**Predicting COVID-19 Death and Recovery Cases in Africa Using Linear Regression**

***Written by*** [***Irene Deodatus Mbeleka***](https://www.linkedin.com/in/irene-mbeleka/)

**Introduction**

The COVID-19 pandemic had a significant impact on the world, including Africa, where health infrastructure faced multiple challenges. This project was developed as the final assignment for my data science boot camp. Its goal is to analyze COVID-19 data from African countries to predict the number of deaths and recoveries based on the number of confirmed cases. By applying linear regression, the project aims to provide insights into the correlation between confirmed cases, fatalities, and recoveries. This analysis can help health authorities make more informed decisions.

**Project Objectives**

* To explore real-world COVID-19 data from Africa.
* To apply linear regression techniques to predict deaths and recoveries.
* To evaluate the model’s performance and interpret key insights.

**Data Collection and Preprocessing**

* **Data Selection:** The dataset contains COVID-19 statistics for each country, including confirmed cases, deaths, and recoveries. It was obtained from publicly available health reports.
* **Data Cleaning:** Missing or inconsistent values were addressed using imputation techniques, including filling gaps with mean or median values.
* **Exploratory Data Analysis (EDA):** Statistical and visual analyses were conducted using histograms, scatter plots, and correlation matrices to identify patterns and relationships among the variables.

**Model Development**

* **Feature Selection:** The number of confirmed cases was identified as the independent variable, while deaths and recoveries were considered dependent variables.
* **Data Splitting:** The dataset was split into training and testing sets, usually in an 80–20 ratio, ensuring that the model was trained on most of the data while keeping a portion for validation.
* **Linear Regression Implementation:** Using Python’s scikit-learn library, a simple linear regression model was developed to establish the relationship between confirmed cases and the number of deaths and recoveries. The model was trained on the dataset and fitted to a linear equation.

**Model Training and Evaluation**

* **Training the Model:** The model was trained using the training dataset, allowing it to learn the correlation between confirmed cases and deaths/recoveries.
* **Evaluation Metrics:** The model’s performance was measured using Mean Squared Error (MSE) to assess prediction accuracy and R-squared to determine how well the independent variable explained the variance in dependent variables.
* **Visualization of Results:** Scatter plots with regression lines were generated to visually compare the actual and predicted values, helping to identify trends and potential outliers in the dataset.

**Results and Discussion**

* The linear regression model demonstrated a strong correlation between confirmed cases and deaths/recoveries, meaning that as the number of cases increased, deaths and recoveries also followed a predictable pattern.
* However, discrepancies were observed, likely due to differences in healthcare quality, intervention strategies, and data inconsistencies across different African countries.
* The model performed reasonably well but had limitations in accounting for external factors such as government policies, hospital capacity, and vaccination campaigns.
* To improve accuracy, it was recommended to integrate more complex models and additional variables that influence death and recovery rates.

**Conclusion**

This project effectively employed linear regression to examine COVID-19 trends, offering valuable insights into the relationships among confirmed cases, deaths, and recoveries. While the model successfully identified key trends, it also highlighted areas for improvement in capturing the complexities of real-world data.

This experience was instrumental in enhancing my skills in data preprocessing, model training, and performance evaluation, providing a solid foundation for my growth in data science.

**Future Work**

* **Real-Time Analysis:** Implementing a dynamic, web-based tool to analyze COVID-19 data in real time would enhance practical applications for decision-makers and researchers.
* **Pandemic Preparedness Modeling:** Applying similar predictive models to analyze and forecast potential outbreaks of emerging diseases could help governments and healthcare institutions prepare better.
* **Integration with AI-Powered Diagnostics:** Combining machine learning predictions with AI-driven diagnostics from medical imaging or biosensors could enhance real-time decision-making during future health crises.