Diabetes risk factors analysis

Bertrand Flanet

Introduction

We were provided with a 2015 dataset from the Behavioral Risk Factor Surveillance System (BRFSS) that records various behaviors, conditions, and symptoms for U.S. individuals with or without diabetes. For this study, the dataset was cleaned and reduced to 70,692 responses, evenly split between those with diabetes and those with prediabetes or no diabetes. Our goal was to identify the most predictive risk factors for diabetes. Given the time constraints, we conducted a brief exploratory data analysis (EDA), categorized features using domain knowledge, performed a Chi-square test for initial selection, and refined our results with a Random Forest model.

1. Exploratory Data Analysis (EDA)

We began by assessing data integrity and confirmed the dataset was in good condition. Our univariate analysis revealed a balanced population with minimal outliers. By splitting the population into those testing positive and negative for diabetes or prediabetes, we identified potential risk factors through overlapping visualizations.

In a subsequent step, We created a heatmap to calculate Pearson correlations and identify relationships between features and diabetes diagnosis. Although most features showed weak correlations, some aligned with our previous observations. We retained features with a correlation above absolute 0.20: Income, Age, Difficulty Walking, Physical Health, General Health, Heart Disease or Attack, BMI, High Cholesterol, and High Blood Pressure.

However, it is important to interpret these results with caution, as they denote correlation rather than causation.
The retained features can be classified into categories with varying influences on diabetes diagnosis.
From domain knowledge, some features are considered symptoms rather than conditions. Conditions can be further broken down into physiological and socio-economic categories.
We classified our features into the following categories:
Physiological Conditions, Socio-economic Conditions, and Symptoms
These distinctions will help us to interpret our result in a nuance manner.

2. Feature Selection Methodology

Chi-Square Statistical Test

Before running our statistical model, we excluded features classified as "symptoms" and some "socio-cultural conditions" since they reflect physiological consequences and may skew results due to redundant intel. We then applied a Chi-square test to identify statistically significant features related to diabetes. The test highlighted eight key features: high cholesterol, heavy alcohol consumption, physical activity, income, and age, with smoking habits, BMI, and education also identified as potential, though lesser, risk factors.

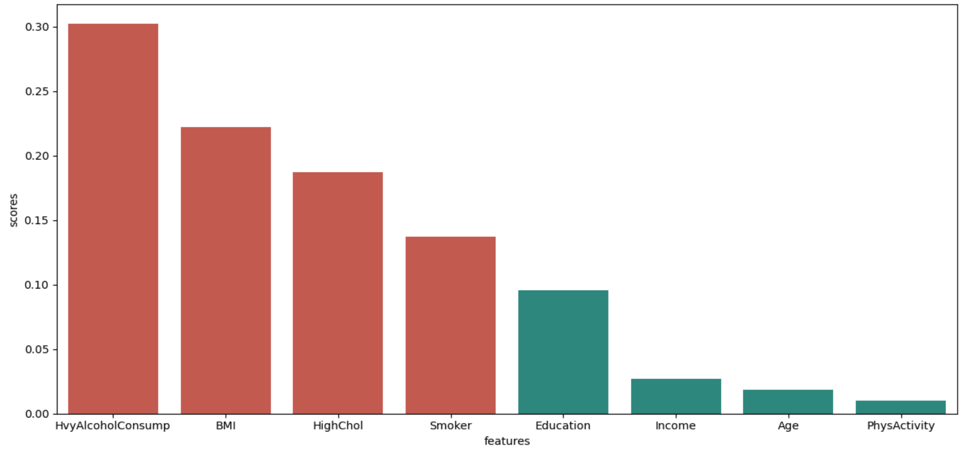
Random Forest Model

Following the Chi-square selection, we refined our feature set using a Random Forest model to assess feature importance. This method further evaluated the significance of the remaining features.
The Random Forest model identified four main features in order of importance: heavy alcohol consumption, BMI, high cholesterol, and smoking habits. The remaining features—education, age, income, and physical activity—were classified as less influential.

Additionally, even though correlations were found between certain features and diabetes status, they do not necessarily indicate direct causation.

Mitigating Results and Considering False Negatives

After identifying potential risk factors, we analyzed a sample of the population that tested negative but had not seen a doctor in the past year due to cost and lacked healthcare coverage. We sought to determine if this group might be at risk or falsely negative for diabetes. We found that this population exhibited higher risk factor metrics compared to the negative group with healthcare access, though cholesterol levels were similar. Notably, heavy alcohol consumption was higher in the at-risk group, suggesting that alcohol might be a risk factor. Once positively diagnosed, individuals may consume less alcohol due to health concerns.



3. Results Interpretation

By combining domain knowledge with statistical and model-based feature selection, we identified four potential risk factors for diabetes.
However, these results should be interpreted cautiously.
First, no single feature stands out as a clearly distinguishable predictor, suggesting that diabetes risk is likely multifactorial.
The relationships and interactions between features may be more significant than the individual features themselves. Future analyses should explore feature combinations that could be more predictive of diabetes.
Among the four remaining features, three fall into the "Physiological Conditions" category, highlighting the connection between individual health habits and diabetes diagnosis. However, each of these habits may be influenced or conditioned by other factors, raising questions about how socio-economic conditions affect physiological conditions and potentially lead to diabetes.

Conclusion

There are elements of our approach that could be refined. The sample population used in this study may not be optimal for such classification and selection, as it was limited to a single year (2015). The BRFSS study began in 1984, so there is more data available that could help us track individual behaviors and pathologies over time.
Observing the evolution of individuals' health might yield more fruitful results by allowing us to compare behaviors before and after diabetes diagnosis while comparing features of individuals who remain healthy.
Recognizing the multifactorial nature of diabetes, other factors such as environmental components (e.g., temperature affecting blood sugar and insulin levels), pollutants, or genetic contributions, which are already known to influence human physiology, should also be considered.