Tracy Cardwell
September 1, 2019

Capstone 1 Project Proposal

Predicting Prediabetes with Machine Learning

Type 2 diabetes affects over 30 million Americans – about 9.4% of the total population. At least 7 million of those are undiagnosed and unaware they have the disease according to the American Diabetes Association. 1.5 million Americans are diagnosed with diabetes every year, with 90-95% of all diabetes cases being type 2. An estimated 84 million or 33.9% of Americans have prediabetes, but only 11.6% or 28.8 million know they do.

Diabetes is the 7th leading cause of death in the United States, taking nearly 80,000 lives each year and contributing to the loss of over 250,000 more. Complications from diabetes include heart disease, stroke, kidney damage, nerve damage, blindness and amputation. Prediabetes increases a person's risk for type 2 diabetes. Studies have also shown it also increases risk for heart attacks and kidney damage, even before it has progressed to type 2 diabetes. Treating diabetes in America cost over $245 billion in 2012.

Type 2 diabetes is thought to be caused by a combination of genetics and lifestyle. It can be controlled with diet, exercise and in some cases, medication. Prediabetes can often be reversed with diet and exercise, thus avoiding a costly and high-risk disease. Reducing the number of prediabetes and type 2 diabetes cases will save billions for insurance companies, reduce the death toll of type 2 diabetes, and improve the quality of life of millions of Americans.

Objective

I propose to use the National Health and Nutrition Examination Survey (NHANES) data from 2013-2014 to predict prediabetes. The NHANES data contains demographic and socioeconomic data as well as lab test results and physical examination results. I will examine the data and look for features that correlate to lab markers indicating prediabetes per the American Diabetes Association: glycohemoglobin values in the range 5.7% - 6.4%, fasting blood glucose in the range 100-125 mg/dl, or an oral glucose tolerance test of 140-200 mg/dl.

Several screening tools exist to predict prediabetes. I will evaluate the NHANES data to determine if there are additional factors that can improve the predictive models and help more people have a chance at a healthier life. Reducing the number of Americans with prediabetes will significantly reduce costs for three groups: medical insurance companies, the uninsured, and the hospitals who treat uninsured emergency patients without reimbursement.

The data is available at https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey.

Approach

First, I will evaluate the completeness of the data and determine a plan to manage missing data. I will then use exploratory data analysis and inferential statistical analysis to investigate which of the hundreds of factors, including known risk factors, have a correlation with prediabetes. I will look at demographic data such as gender, age, race, education level, marital status, income level; physical examination data such as blood pressure, pulse, waist size, body mass index, dental health; lab test data such as cholesterol, albumin, creatinine, blood count values, liver function values; and questionnaire data such as activity level, alcohol use, self-perceived diabetes risk, family history of diabetes, gestational diabetes, overweight baby at birth, hysterectomy, hormone intake, food security and daily hours of TV watching.

Next, I will develop machine learning models to predict prediabetes based on the identified features. My target variable will be a new feature I define using the criteria mentioned previously: glycohemoglobin values in the range 5.7% - 6.4%, fasting blood glucose in the range 100-125 mg/dl, or an oral glucose tolerance test of 140-200 mg/dl. I will evaluate the models and select the most accurate.

Finally, I will look at why the model predicted as it did.

Deliverables

The following will be delivered on a GitHub repository:

- Jupyter notebook including all my code,
- Final report, and
- Powerpoint slide presentation.

References

Centers for Disease Control and Prevention. Diabetes Report Card 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2018, Retrieved from https://www.cdc.gov/diabetes/pdfs/library/diabetesreportcard2017-508.pdf

Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2017, Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

Prediabetes, Retrieved from https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278

Statistics About Diabetes (2015), Retrieved from https://www.diabetes.org/resources/statistics/statistics-about-diabetes

Type 2 Diabetes, Retrieved from https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193