaggle (CORD-19)

Steps in Data Preparation:

```
i. Cleaning:
    - Addressed missing values by filling them with zeros where applicable.
    - Removed duplicate entries to ensure data integrity.
    - Standardized date formats using pd.to_datetime.

ii. Feature Engineering:
    - Derived Variables:
        - Daily Growth Rate (%): Percentage change in confirmed cases for each country.
        - Mortality Rate (%): Deaths as a percentage of confirmed cases.
        - Cases per Population: Confirmed cases per unit population (assumed population for simplicity).

iii. Transformation:
    - Normalized numerical features (Confirmed, Deaths, Recovered, Active) using Min-Max Scaling.
```

Outcome: Data cleaning and transformation ensured consistency and reliability for analysis and modeling.

3. Exploratory Data Analysis (EDA)

Objectives:

- Identify trends, correlations, and outliers.
- Understand key factors affecting COVID-19 transmission and severity.

Key Insights:

```
i. Global Trends:
    - Steady growth in confirmed cases during early pandemic phases, followed by periodic waves.
    - Mortality rates showed gradual decline over time, possibly due to improved treatment protocols.

ii. Top 5 Affected Countries:
    - Countries with the highest confirmed cases demonstrated varied trends in mortality and recovery rates.

iii. Correlations:
    - Strong positive correlation between Active Cases and Confirmed Cases.
    - Negative correlation between Mortality Rate and Daily Growth Rate.

iv. Outliers:
    - High mortality rates observed in a few countries with older populations or overwhelmed healthcare systems.
    - Daily growth rates show outliers where specific days had spikes in confirmed cases.
```

Visualizations:

- Line plots for global trends (Confirmed, Deaths, Recovered, Active).

- Correlation heatmaps.
- Boxplots for outlier analysis.

4. Model Development and Performance

```
4.1 Time-Series Forecasting (ARIMA Model):
    - Objective: Predict future confirmed cases globally.
    - Model Parameters: ARIMA (5, 1, 0)
    - Evaluation Metric: Root Mean Squared Error (RMSE)
        - RMSE: 4.83 (indicative of moderate predictive performance).
    - Findings: The model effectively captured historical trends but struggled with unexpected
variations due to policy changes or variants.

4.2 Classification Model (Random Forest):
    - Objective: Predict mortality classes (High vs. Low Mortality Rate).
    - Features: Cases per Population, Active Cases
    - Performance Metrics:
       - Accuracy: 76%
       - Precision: 79%
       - Recall: 80%
       - F1-Score: 77%
   - Feature Importance:
       - Cases per Population had the highest impact on classification.
   - Findings: The model demonstrated strong performance in identifying mortality risk based
on key features.
```

5. Public Health Insights and Recommendations

```
i. Resource Allocation:
    - High Cases per Population regions should receive priority in healthcare resource
distribution.

ii. Mortality Risk Mitigation:
    - Countries with high mortality risk should focus on vaccination campaigns and enhancing
healthcare infrastructure.

iii. Growth Rate Monitoring:
    - Continuous tracking of Daily Growth Rate can help predict and respond to outbreak
surges.

iv. Data-Driven Policies:
    - Insights from time-series trends can guide decisions on lockdowns and social distancing
measures.
```

6. Visualizations Summary

```
i. Global Trends:
    - Line plots of confirmed, active, recovered, and death cases.

ii. Mortality and Growth Rates:
    - Scatter plots to identify relationships and trends.

iii. Model Interpretability:
    - Feature importance bar chart for the classification model.
    - Confusion matrix for model evaluation.
```

7. Conclusion

The project successfully achieved its objectives of deriving actionable insights and building predictive models to assist public health organizations. Future improvements could include integrating demographic data and advanced deep learning models to enhance predictive accuracy.

In [ ]:

In [ ]: