**Data Preparation/Feature Engineering**

**1. Overview**

Data preparation and feature engineering are critical phases in any machine learning project. They ensure the dataset is clean, relevant, and optimized for the algorithms used. Proper data preparation improves the model's accuracy and efficiency.

**2. Data Collection**

The datasets used include population data, GDP per capita, malaria cases, malaria deaths, and sanitation services for various African countries from 2014 to 2020. Preprocessing steps involved filtering data for the relevant years and African countries only.

**3. Data Cleaning**

Cleaning involved handling missing values, removing duplicates, and ensuring consistency in country names across datasets. Null values were filled with appropriate measures like mean, median, or zeros, depending on the dataset's nature.

**4. Exploratory Data Analysis (EDA)**

EDA involved visualizing trends and distributions of the key variables. Some visualizations included:

- Population Trends: Line plots showing population changes from 2014 to 2020.

- GDP per Capita: Bar charts to compare GDP per capita across different countries.

- Malaria Cases and Deaths: Heatmaps to show the prevalence of malaria in different regions.

- Sanitation Services: Pie charts or bar graphs to show the proportion of the population with access to sanitation services.

**5. Feature Engineering**

New features were created, such as the ratio of malaria cases to population, and transforming GDP data into growth rates. These features provide additional insights and improve model predictions.

**6. Data Transformation**

Data scaling and normalization were performed to standardize the range of features, ensuring all variables contribute equally to the model. Encoding categorical variables was also necessary.

**Model Exploration**

**1. Model Selection**

The choice of model was based on the nature of the data and the problem. For instance, regression models were chosen for predicting continuous outcomes like GDP per capita, while classification models were used for categorical outcomes like malaria presence.

**2. Model Training**

Models were trained using cross-validation techniques to ensure robustness. Hyperparameters were tuned using grid search methods to find the optimal configuration for each model.

**3. Model Evaluation**

Evaluation metrics included RMSE for regression models and accuracy, precision, recall, and F1 score for classification models. Confusion matrices and ROC curves were also used to assess performance.

**4. Code Implementation**

Relevant code snippets were included in the document to demonstrate data preparation, feature engineering, and model training processes. Comments in the code provided explanations of key steps.

**Please provide images from your data and models (I want to see different visualizations in EDA part as we did pair coding session )!!!!**

1. **Line plot for population over the years**

**A screen shot of a graph

Description automatically generated**

1. **Bar plot for GDP per capita**

**A graph with blue bars

Description automatically generated with medium confidence**

1. **Heatmap for malaria cases.**

**A table with numbers and symbols

Description automatically generated**

1. **Pie chart for sanitation services**

**A colorful circle with numbers and text

Description automatically generatedA colorful circle with numbers and text

Description automatically generated**