Abstract

1. Introduction

The growing prevalence of Type 2 Diabetes and its precursor, prediabetes, represents a significant and growing public health challenge in the United States that has touched many, including my family. Prediabetes is characterized by elevated blood glucose levels that do not yet meet the threshold for type 2 diabetes. It affects millions of individuals, some whom don't even know of their prediabetes status, placing them at heightened risk for developing the full-blown disease and associated co-morbidities. Early and accurate detection is critical for timely lifestyle interventions, which align with current medical guidance to delay or prevent disease progression.

To address the need for improved diabetes screening tools, my project uses machine learning classification techniques to predict an individual's risk of having prediabetes or undiagnosed diabetes. The analysis uses data from the National Health and Nutrition Examination Survey (NHANES), a program operated by the Centers for Disease Control and Prevention (CDC) that assesses the health and nutritional status of adults and children across the United States. This study uses publicly available NHANES Body Measures (BMX) and Plasma Fasting Glucose (GLU) data to train and test the created models.

The project explores the potential effects of the COVID-19 pandemic on prediabetes risk factors. Given that the pandemic period encouraged widespread sedentary behaviors and altered dietary patterns due to fears and restrictions, this project focuses on the mid- to post-pandemic pandemic, from August 2021 to August 2023. By analyzing data, the project aims to identify and classify patterns associated with increased risk of prediabetes, thereby providing timely insights into the shifts in population health following a major global event. The resulting models will serve as tools for identifying individuals who may want further clinical evaluation to determine their diabetes status.

2. Methods

2.1 Data Preprocessing

The variables of interest in the project come from two separate NHANES datasets during the August 2021 to August 2023 period: GLU_L and BMX_L.

1. *Merge Datasets*

The datasets were first matched based on the respondent ID using the column "SEQN" found in both datasets and then merged. BMX_L originally had 8,860 observations, and only matched 3,996 rows to the GLU_L dataset.

2. *Feature Selection*

The merged dataset originally contained 25 columns. A majority of columns were dropped due to irrelevance to the analysis or redundancy, including the following:

- SEQN: Respondent sequence number
- WTSAF2YR: Sampling Weight
- LBDGLUSI: Fasting blood glucose in SI units
- Comment Codes: BMIWT, BMIRECUM, BMIHEAD, BMIHT, BMILEG, BMIARML, BMIARMC, BMIWAIST, BMIHIP

This initial feature cleaning reduced the final dataset only to include relevant body measurement and glucose columns.

3. *Handle Missing Values*

Since the LBXGLU (Fasting Glucose in mg/dL) feature is central to defining the dependent variable (prediabetes/undiagnosed diabetes status), rows with missing values for this feature were removed. After this process, the cleaned dataset contained 3672 observations.

4. *Define the Target Variable*

Based on established medical thresholds for fasting glucose and their corresponding classifications of diabetes status, the LBXGLU column was used to create a new binary target variable, "prediabetes_or_undiagnosed", based on the following clinical guidelines.

- Normal: Fasting Glucose < 100 mg/dL
- Prediabetes/Diabetes: Fasting Glucose ≥ 100 mg/dL

The target variable is recorded as:

- 0: Normal (LBXGLU < 100 mg/dL)
- 1: Prediabetes/Diabetes (LBXGLU ≥ 100 mg/dL)

Binary classification allows the model to focus on identifying the population at risk (Group 1) that requires clinical follow-up.

5. *Scale the Features*

Prior to model training, the numerical features were scaled to prevent variables with larger magnitudes from influencing the learning process. The `StandardScalar` from the scikit-learn library was used. Z-score normalization here standardizes features by removing the mean and scaling to unit variance. The scaler fit was only on the training data. The resulting scaled parameters were then applied to the testing set to ensure that the test data remained unseen for the evaluation of the final model.

2.2 Modeling