Health Data Analysis

Milestone: Project proposal

Group: Data Dynamos
Student 1: Muskan Sharma
Student 2: Hemalikaa Thirumavalavan

778-869-1739 (Tel of Student 1)
437-217-5392 (Tel of Student 2)

sharma.muskan@northeastern.edu
thirumavalavan.h@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%
Percentage of Effort Contributed by Student 2: 50%
Signature of Student 1: Muskan Sharma
Signature of Student 2: Hemalikaa Thirumavalavan
Submission Date: 23 January 2025

Problem Setting: The health sector receives extensive study including the identification of health determinants across chronic health issues like diabetes blood pressure and obesity. The health care costs continue to go up while more people around the world suffer from health problems linked to their daily choices. The dataset under consideration offers a set of various demographic, behavioural and health-oriented assessment metrics of the population of 25 countries. In fact, the problems associated with the incidence and interaction of such factors as smoking, alcohol consumption and physical activity have not been solved even with the progress made in medical research. These challenges further emphasize the importance of getting back to the fundamentals of applied data analyses for identification of recommended interventions, maintenance/improvement of population health, and elimination/minimization of variations in care delivery.

Problem Definition: This study aims to address the following questions through data mining techniques:Major demographic and behavioral features of the dataset relating to obesity and diabetes are as follows:Examining the relationship between some of the broad-reaching patterns of daily living and biochemical values of cholesterol and blood pressure: smoking, alcohol use, and exercise.Following on from the above, are there clear variations in health outcomes (e.g., obesity and diabetes existence) between different nations?

Which determinant factors have the highest odds ratios by disease, including high cholesterol and high blood pressure?

Is there anything that can be recommended from the patterns obtained from this dataset with regard to public health?

Data Sources:
The health dataset employed for this assessment is a random dataset developed specifically for educational and analysis purposes. It also encompasses data obtained from numerous countries with the members of various populations and health-related conditions. The dataset was collected from [https://www.kaggle.com/datasets/ankushpanday1/heart-attack-risk-predictions] with favourable emphasis on the proper uses that have been permissible according to the trends on data privacy.

Data Description:
The dataset contains the following characteristics:

- Number of Columns: 30
- Number of Rows: 623028
- Variable Names: Country, Age, Gender, Cholesterol, Blood Pressure, Smoking, Drinking, Exercise, Obesity, Diabetes status, Family History, Stress Levels, Diabetic habits, heart attack history, chest pain, Exercise-Induced Angina, resting ecg, maximum heart rate achieved, Thalassemia, HDL Cholesterol, LDL Cholesterol, Triglycerides, Heart attack risk, medication adherence, Urbanisation level, air pollution exposure, access to healthcare, education level, income level, heart attack outcome.

Sample Variables:

- Country of Residence: Representing 25 countries.

- Age: Range from 18 to 90 years.
- Gender: Half and half (50% Male and 50% Female).
- Cholesterol Level: Ranges from 150 to 300 mg/dL.
- Smoking History: Never: 33%, Former: 33%, and Current: 33%.
- Physical Activity Level: Sedentary (33%), Moderate (33%), Active (33%).

Summary Statistics:

- Gender distribution: 50% Male, 50% Female.
- Smoking history: balanced in the given categories.
- Key Health Conditions: Obesity and diabetes are modelled as binary dummy variables (1/0 = Yes/No).

Challenges in the Data:

- Some gaps in date or date differences may be a kind of missing data in the dataset.
- Differences in people's health that are clustered by countries may be different and hence need adjustment.
- The extent of information that may encapsulate the social determinants of health that may affect the health outcomes is also limited.

Methodology:

The application of the respective research questions will involve the use of data mining techniques to analyse the given set of data. The proposed methodology includes:

1. **Data Preprocessing:**

- Describing how to transform numeric variables, how to handle missing values, and outliers.
- Data transformation in order to be able to compare countries.
- It also involves encoding the categorical variables if and only if you see the need to.

2. **Exploratory Data Analysis (EDA):**

- Exploratory data visualization to get a first impression of distributions of the primary variables.
- Looking for a relationship between demographic, behavior, and health data and performances.

3. **Pattern Recognition:**

- Our research groups similar individuals with their specific traits.
- Data mining to generate common ALERMS, where ALERM stands for Association of Large, Extensive, Related and Multiple risk factors.

4. **Predictive Modeling:**

- By applying decision trees and logistic regressions to our dataset features we aim to differentiate between obesity and diabetes conditions.

- We evaluate different models and interpret how well they perform compared to standard measures of accuracy precision recall.

5. **Geographical Analysis:**

- To investigate regional disparities in the provision of preventive and curative services among the 25 countries in the European Union.
- The findings of the analysis will be purposefully designed to help with the formulation of recommendations for interventions in public health. Activities are planned to be done in a recursive manner until the most suitable techniques are identified from the exploratory results and the nature of the given dataset.