Predictive Medical Analysis

Team Members

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

As our data primarily consists of numerical data as input and output, we will predominantly use regression models, including both linear and multilinear regression models. For the classification part, we will use logistic regression models, and as a supportive model, we will use the k-nearest neighbors (KNN) model.

Additionally, because our data includes trends in population and resources across various countries over time, we will apply Long Short-Term Memory (LSTM) models for time series analysis.

Software Tools

We plan to use Python as our primary programming language. The following libraries and tools will be employed throughout the project:

- **GitHub**: For project management and version control.
- **Selenium**: For data scraping.
- **Power BI**: For data visualization.
- **Matplotlib**: For data exploration and graphical analysis.
- **Pandas and NumPy**: For data manipulation, processing, and numerical computations.
- **Scikit-learn (Sklearn)**: For implementing Machine learning algorithms and statistical modeling.

Expected Results and Evaluation

This project will primarily focus on regression tasks, supplemented by classification tasks. We anticipate producing regression results that reflect various health and demographic statistics across all countries. For the classification component, we expect to categorize countries based on income levels or disease prevalence.

Evaluation Metrics:

- **Regression**: Mean Squared Error (MSE) and Root Squared Error (RSE).
- **Classification**: Accuracy, sensitivity, specificity, and F1 scores.

These metrics will provide insights into model performance and effectiveness in predicting health-related outcomes and demographic classifications.

Preliminary Results and Dataset Exploration

We have explored the WHO's website and a dataset with approximately 10,000 rows of data for 198 countries can be created by scraping. Initial exploratory data analysis (EDA) revealed a normal distribution among countries in specific regions or income levels. Additionally, we observed a steep increase in populations and life expectancy trends across the dataset.

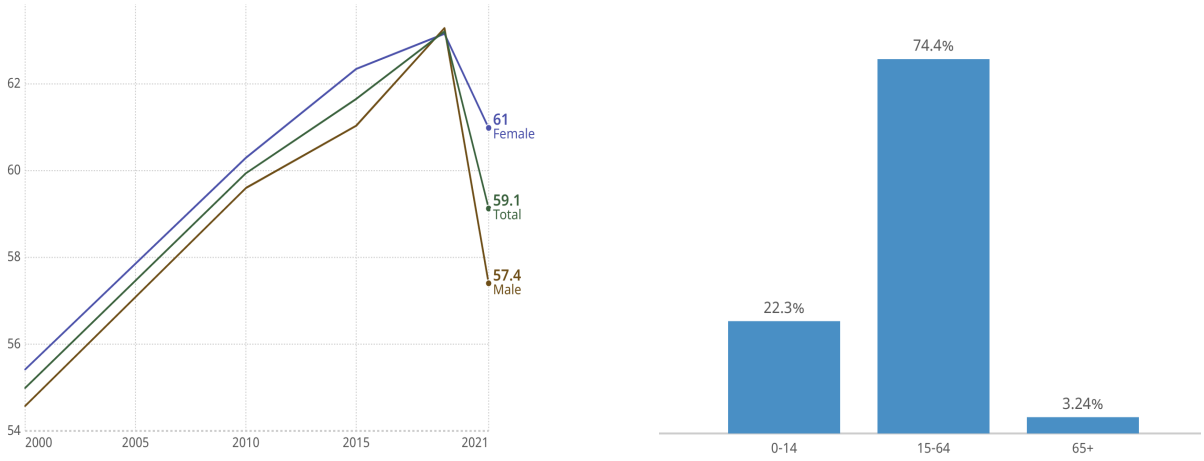

Figure 1: Graph 1 and Graph 2 showing population trends and life expectancy

Outline of the Work-to-Do

- **Dividing the Modules**: Assigning tasks and dividing the modules among all team members on GitHub.
- **Data Scraping**: Scraping data from WHO for all countries.
- **Exploratory Data Analysis (EDA)**: Understanding the dataset using summary statistics, data visualization, correlation analysis, and distribution analysis.
- **Feature Engineering**: Normalization, One-Hot Encoding, binning, and polynomial feature generation.
- **Data Visualization**: Using Power BI to present insights effectively.
- **Machine Learning Models**: Applying machine learning models for prediction.
- **Model Evaluation**: Iterative evaluation of models to refine performance.

Reference

- `https://data.who.int/countries/`: A complete website of global scale medical data (open-source).