Capstone 1 Milestone Report

Predicting Prediabetes and Diabetes with Machine Learning

National Health and Nutrition Examination Survey (NHANES) Data 2007 – 2016

Introduction

The Centers for Disease Control and Prevention estimated in 2017 that 9.4% of the U.S. population had diabetes and nearly a quarter of those were undiagnosed. Just over a third of U.S. adults had prediabetes in 2015 and only 11.6% of them had been told by their doctor.

Damage to body systems begins with prediabetes, as risk of heart attack, stroke and kidney disease increase even before blood sugar reaches diabetic range. Diabetes is the 7th leading killer of Americans, taking 80,000 lives each year and contributing to the loss of over 250,000 more. It is the leading cause of blindness, amputation, end-stage kidney disease, and causes about a two to four-fold increased risk of heart attack and stroke (Brannick, et al., 2016; Centers for Disease Control and Prevention, 2017).

There are huge benefits to the population and to the health care system to increasing the number of people with prediabetes and diabetes who become aware of their condition and begin treatment. Prediabetes can be reversed with diet and lifestyle change. Diabetes can be managed, and side effect risk dramatically reduced by controlling blood sugar level. First you must know you have the diagnosis.

Objective

The goal of this project is to develop models using NHANES data to predict prediabetes and diabetes, then look at the main factors influencing those models to see if any new insights can be found to apply to existing screening tools.

Data

This study uses NHANES data collected between 2007 and 2016. NHANES stands for National Health and Nutrition Examination Survey and comprises demographic, socioeconomic, dietary and health related information. It also includes physical examination measurements and laboratory tests conducted by medical personnel. NHANES is a long-term health survey conducted by the National Center for Health Statistics, which is a part of the Centers for Disease Control and Prevention (CDC). The survey collects information from about 5,000 persons each year. Ten years of data yields around 50,000 observations.

Data Process

The NHANES survey is not static. Survey data and questions are reviewed and changes made between two-year survey cycles. Data elements are sometimes moved from one file to another as the survey is modified. NHANES data are made available in small files, each containing data relating to one topic for one two-year survey cycle. 314 files were downloaded for this analysis from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx. The files were SAS transport format and were first converted to CSV format for processing.

Data for survey cycles were combined by topic. Several required special handling due to elements being moved during the ten-year period. String data also required special handling as the Python package used to convert SAS transport format to CSV format did not wrap strings in quotes, causing unpredictable results. Null heatmaps were generated to visualize missing data. Columns that did not span the entire ten years were dropped. Data was combined into six categories and saved in CSV format files. The categories are nearly identical to those on the NHANES website with one addition:

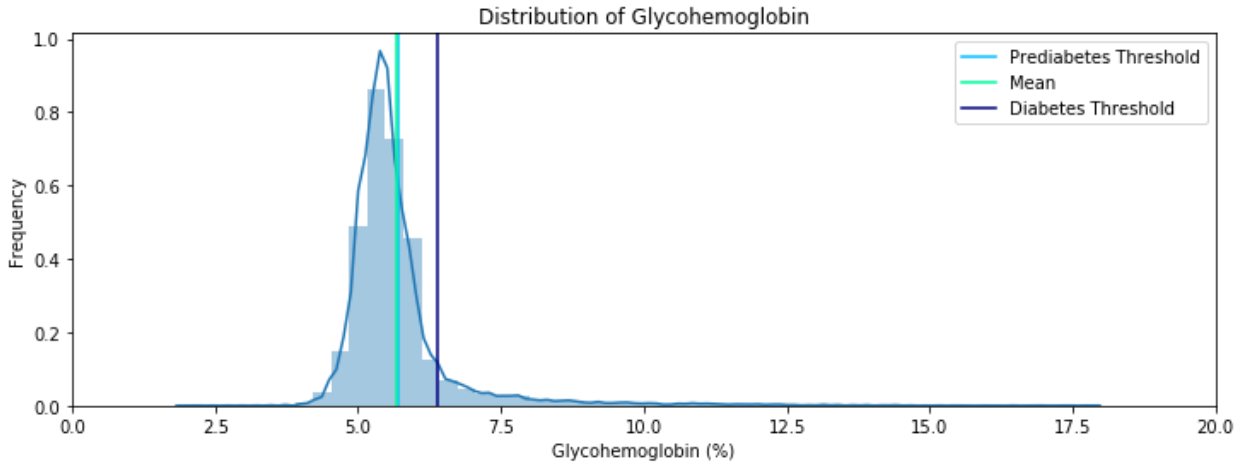
- Demographics: demo.csv
- Dietary intake: diet.csv
- Physical examination: exam.csv
- Laboratory test results: labs.csv
- Questionnaire answers: ques.csv
- Prescription medications (from Questionnaire category): meds.csv

The prescription data was split from the questionnaire data as there is one row for each prescription medication taken by an individual. Separating it simplified processing.

Exploratory Data Analysis

Initial exploration was directed at statistics and known risk factors for prediabetes and diabetes. Would the study data reflect the statistics and trends reported by the American Diabetes Association (ADA)? (American Diabetes Association, 2017)

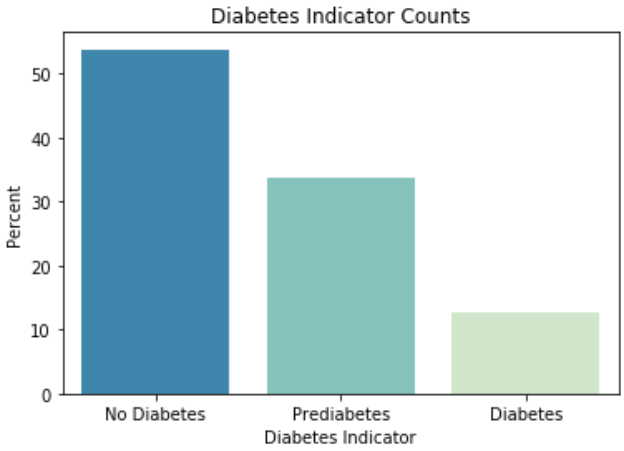
The graph below shows the distribution of glycohemoglobin or HbA1c, a lab value used to diagnose prediabetes and diabetes. Interestingly, the mean is equal to the threshold for diagnosing prediabetes!



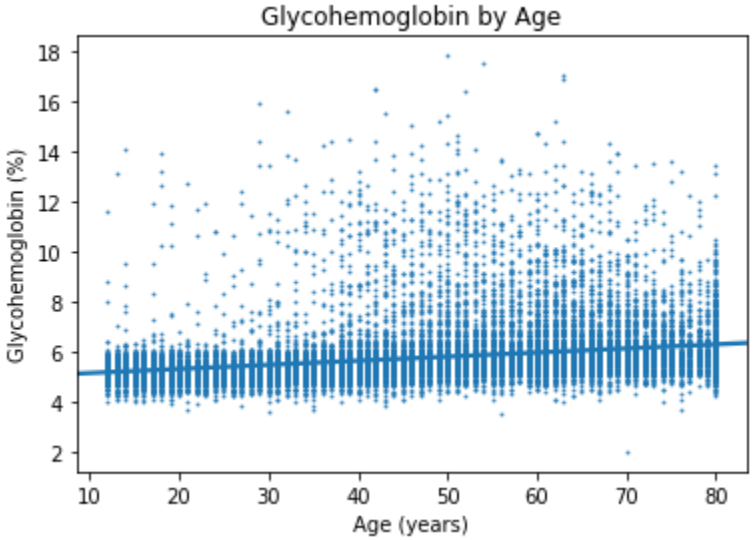
Also of note is the right skew of many of the lab results, with long right tails. This pattern is clearly illustrated in the glycohemoglobin plot. These outliers are very high lab test results and are valid values.

Below is the distribution of diabetes categories in the study data. Lab values used to diagnose prediabetes and diabetes were evaluated according to ADA diagnostic criteria to categorize study participants (American Diabetes Association, n.d.). The rate of prediabetes was very close to that reported by the ADA. The rate of diabetes was a bit higher than the reported rate, most likely due to

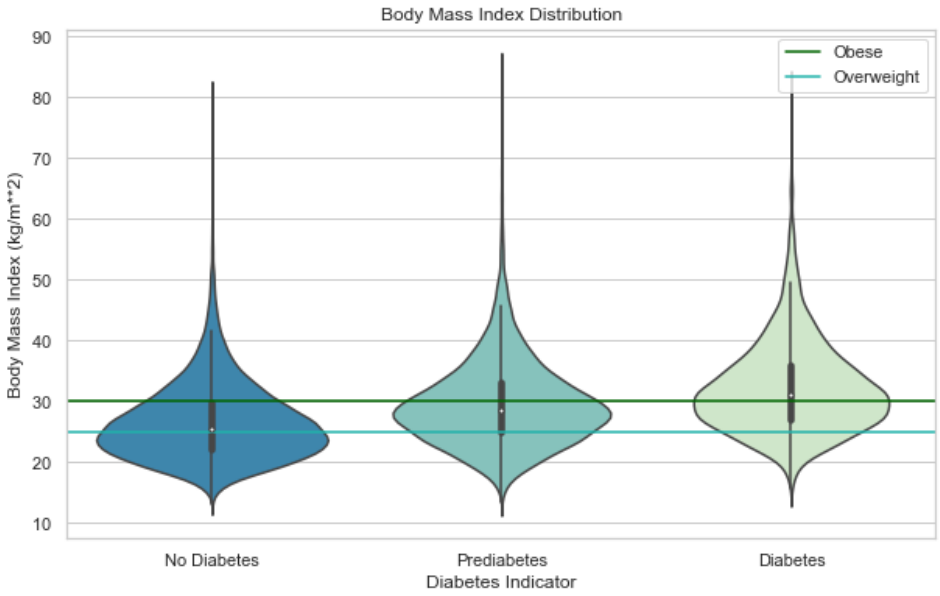
NHANES oversampling of higher risk categories (over 60, African American, Hispanic). This chart shows unweighted proportions so does not adjust for oversampling.



Below is a chart showing the relationship between glycohemoglobin and age. There is a clear trend upward with increasing age, which is in sync with the increased risk of diabetes. Type 2 diabetes was originally called adult-onset diabetes but is now frequently being diagnosed in teens and children. The chart below includes ages 12-80 and shows several teens well into diabetic range.



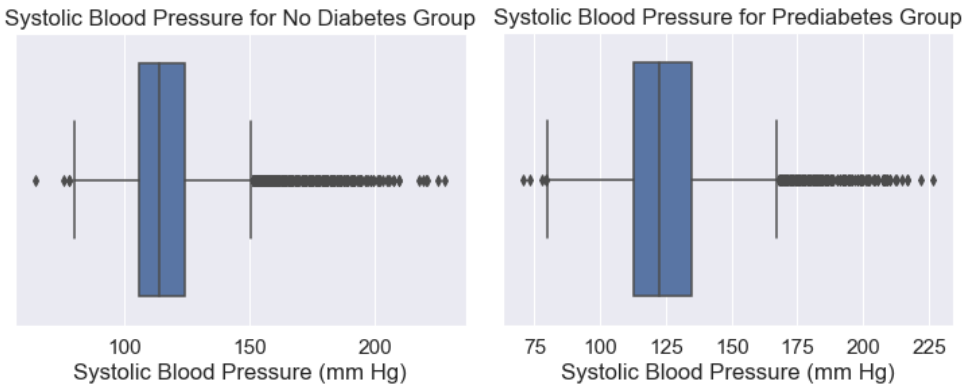
Finally, the following chart shows the distribution of Body Mass Index (BMI) for the three study categories. There is a clear trend toward higher BMI as we progress from healthy blood sugar to diabetic range. Note that the median BMI of the diabetes category is just above the obese BMI threshold. Even the non-diabetic category median is at the overweight threshold.



Statistical Analysis

Various statistical tests were run to determine whether relationships exist in the data between the study categories and known risk factors as well as some factors not recognized as risk factors. All tests resulted in finding for a statistically significant relationship between the tested factor and study categories. It should be noted that due to the complex design of NHANES, adjustments to standard statistical tests are recommended. Python statistics packages do not have the capability to account for such complex survey designs so these statistical analyses may report p-values that are smaller than would be calculated were we able to account for the survey design.

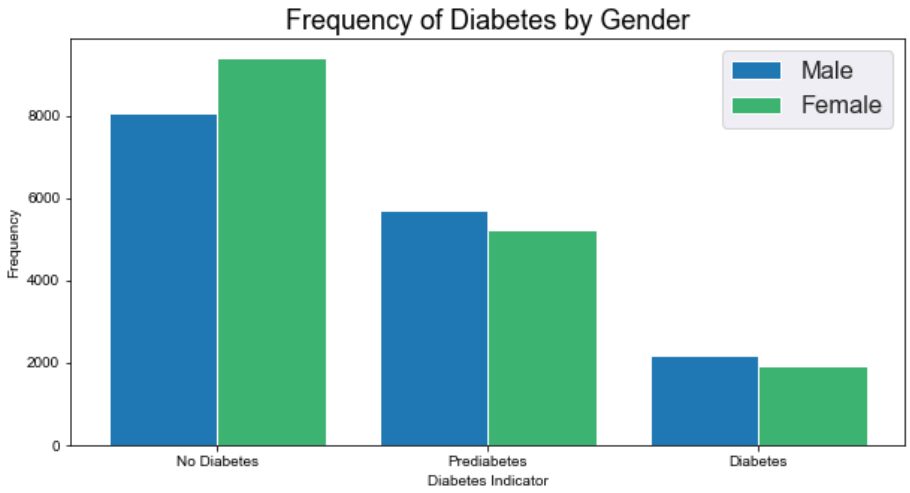
A student's t-test was run comparing the mean values of systolic blood pressure for the no diabetes and prediabetes groups. High blood pressure is a known risk factor for diabetes. The null hypothesis was no difference between the means, the alternative a statistically significant difference.



The t-statistic was -40, the p-value 0.0. The null hypothesis was rejected and we conclude that systolic blood pressure means are statistically different between non-diabetic and prediabetic categories.

A chi-squared test for independence was run on gender and diabetes category. The null hypothesis was no relationship, the alternative a statistically significant relationship. Gender is not a known risk factor for diabetes, although some studies show different risk of complications and different rates of diabetes (Kautzky-Willer, et al., 2016).

The χ^2 value was 131 with a p-value near zero (10^{-29}). This gave evidence against the null hypothesis, and we conclude that gender and diabetes category are related.



Summary of Statistical Tests

Chi-square Tests	**Result**
Gender relationship to diabetes category	Statistically significant relationship (χ^2 = 131; p < 0.0001)
Hypertension diagnosis relationship to diabetes category	Statistically significant relationship (χ^2 = 3269; p = 0.0)
Moderate activity relationship to diabetes category	Statistically significant relationship (χ^2 = 381; p < 0.0001)
Marital status relationship to diabetes category	Statistically significant relationship (χ^2 = 17; p = 0.0002)

T-Tests	**Result**
Means of triglycerides of no diabetes and prediabetes categories	Statistically different (t = -21; p < 0.0001)
Means of systolic blood pressure of no diabetes and prediabetes categories	Statistically different (t = -40; p = 0.0)
Means of body mass index of no diabetes and prediabetes categories	Statistically different (t = -36; p < 0.0001)
Means of serum iron of no diabetes and prediabetes categories	Statistically different (t = 10; p < 0.0001)
Means of asparate aminotransferase (AST) of no diabetes and prediabetes categories	Statistically different (t = -7; p < 0.0001)

References:

1. Centers for Disease Control and Prevention. (2017) *National Diabetes Statistics Report.* American Diabetes Association. https://dev.diabetes.org/sites/default/files/2019-06/cdc-statistics-report-2017.pdf
2. Brannick, B., Wynn, A., & Dagogo-Jack, S. (2016) Prediabetes as a toxic environment for the initiation of microvascular and macrovascular complications. *Experimental Biology and Medicine*.
3. American Diabetes Association. (n.d.) *Diagnosis.* https://www.diabetes.org/a1c/diagnosis
4. Kautzky-Willer, A., Harreiter, J., & Pacini, G. (2016). Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. *Endocrine reviews, 37(3), 278–316.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890267/