Machine Learning Model For Diabetes, Heart Attack, and Stroke Prediction

Final Project Proposal
CMPS320 - Prof. Tweneboah
Bijay Adhikari; Merit Kayastha; Nischal Bhandari; Saurav Dahal

Table of Contents

a) Introduction and Data Source.....1
b) Research Questions.....2
c) Exploratory Data Analysis & Data Preprocessing.....2
d) Machine Learning Models & Techniques.....3
e) Tasks for each group member.....3

a) Introduction and Data Source

Diabetes is a chronic health condition that occurs when the blood glucose is too high. According to the CDC, diabetes is the 8th leading cause of death in the United States, and 1 in 5 of the adults with diabetes in the US do not even know that they have it. If unmanaged, diabetes can lead to various complications including heart disease and stroke, neuropathy, compromised immune system, and kidney damage.

Our dataset comes from Kaggle (https://www.kaggle.com/datasets/prosperchuks/health-dataset/data), which has been compiled from the Behavioral Risk Factor Surveillance System (BRFSS) 2015. BRFSS is a collaborative project between all the states in the United States and the Centers for Disease Control and Prevention (CDC) designed to measure behavioral risk factors for the adult population (aged 18 years of age and older) residing in the United States. Our dataset is a collection of 70,692 survey responses with 17 feature variables and 1 target variable.

The features of our dataset are as follows:

1) **Age:** 13 different age categories starting from (1=18-24 yrs).
2) **Sex**
3) **HighChol:** 0 for no high cholesterol, 1 for high cholesterol.
4) **CholCheck**: 0 for no cholesterol check in 5 yrs, 1 for yes.
5) **BMI**
6) **Smoker:** 0 for not smoked 5 packs of cigarettes in lifetime, 1 for yes.
7) **HeartDiseaseorAttack**: 0 for no, 1 for yes.
8) **PhysActivity:** in the past 30 days, 0 for no and 1 for yes.
9) **Fruits:** 1 for fruit consumed 1 or more times a day, 0 for no.
10) **Veggies**: 1 for vegetables consumed 1 or more times a day, 0 for no.
11) **HvyAlcoholConsump:** 0 for no, 1 for yes.
12) **GenHlth:** 1 =excellent, 2=very good, 3=good, 4=fair, 5=poor.
13) **MentHlth**:1-30 (days of poor mental health in a month).
14) **PhysHlth**: 1-30
15) **DiffWalk**: serious difficulty while walking or climbing stairs (0 for no, 1 for yes).
16) **Stroke:** 0 for no, 1 for yes.
17) **HighBP**: 0 for no, 1 for yes.
18) **Diabetes:** 0 for no, 1 for yes.

b) Research Questions

Some of the preliminary research questions that the team is considering to address with our dataset are listed below:

1. What are the key predictors of diabetes, hypertension, and stroke?
2. How do the factors interact with each other in influencing the likelihood of these health conditions?
3. Can we develop a predictive model that accurately identifies individuals at risks of these conditions?
4. What is the impact of lifestyle factors (like smoking, alcohol consumption, physical activity) on the likelihood of these health conditions?
5. Are there significant differences in the prevalence of these health conditions across different demographic groups (like age, gender)?

c) Exploratory Data Analysis & Data Preprocessing

- Handling missing values and categorical data: We will first check for any missing values in the dataset and get rid of the row. We will also handle categorical data (sex) into binary representations. Outliers can disproportionately affect machine learning algorithms, and thus we will remove or transform them to minimize their effect.

- Feature selection and feature engineering: Our dataset contains 17 feature variables, and all of them may not play considerable roles in predicting diabetes of an individual. So, we will explore the possibility of reducing the features in the dataset using techniques such as Principal Component Analysis (PCA) or feature selection methods. We will also look at the correlation matrix to see if two or more features in the dataset are collinear. In such a case, we will combine existing features to improve model performance.

- Feature Scaling: We will also standardize our data to ensure that the machine learning algorithms like support vector machine (SVM) and K-Nearest Neighbor (KNN) are effective.

d) Machine Learning Models & Techniques

The dataset has labeled data for our target variables, *Diabetes, Strok, and HeartDiseaseorAttack*. So most of the techniques we will be using to answer our research questions listed above will be using supervised Machine Learning models.

However, to get the overview of our data, we will use dimensionality reduction method (PCA). In addition to the dimensionality reduction, we will also remove our target variables and look for the clusters within the dataset using K-means or hierarchical clustering.

After obtaining the information on the dataset, we will use the following methods listed in an order to match our research questions mentioned above:

1. **Methods for #1**: We will use correlation analysis and possibly more advanced feature selection methods available in scikit-learn (like Best Subset Selection) to identify the key predictors.

2. **Methods for #2**: Decision Trees and Random Forests will be used to capture complex interactions between different predictors.

3. **Methods for #3**: We will experiment with different machine learning models such as logistic regression, decision trees, random forests, and gradient boosting to build a predictive model in our dataset. We will use the ensemble of different combinations of classifiers to get the maximum accuracy of our predictions.

4. **Methods for #4**: Logistic Regression will be used to understand the impact of individual predictors like lifestyle factors on the likelihood of a binary outcome.

5. **Methods for #5**: In order to check for the significant differences in the prevalence of these health conditions across different demographic groups (like age, gender), we will use statistical tests available in scipy to validate our findings.

e) Tasks for each group member

Bijay Adhikari: Exploratory Data Analysis & Data Processing
Merit Kayastha: Research Question 1 & 2
Nischal Bhandari: Research Question 3 & 4
Saurav Dahal: Research Question 5 and Recommendations and conclusions