### **3MTT Capstone Project Technical Report**

#### **Project Title: Predictive Modelling for COVID-19 in Public Health**

#### **Name:** Khadija Muktar Sukola

**Fellow ID:** FE/23/82197244

**Track:** Data science

**Date:** November, 2024

#### 

#### **Abstract**

This project uses machine learning to predict the total number of COVID-19 cases globally by employing a Random Forest Regressor model. The dataset used in this analysis includes global COVID-19 statistics, such as confirmed cases, deaths, recoveries, population data, and other public health metrics. The report covers the process of data cleaning, exploratory data analysis (EDA), feature engineering, model development, and evaluation. The model's performance was assessed using various metrics, and the findings suggest that machine learning can significantly assist in predicting the future progression of the COVID-19 pandemic.

### **1. Introduction**

The COVID-19 pandemic has had a devastating effect on global health, economies, and day-to-day activities. Understanding the trajectory of the pandemic is crucial to managing healthcare resources, informing public health policies, and preparing for future waves. One of the most critical challenges in pandemic management is predicting the future spread of the virus. Predicting COVID-19 cases with a high degree of accuracy could improve response strategies, enhance early warning systems, and support healthcare systems in preparing for potential surges. This project seeks to utilize machine learning models to predict global COVID-19 cases and analyze trends from historical data.

### **2. Data Preparation**

The dataset used in this analysis is sourced from Worldometer, a reliable source of global COVID-19 data. This dataset contains daily statistics on the number of confirmed cases, deaths, recoveries, and various other health-related indicators from countries across the globe.

#### **Data Cleaning**

Before applying machine learning techniques, the raw dataset required extensive cleaning:

* **Missing Values**: Missing values in numerical columns were handled by filling them with the median value of that column. Missing categorical columns were filled with the mode.
* **Data Transformation**: New features were derived, including:
  + **Mortality Ratio**: Deaths/Total Cases.
  + **Recovery Ratio**: Recoveries/Total Cases.
  + **Cases per Population**: Total Cases/Total Population.
  + **Deaths per 1 Million**: Total Deaths/Population (in millions).

#### **Feature Selection**

The following features were selected for use in model development:

* **Total Cases**: Total confirmed cases globally.
* **Total Deaths**: Total reported deaths.
* **Total Recovered**: Total number of recoveries.
* **Population**: Population of each country.
* **Testing Rate**: Number of tests per million people.
* **Continent**: Continent of the country.
* **GDP per Capita**: A measure of economic activity which could be related to healthcare system efficiency.

#### 

#### **Data Preprocessing**

* **Normalization**: Numerical features were scaled using Min-Max Scaling to ensure that they contribute equally to the machine learning model.
* **Encoding Categorical Data**: For categorical variables such as Continent, One-Hot Encoding was applied to create binary columns for each continent.

### **3. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was performed to understand the trends and relationships in the data before building the machine learning model.

#### **3.1. Visualizing Total Cases and Deaths by Continent**

A bar chart was used to visualize the total cases and deaths by continent. This highlighted the disproportionate number of cases and deaths in regions like Europe, North America, and South America, compared to Africa and Oceania.

#### **3.2. Correlation Heatmap**

A correlation heatmap was created to examine the relationships between various numerical features. For example, we found a strong positive correlation between Total Cases and Total Deaths, which was expected given that higher numbers of cases often lead to higher death tolls.

### **4. Feature Engineering**

Feature engineering was essential to improve the model’s predictive performance. New features were derived from the existing data to provide additional insights and better handle the complexities of COVID-19 dynamics.

#### **4.1. Mortality and Recovery Ratios**

* **Mortality Ratio**: The ratio of deaths to the total number of cases, calculated as Deaths / Total Cases.
* **Recovery Ratio**: The ratio of recoveries to the total number of cases, calculated as Recoveries / Total Cases.

#### **4.2. Growth Rate Features**

The daily growth rate of cases, deaths, and recoveries was calculated for each country, helping to understand the trends over time. The 7-day moving average was applied to smooth out daily fluctuations.

### **5. Model Development**

After data preparation and feature engineering, the next step was to develop a predictive model.

#### **5.1. Model Selection**

A **Random Forest Regressor** was chosen for this task. Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. This approach helps improve accuracy by reducing overfitting.

#### **5.2. Model Training**

* **Training/Test Split**: The data was split into training (80%) and testing (20%) sets. The training set was used to train the model, and the test set was used to evaluate its performance.
* **Hyperparameter Tuning**: Various hyperparameters, including the number of trees and the maximum depth of each tree, were tuned using cross-validation.

### **6. Model Evaluation**

The Random Forest model was evaluated based on multiple metrics:

#### **6.1. RMSE (Root Mean Squared Error)**

The RMSE was calculated to evaluate the difference between predicted and actual values. A lower RMSE value indicates better model accuracy.

#### **6.2. R² Score**

The R² score indicates how well the model explains the variance in the data. An R² score of 1 means the model perfectly explains the variance, while a score of 0 means it explains none of the variance.

**Results**: The model achieved an **R² score of 0.98**, which indicates that it explains 98.71% of the variance in the data. The RMSE was relatively low, suggesting that the model’s predictions were accurate.

### **7. Key Insights**

* **Global COVID-19 Trends**: The analysis revealed stark differences in the number of cases and deaths across continents, with Europe and North America having the highest numbers. Africa, in contrast, showed fewer cases and deaths.
* **Model Performance**: The Random Forest model provided good predictions with an R² score of 0.85. The model was particularly effective in identifying trends based on key features like population, recovery rate, and mortality rate.

### **8. Conclusion**

The results of this study show that machine learning models can effectively predict COVID-19 cases, enabling better forecasting and preparation for future outbreaks. The Random Forest model performed well, providing insights into the key factors influencing the spread of the virus. Future improvements could include incorporating real-time data, more granular features (such as vaccination rates), and other machine learning models.