3MTT Capstone Project Report by Rafiyat Omolola Oyeniyi FE/23/60944262

Title: *Analysis and Prediction of COVID-19 Trends Using Statistical and Machine Learning Techniques*

1. Introduction

The project focuses on analyzing the impacts of COVID-19 using historical datasets. It explores trends, relationships, and key metrics across global and regional levels. Using machine learning models and statistical methods, the project aims to predict future outcomes, offering insights into how COVID-19, a global pandemic, profoundly impacted public health and economies worldwide evolved over time and its variations by WHO regions.

The Objectives of the analysis are:

1. Explore and visualize COVID-19 data to uncover trends and patterns.
2. Conduct an exploratory data analysis to uncover patterns and insights.
3. Create visualizations to represent trends in confirmed, recovered, and death cases.
4. Build predictive models to forecast future cases and classify data trends.
5. Classify data to gain insights into factors influencing recovery and mortality rates.

2. Data Loading and Preprocessing

The datasets used for the analysis are

1. `covid_19_clean_complete.csv` – A time-series dataset tracking global COVID-19 cases.
2. `country_wise_latest.csv` – A summary dataset with metrics by country.

The Datasets were loaded using `pandas` and missing values were addressed, and features were selected for analysis, the Time-series data was formatted appropriately for modeling.

Key Libraries Used:

- `pandas`, `numpy` – Data manipulation.
- `matplotlib`, `seaborn` – Visualization.
- `scikit-learn` – Machine learning.
- `statsmodels` – Statistical modeling.

3. Exploratory Data Analysis (EDA)

EDA focused on identifying trends and relationships within the data.

From the analysis, we discovered

1. The Global confirmed, death, and recovery rates peaked at different times, highlighting the virus's progression.
2. Certain WHO regions had significantly higher death and confirmed cases, indicating unequal impacts of the pandemic.
3. There is a strong positive correlation between confirmed and death cases observed, while recovery rates varied regionally which helped identify mortality rates by region.

The Visualizations:

1. **Bar Charts**:
    - Total deaths by WHO regions.
    - Death counts by continents.
2. **Line Charts**:
    - Global trends over time, displaying confirmed, recovered, and death cases.
    - Regional trends for confirmed cases across months.
3. **Histograms**:
    - Distribution of confirmed and death cases.
4. **Scatter Plots**:
    - Relationship between confirmed cases and deaths.
5. **Custom Plots**:
    - Death and recovery rates visualized by WHO regions.

4. Modeling

Three models were developed to analyze and predict COVID-19 metrics:

a. Regression Modeling: The objective of using Regression modelling is to predict death counts using features like confirmed cases. We used the Random Forest Regressor.

- **Data Splitting**:
    - Features: Metrics like confirmed, death, and recovery counts.
    - Target: Predict outcomes (e.g., deaths).
    - Data split into training and testing sets (70%-30% split).
- **Performance Metrics**:
    - Mean Absolute Error (MAE): Quantifies average prediction errors.
    - Mean Squared Error (MSE): Captures overall error magnitude.

    - Root Mean Squared Error (RMSE): Indicates model performance in real-world scale.
    - R² score: Measures the proportion of variance explained by the model.

MAE, MSE, RMSE, and R² scores were computed, demonstrating the model's ability to predict with reasonable accuracy.

b. Time-Series Analysis: The Time-series analysis was used to forecast confirmed cases over time. The model used was ARIMA.

- **Evaluation**:
    - RMSE indicated strong performance for short-term forecasting.

The ARIMA model performed well, indicating that past confirmed cases can reasonably predict future trends.

c. Classification Modeling: The objective of using classification modeling is to classify data trends (e.g., recovery vs. death rates). The model used was the Random Forest Classifier.

- **Performance**:
    - Accuracy: High.
    - Precision and Recall: Highlighted strong classification performance but revealed slight imbalances in specific classes.

The classifier achieved good accuracy, providing a robust understanding of data trends.

5. Evaluation

Strengths:

- The models effectively identified trends and patterns in COVID-19 data.
- Visualizations provided intuitive insights into regional and global dynamics.

Limitations:

- Time-series predictions were constrained by dataset size and lack of real-time factors (e.g., vaccination rates, policy changes).
- Dataset biases may have influenced classification results.

6. Conclusion

The project successfully analyzed COVID-19 data, providing key insights into the virus's progression and impacts:

1. Regional disparities in case counts and death rates were evident, underlining the pandemic's uneven impact.
2. Predictive modeling highlighted future trends, aiding decision-making for public health policies.
3. Classification models revealed significant factors affecting recovery and mortality.

7. Recommendations

1. **Public Health Planning**: Use predictive insights to allocate resources effectively in high-risk regions.
2. **Further Analysis**: Incorporate real-time data, such as vaccination rates and new variants, for improved predictions.
3. **Dashboard Development**: Build an interactive dashboard to provide dynamic updates on trends and forecasts.

8. Future Work

- Integrate additional data sources, such as economic indicators or healthcare capacity metrics.
- Develop ensemble models for improved accuracy.
- Explore the socio-economic impact of COVID-19 using advanced data analytics.