Data Mining-Project Proposal

Adrià Irigaray, Camilla Tartaglia, Esteban Tortosa, Jordi Catafal
Group 32

Research question and goal

This project intends to find an answer to the following question: Which health and Circumstantial factors influence the most on cases of diabetes and pre-diabetes?
Diabetes is a chronic condition that affects many millions of people around the world and therefore, it is truly important to be aware of the causes that might lead to a person having such a disease.
Type 1 and Type 2 diabetes can lead to organ failure, neuropathy or cardiovascular diseases, and an early prevention of a pre-diabetes (where the sugar levels are high but not yet classifiable as diabetes) can help reduce the effects and risks of an advanced diabetes.
By studying the factors that could play a part in developing a full-blown diabetes, we could come to shed some light on possible measures and habits that could be scaled and given to a general population of people in order to prevent in some degree the appearance of diabetes.
Understanding these influences and this correlation between the components could provide a better focus for higher organizations to serve as a general guideline for educating people with the aim of avoiding this kind of disease, which would lead to a mitigation of the impact of diabetes and the complications aforementioned that such illness implies.

Dataset description

The data comes from the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the CDC.
Key health and circumstantial factors include:

- Indicator of whether the individual has no diabetes, prediabetes, or diabetes.
- Indicator of whether the individual has high blood pressure.
- Indicator of high cholesterol levels.
- Whether the individual has had a cholesterol check in the past 5 years.
- Body Mass Index, a measure of body fat based on height and weight.
- Whether the individual has smoked at least 100 cigarettes in their lifetime.
- Indicator of whether the individual has had a stroke.

- Indicator of coronary heart disease or myocardial infarction.
- Whether the individual has engaged in physical activity in the past 30 days (excluding job-related activity).
- Daily consumption of fruits and vegetables.
- Indicator of heavy alcohol consumption (more than 14 drinks/week for men, and more than 7 for women). Circumstantial and lifestyle factors:
- Indicator of whether the individual has any form of healthcare coverage.
- Whether the individual was unable to see a doctor in the past 12 months due to cost.
- General, mental, and physical health measures based on self-reported data.
- Whether the individual has difficulty walking or climbing stairs.

Demographic variables:

- Sex
- Age category, ranging from 18-24 to 80 or older.
- Education level, ranging from no formal education to college graduate.
- Income level, ranging from less than $10,000 to 75,000 or more.

Data mining approach

To accomplish the proposed task and attempt to retrieve an answer for the project,we will begin by preprocessing the collected data to address any potential missing values or outliers and ensure that the input data is in the correct format for analysis. Following this, we plan to apply a sampling method to reduce the number of data points. Next, we will perform Exploratory Data Analysis (EDA) to uncover patterns within the dataset, using correlation and association indices to identify relationships between variables.
After examining multicollinearity between features and scaling the numerical features, if necessary, we may also apply dimensionality reduction techniques, such as Principal Component Analysis (PCA), to further streamline the dataset and enhance the clustering process. Once the data is preprocessed, we will encode categorical variables and employ **K-Means clustering** combined with the elbow method or use a hierarchical clustering algorithm to determine the optimal number of clusters (K). This approach will allow us to uncover patterns and group similar data points without relying on predefined labels.