Project Proposal - Predicting Diabetes

Course Number: ORIE 5741
Github link: https://github.com/karenSashimi/predicting-diabetes
Members: Chun Wang (cw954), Karen Kuang (yk499), Jiahao Li (jl4363)

Introduction and Problem Statement

Background Information
Diabetes is one of the widespread challenges that impacts millions of lives across the United States. As a potentially life threatening disease, diabetes affects how our bodies handle sugar, leading to a range of serious health issues, from heart disease to heart problems. Recognizing and addressing this disease in the early stage could significantly improve the quality of life for the patients.

Problem Statement
This project aims to identify individuals at high risk of developing diabetes based on a comprehensive analysis of health indicators, lifestyle factors, and the broader context of their living environments. Through the analysis and constructing predictive modeling, we seek to bring a personalized approach to diabetes prevention, or at least ensure early detection and awareness, making it possible to offer advice and interventions.

Importance of the Problem
From the perspective of patients, diabetes is a condition affecting millions worldwide and affecting not only the patient oneself, but also their families, communities and healthcare systems. Early detection and prevention of diabetes can possibly save their lives and significantly reduce the financial burden associated with the long-term management of its complications.

From a corporate perspective, addressing diabetes early detection is not just a health imperative but a strategic investment, representing an opportunity to lead in innovation and corporate social responsibility. The predictive approach aligns with growing customer demand and enhances brand loyalty and trust.

Dataset effectiveness
The Behavioral Risk Factor Surveillance System dataset provides a comprehensive platform for analyzing the factors influencing diabetes due to its extensive coverage of risk behaviors and health-related variables among a large adult population. With over 279 variables and data collected from 464,644 subjects across various states, the dataset offers a rich and diverse pool of information. Factors such as tobacco use, exercise habits, health status, access to healthcare, chronic health conditions, and alcohol consumption are among the many variables assessed. Additionally, the dataset is carefully weighted, ensuring the accracy of demographic representation, bias reduction and analysis reliability. By leveraging this dataset, our team can explore model training and fitting, and dive deep into the feature correlations and identify potential causal relationships between these features, offering insights for diabete prevention.

Dataset Overview

We obtained access to the 2014 Behavioral Risk Factor Surveillance System (BRFSS) dataset, comprising 464,644 subjects and encompassing 279 variables. (https://www.cdc.gov/brfss/annual_data/annual_2014.html) The BRFSS was established in 1984, initially involving 15 states conducting monthly telephone interviews to gather surveillance data on risk behaviors. BRFSS collects consistent, state-specific information on preventive health practices and behaviors associated with chronic diseases, injuries, and preventable infectious diseases affecting the adult population.

The BRFSS dataset is weighed and has minimized bias stemming from unequal probability of selection. Additionally, the dataset utilizes iterative proportional fitting to account for demographic variations between the sampled individuals and adjusts for discrepancies in demographics to ensure the sample accurately reflects the broader population.

Within this dataset, 61,118 respondents were diagnosed with diabetes, 12,699 with prediabetes, and 390,827 had neither condition. The majority of variables were related to various chronic health conditions, such as cancer and asthma, aside from diabetes. Type 2 diabetes classification was based on age (>30 years), non-pregnancy status, and affirmative response to the survey questions.

Methodology

In our proposed study, we aim to leverage machine learning techniques to build risk prediction models for Type 2 Diabetes (T2D) by analyzing dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2014. Our approach will encompass several stages, starting with data preprocessing to handle missing values and normalize the dataset, ensuring it is adequately prepared for model training and testing. Given the success of various machine learning models in identifying risk factors and predicting T2D as outlined in the literature, we plan to employ a diverse set of algorithms including Support Vector Machines (SVM), Decision Trees, Logistic Regression, Random Forests, Gaussian Naive Bayes classifiers, and Neural Networks. This multi-model strategy will allow us to compare their performance in terms of accuracy, sensitivity, specificity, and Area Under the Curve (AUC) metrics, with a particular focus on achieving high sensitivity for early detection of at-risk individuals.

Moreover, we will explore the use of some data rebalancing methods to address any imbalances in the dataset, which is crucial for improving the predictive performance of our models given the relatively lower prevalence of T2D among the population. Through univariable and multivariable weighted logistic regression analyses, we aim to investigate associations between potential risk factors and T2D, adjusting for covariates to identify both well-established and novel risk factors. The ultimate goal of our project is not only to enhance the predictive accuracy of T2D risk models but also to contribute to public health strategies by enabling early diagnosis and intervention, thereby reducing the prevalence and impact of this chronic condition.