Health Determinants and Chronic Disease Analysis

Milestone: Final 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: 6 Feburary, 2025

Introduction

Background

Healthcare facilities worldwide experience an increasing problem with chronic diseases including diabetes and hypertension and obesity. Healthcare expenses continue to rise while numerous people face healthcare risks stemming from their life choices despite medical advancements. The research analyzes health determinants through data from 25 countries to generate practical insights which support public health actions and enhanced healthcare results and minimized care delivery inequalities.

Problem Statement

Lifestyle diseases have become more common, so health professionals need data-driven solutions to recognize essential risk variables as well as national differences and prevention methods. This study will focus on:

a. Healthcare professionals need to recognize the primary behavioral and population characteristics linked to obesity and diabetes.
b. The analysis evaluates the relationships which form between lifestyle choices like smoking habits alcohol habits and exercise and the biochemical markers of blood pressure and cholesterol measurement.
c. A study of health outcome and care accessibility disparities that exist among different countries at geographical locations.
d. Determining which factors at a population level produce the highest statistic values for high cholesterol and high blood pressure conditions.
e. Health professionals should introduce evidence-based public health interventions through analysis of data.

Data Sources and Description

Data Sources

The research draws its data from Kaggle https://www.kaggle.com/datasets/ankushpanday1/heart-attack-risk-predictions?resource=download where educators and analysts developed the information for educational and analytical purposes. The dataset comprises thorough health-related information which comes from 25 different countries.

Number of Columns: 30
Number of Rows: 623,028

Data Description

Key Variables:

a. The study includes data about six demographic aspects: Country of Origin, Age Group, Gender, Education Level, Income Level and Urbanization Level.
b. The five behavioral elements include smoking alongside alcohol consumption and physical activity together with stress evaluation and diabetic behavioral patterns.
c. Health Metrics: Cholesterol, Blood Pressure, Obesity, Diabetes Status, Heart Attack History, Medication Adherence.
d. The dataset contains information about air pollution exposure as well as healthcare accessibility.

Challenges in the Data

- Some variables contain missing data that needs replacement before analysis.
- Standardization is necessary because various healthcare systems present different levels of variability between countries.
- Limited representation of social determinants of health.

Research Questions and Objectives

Research Questions

1. What demographic characteristics along with behavior patterns play crucial roles in the occurrence of obesity combined with diabetes?
2. Lifestyle behaviors influence which biochemical markers in the body indicate health condition.
3. Chronic disease occurrence shows substantial variations between different nations.
4. Who or what exercises the most influence in causing high cholesterol levels and blood pressure elevations?

The research provides which information could guide public health leadership in developing intervention programs and approaches.

Objectives

- The initial step includes a complete review of data through exploratory data analysis (EDA) to gain insight into variable distributions together with relationship patterns.
- To recognize risk patterns both clustering methods and statistical correlations should be implemented.
- Authors must create predictive models that help identify high-risk subjects.

- Use geographic analysis methods to display variations between healthcare result areas.
- The team should develop policy suggestions to enhance population health status.

Methodology and Analysis

Methodology

1. **Data Preprocessing**
    a. Handling Missing Values:
        1. Mean/median imputation for numerical variables.
        2. Mode-based imputation for categorical variables.
    b. Standardization and Normalization:
        1. The research normalizes BP and cholesterol values for international comparison of results.
    c. Encoding Categorical Variables:
        1. One-hot encoding for nominal variables (e.g., country, gender).
        2. Label encoding serves as the method to transform ordinal variables including education level and income level.
    d. Outlier Detection and Removal:
        1. Identification of extreme values was possible through the combination of Z-scores and IQR methods.
2. **Exploratory Data Analysis (EDA)**
    a. The investigation uses histograms and boxplots and summary statistics during univariate analysis.
    b. The analysis includes Bivariate Analysis through correlation heatmaps together with scatterplots and grouped bar charts.
    c. The research employs Principal Component Analysis (PCA) as a multivariate technique to decrease dimensionality while finding hidden patterns.
3. **Pattern Recognition**
    a. Clustering: K-Means and Hierarchical Clustering to group individuals with similar health profiles.
    b. Association Rules allows extraction of mutual risk factors by utilizing Apriori and FP-Growth algorithms.

Predictive Modeling

1. **Classification Models:**

    a. Logistic Regression (baseline model for diabetes and obesity prediction).
    b. The model uses Decision Trees to display important disease-driving factors.
    c. Random Forest (for improved predictive accuracy).
    d. Support Vector Machines serve to separate different risk categories.

e. Gradient Boosting Models (for advanced prediction performance).

2. **Model Evaluation Metrics:**

a. Accuracy (Overall model performance).
b. Precision together with Recall and F1-Score measure how effective the model is at identifying disease cases.
c. ROC-AUC Curve (Model discrimination power).
d. Confusion Matrix (True Positive/False Positive analysis).

3. **Geographic Analysis**

The application of choropleth maps helps display national health variations in different countries.

Regional Disparities: Identifying areas with high disease burden and limited healthcare access.

Healthcare System Impact: Assessing the effect of infrastructure and policies on health outcomes.

Expected Outcomes

1. The research examines the principal elements that affect obesity together with diabetes.
2. Research examines how smoking behavior and alcohol consumption as well as physical movement affect significant health indicators.
3. A comprehensive comparison of health outcomes across different nations.
4. A robust predictive model for early disease detection.
5. Data-driven recommendations for policymakers and healthcare professionals.

Tools and Technologies

1. **Programming Language:** Python (pandas, NumPy, scikit-learn, matplotlib, seaborn, geopandas)
2. **Data Visualization:** Tableau, Power BI
3. **Statistical Analysis:** Python
4. **Geospatial Analysis:** Python

Conclusion

A data-centered analysis of chronic disease determinants in 25 different countries uses the present study. The study uses three methods including exploratory analysis and predictive modeling alongside geographic assessment to raise public health effectiveness and healthcare availability standards. The research outcomes will help authorities develop fundamental policies with resource allocation strategies that fight both diabetes and obesity across the world.