Diabetes Prediction model: NHANES data 2013-2014

Hannah Rosenblum, James Ng, Purnima Sharma

Contents

Introduction

This project aimed to study any association between diabetes and several covariates in participants, ages 12 and older, using NHANES data. Effort was made to build an optimal prediction model to identify key variables within the dataset for a categorical outcome of diabetes, objective being a binary classification model with supervised learning. Certain factors of special interest were any association with participant's race, age, cholesterol and fasting glucose levels, among others. Specifically, association was assessed between diabetes and plasma fasting glucose, body mass index, blood pressure, hours of fasting, weekly workload, cholesterol, triglycerides, insulin levels, 2 hour glucose tolerance, glucose challenge and other demographic factors such as age, gender, race, using the NHANES data for the year 2013 - 2014 from the cdc.gov website, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013 .

Motivation was provided by the fact that diabetes is one of the major leading causes of death in the United States. As stated by the CDC site's National Diabetes Statistics Report of 2020, 34.2 million Americans are diabetic, while 7.3 million were undiagnosed. With prevalence of diabetes and prediabetes on the rise, it was of interest to find factors that might affect the diabetes status. Later years post-2013 were tried for the data, however were unavailable for the variables of interest possibly due to continuing updates.

After extracting and merging the necessary files by participant's Id number, variables of interest were retained in a dataframe, and saved as a dataset file within the project for easy access. Gender, Race and the response variable Diabetes were converted to factors from numeric data type. Missing entries for the response of diabetes status were removed. 185 "borderline " reported cases, 5 with "don't know" responses and 1 with "refused" response were also removed given the small scale of these categories, which accounted for less than 2% of the data, and in order to focus on the majority of binary responses of presence or absence of diabetes. The cleaned data contained 9,578 observations of 15 variables, including the binary outcome variable diabetes.

EDA

. . .

Data Frame Summary

raw_data
Dimensions: 9578 x 15
Duplicates: 506

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. 1 2. 2	4706 (49.1%) 4872 (50.9%)	IIIIIIII IIIIIIIII	0 (0.0%)
2	age [numeric]	Mean (sd) : 32.4 (23.9) min < med < max: 1 < 28 < 80 IQR (CV) : 41 (0.7)	80 distinct values	: : : : : . : : : : : : : : : . : : : : : : : : : :	0 (0.0%)
3	race [factor]	1. 1 2. 2 3. 3 4. 4 5. 6 6. 7	1616 (16.9%) 893 (9.3%) 3449 (36.0%) 2148 (22.4%) 1033 (10.8%) 439 (4.6%)	III I IIIIIII IIII II	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
4	bmi [numeric]	Mean (sd) : 25.6 (7.9) min < med < max: 12.1 < 24.6 < 82.9 IQR (CV) : 10.4 (0.3)	436 distinct values	: . : : : : : : : : . : : : : .	706 (7.4%)
5	fasting_hrs [numeric]	Mean (sd) : 10.9 (3.5) min < med < max: 0 < 11 < 34 IQR (CV) : 3 (0.3)	32 distinct values	: : : : : : . : : : :	6374 (66.5%)
6	workload [numeric]	Mean (sd) : 209.5 (3989.3) min < med < max: 1 < 40 < 99999 IQR (CV) : 15 (19)	90 distinct values	: : : : :	6312 (65.9%)
7	hdl [numeric]	Mean (sd) : 53.2 (15.2) min < med < max: 10 < 51 < 173 IQR (CV) : 19 (0.3)	116 distinct values	: : . : . : : : : : : .	2128 (22.2%)
8	ldl [numeric]	Mean (sd) : 106 (34.9) min < med < max: 14 < 103 < 375 IQR (CV) : 46 (0.3)	194 distinct values	: . : : : . : : : . : : : .	6553 (68.4%)
9	triglyceride [numeric]	Mean (sd) : 111.7 (115.9) min < med < max: 13 < 88 < 4233 IQR (CV) : 73 (1)	344 distinct values	: : : : :	6515 (68.0%)
10	insulin [numeric]	Mean (sd) : 13.4 (18.7) min < med < max: 0.1 < 9.3 < 682.5 IQR (CV) : 9.1 (1.4)	1716 distinct values	: : : : :	6567 (68.6%)
11	hr2_glucose_tolerance [numeric]	Mean (sd) : 114 (45.5) min < med < max: 40 < 104 < 604 IQR (CV) : 44 (0.4)	227 distinct values	: : : : : : : : : :	7294 (76.2%)
12	glucose_challenge [numeric]	Mean (sd) : 6 (3.4) min < med < max: 1 < 5 < 74 IQR (CV) : 3 (0.6)	26 distinct values	: : : : : :	7232 (75.5%)
13	glucose_tolerance [numeric]	Mean (sd) : 118.4 (6.4) min < med < max: 42 < 117 < 183 IQR (CV) : 7 (0.1)	42 distinct values	: : : : : : .	7293 (76.1%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
14	glucose [numeric]	Mean (sd) : 104.6 (31.2) min < med < max: 51 < 98 < 421 IQR (CV) : 15 (0.3)	184 distinct values	: : : . : : : : .	6490 (67.8%)
15	diabetes [factor]	1. yes 2. no	737 (7.7%) 8841 (92.3%)	I IIIIIIIIIIIIIIII	0 (0.0%)