**Medical Statistics Analysis**

**Name Student ID**

Umair Imran L22-8370

Bilal Ahmad L22-7472

Muhammad Shahzad L22-7530

Najam Ul Islam L22-7497

**Problem and Data Set Description**

In analyzing country-wise data, a significant challenge arises in conducting collective analysis to determine which country requires specific resources. While individual country analyses provide valuable insights, performing a comprehensive analysis on a collective scale proves difficult.

This project addresses the challenge by predicting population growth rates and life expectancy on a larger scale. Using a drill-up approach, we will begin with country-level data and progress to continent and region-level analysis. The dataset spans from 2000 to 2024 and includes dimensions such as population, life expectancy, and trends in diseases across various countries.

The primary objective of this project is to offer a unified solution for collective analysis of health and resource needs across countries. It will also enable predictions of specific variables for individual countries or regions.

The dataset includes key dimensions such as population, life expectancy, trends of diseases over time, causes of mortality, and health expectancy rates for both genders. We will employ unsupervised learning techniques to effectively analyze and predict these factors.

# Preliminary ideas on how you plan to address it (models/algorithms/techniques)

As our data primarily consists of the numerical data as input and numerical data as output we would majorly use regression models including both the linear as well as multilinear regression models moreover for the classification part we would use the logistic regression models and as a supportive model we would use the k nearest neighbors model ,As our data also compromises the trends of population and resources of various countries across a range of years to we would apply LSTM’s for time series analysis .

**Software Tools**

We plan to use Python as our primary programming language. Below is a list of specific libraries and tools that will be employed throughout the project:

* **GitHub**: For efficient project management and version control.
* **Selenium**: As a data scraping tool for gathering required information.
* **Power BI**: For performing data visualization tasks.
* **Matplotlib**: To assist with data exploration and graphical analysis.
* **Pandas** and **NumPy**: For data manipulation, processing, and numerical computations.
* **Sklearn**: For implementing unsupervised learning algorithms and statistical modeling.

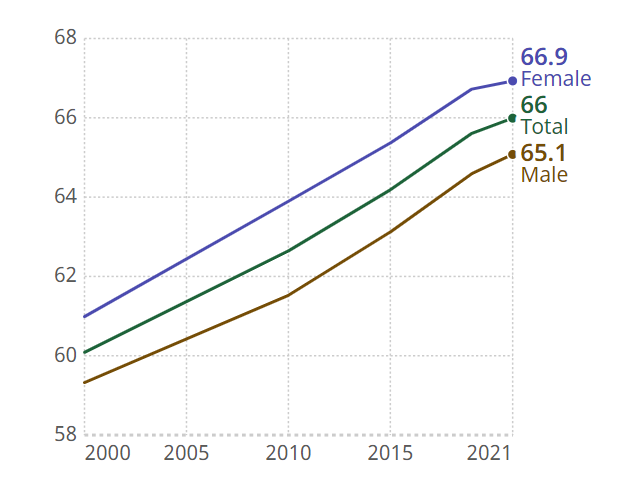
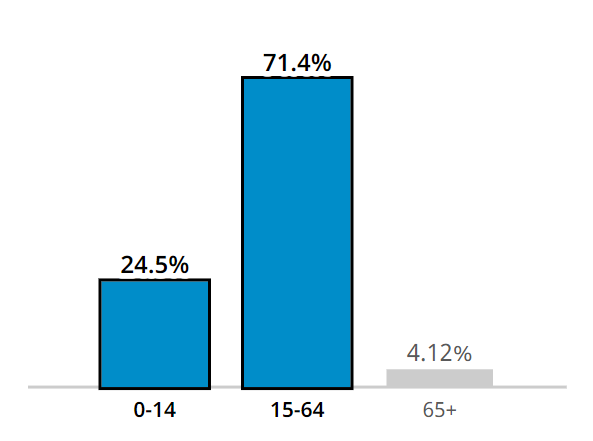
**Expected Results and Evaluation**

This project primarily focuses on regression tasks, supplemented by some classification tasks. We anticipate obtaining regression results that reflect various health and demographic statistics across all countries. For the classification component, we expect to classify countries into categories such as low income or high income, as well as to categorize countries based on specific disease prevalence.

To evaluate our regression results, we will use metrics including Mean Squared Error (MSE) and Root Squared Error (RSE). For the classification tasks, we plan to assess our results using several key performance metrics, including accuracy, sensitivity, specificity, and F1 scores. These evaluations will provide insights into the model's performance and its effectiveness in predicting health-related outcomes and demographic classifications.

**Preliminary Results and Data set explored.**

For this project, we have explored the data set scraped from WHO.com . Our dataset majorly contained almost 13000 rows for around 198 countries world wide . We performed initial EDA before any removal of outliers and nulls and found out that there exist a normal distribution among the countries in a specific region or falling in specific income level . Moreover regarding the population trend there was also a steep increase in populations as well as the life expectancy.



**Outline of the Work-to-Do**

1. **Dividing the Modules**
   * Assigning tasks and dividing the modules among all team members on GitHub.
2. **Data Scraping**
   * Scraping data from the major website (WHO) for all countries.
3. **Exploratory Data Analysis (EDA)**
   * Performing exploratory data analysis to understand the dataset.
   * **Major Techniques**: [e.g., Summary Statistics, Data Visualization, Correlation Analysis, Distribution Analysis]
4. **Feature Engineering**
   * Implementing feature engineering based on the following techniques:
   * **Major Techniques**: [e.g., Normalization, One-Hot Encoding, Binning, Polynomial Features]
5. **Data Visualization**
   * Using Power BI for data visualization to present insights effectively.
6. **Machine Learning Models**
   * Applying machine learning models to make predictions based on the data.
7. **Model Evaluation**
   * Conducting an iterative evaluation of models to refine performance.
8. **Application Development**
   * Converting the project into a real-world application using Streamlit and Flask (tentative).
9. **Chatbot Development**
   * Developing a chatbot that can retrieve data from the WHO in real time based on user queries for better analysis (tentative).

# Reference

1. https://data.who.int/countries/: A Complete website of global scale medical data open source