Diabetes Prediction model: NHANES data 2013-2014

Hannah Rosenblum, James Ng, Purnima Sharma

Contents

Introduction

This project aims to study any association between diabetes and several covariates in participants ages 1 and older, using NHANES data, and selecting an optimal prediction model among linear, non-linear, parametric and non-parametric models. The main objective is building a binary classification model with supervised learning. Certain factors of special interest were any association with participant's race, age, cholesterol and lifestyle factors, among others. Data was extracted for the year 2013 - 2014 from the cdc.gov website, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013.

Data description

Specifically, association was assessed between diabetes and the following covariates:

- Gender: Participant's gender (male or female)
- Age: Age at screening, with possible values of 0 to 79, or 80+ (years)
- Race: 6 categories for race include Mexican American, other Hispanic, White, Black, Asian and other.
- bmi: body mass index (kg/mˆ2)
- hdl: High-density lipoprotein (mg/dL)
- Blood pressure (mm Hg): Both systolic and diastolic, first-round measurements
- waist: Waist circumference measurement (cm)
- Sedentary activity (lifestyle, minutes): time spent sitting in a given day, not including sleeping.
- Education level: highest degree of adults 20+ years of age, with 7 categories.
- Marital status: Categories include married, widowed, divorced, separated, never married, living with partner, refused, and don't know
- Depression: severity on a scale of 0 to 3 treated as a continuous variable, with 0 as not at all depressed
- Sleep: amount of sleep in hours on a given night on weekdays or workdays

The outcome of "diabetes" dependent-variable was based on classification of the participants into two groups of those with diabetes and those who did not have diabetes. Individuals answered the question "other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?", and were classified as having diabetes if they answered yes.

Motivation

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. Furthermore, increase in type 2 diabetes among children is a growing concern according to the CDC. 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.

Data cleaning

After extracting and merging the necessary files by participant's Id number, variables of interest were retained in a dataframe. Gender, race, education level, marital status 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 dataset contained 9,578 observations of 18 variables, including the binary outcome variable diabetes.

EDA

Exploratory data analysis was performed using density plots and bar graphs. Summary statistics were analyzed for all variables, to get an overview of the data and check for extent of missing values. Density plots were used to check for relationships between diabetes and other numeric variables. Categorical variables were visualized separately, using bar graphs instead.

summary statistics

Data Frame Summary

raw_data
Dimensions: 9578 x 18
Duplicates: 319

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. 1 2. 2	4706 (49.1%) 4872 (50.9%)	IIIIIIIII 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%)
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	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%)
6	ldl [numeric]	Mean (sd) : 106 (34.9) min < med < max: 14 < 103 < 375 IQR (CV) : 46 (0.3)	194 distinct values	: . : : : . : : : . : : : .	6553 (68.4%)
7	triglyceride [numeric]	Mean (sd) : 111.7 (115.9) min < med < max: 13 < 88 < 4233 IQR (CV) : 73 (1)	344 distinct values	: : : : :	6515 (68.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	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%)
9	glucose [numeric]	Mean (sd) : 114 (45.5) min < med < max: 40 < 104 < 604 IQR (CV) : 44 (0.4)	227 distinct values	:	7294 (76.2%)
10	bp_systolic [numeric]	Mean (sd) : 117.9 (18) min < med < max: 66 < 116 < 228 IQR (CV) : 20 (0.2)	71 distinct values	:	2571 (26.8%)
11	bp_diastolic [numeric]	Mean (sd) : 65.7 (15) min < med < max: 0 < 66 < 122 IQR (CV) : 16 (0.2)	59 distinct values	:	2571 (26.8%)
12	waist [numeric]	Mean (sd) : 86.9 (22.5) min < med < max: 40.2 < 87.4 < 177.9 IQR (CV) : 31.6 (0.3)	1030 distinct values	:	1091 (11.4%)
13	lifestyle [numeric]	Mean (sd) : 478.5 (642.1) min < med < max: 0 < 480 < 9999 IQR (CV) : 300 (1.3)	36 distinct values	:	2625 (27.4%)
14	education [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 7 7. 9	442 (7.9%) 761 (13.6%) 1261 (22.6%) 1715 (30.7%) 1406 (25.1%) 2 (0.0%) 5 (0.1%)	I II IIII IIIIII IIIII	3986 (41.6%)
15	married [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 77 8. 99	2866 (51.3%) 419 (7.5%) 637 (11.4%) 170 (3.0%) 1096 (19.6%) 401 (7.2%) 2 (0.0%) 1 (0.0%)	IIIIIIIII I II III I	3986 (41.6%)
16	depression [numeric]	Mean (sd) : 0.4 (0.8) min < med < max: 0 < 0 < 9 IQR (CV) : 0 (2.1)	0 : 3955 (75.5%) 1 : 876 (16.7%) 2 : 205 (3.9%) 3 : 194 (3.7%) 7 : 2 (0.0%) 9 : 3 (0.1%)	IIIIIIIIIIIIII III	4343 (45.3%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
17	sleep [numeric]	Mean (sd) : 7 (3.2) min < med < max: 2 < 7 < 99 IQR (CV) : 2 (0.5)	12 distinct values	: : : : :	3300 (34.5%)
18	diabetes [factor]	1. yes 2. no	737 (7.7%) 8841 (92.3%)	I IIIIIIIIIIIIIIII	0 (0.0%)

Data Frame Summary

raw_data
Dimensions: 9578 x 18
Duplicates: 319

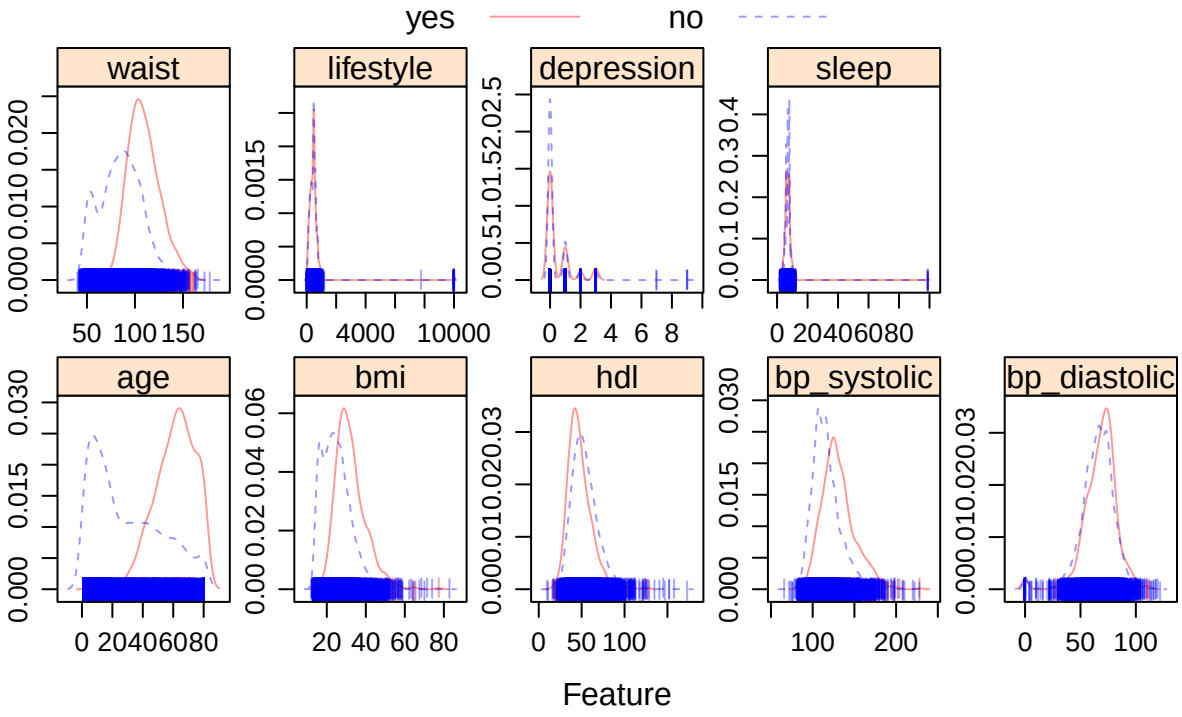
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. 1 2. 2	4706 (49.1%) 4872 (50.9%)	IIIIIIIII IIIIIIIII	0 (0.0%)
2	age [numeric]	Mean (sd) : 32.4 (23.9) min < med < max: 1 < 28 < 80 IQR (CV) : 41 (0.7)	80 distinct values	: : : : : . : : : : : : : : : . . : : : : : : : : : : :	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%)
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	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%)
6	ldl [numeric]	Mean (sd) : 106 (34.9) min < med < max: 14 < 103 < 375 IQR (CV) : 46 (0.3)	194 distinct values	: . : : : . : : : . : : : .	6553 (68.4%)
7	triglyceride [numeric]	Mean (sd) : 111.7 (115.9) min < med < max: 13 < 88 < 4233 IQR (CV) : 73 (1)	344 distinct values	: : : : :	6515 (68.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	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%)
9	glucose [numeric]	Mean (sd) : 114 (45.5) min < med < max: 40 < 104 < 604 IQR (CV) : 44 (0.4)	227 distinct values	: : : : : : : : : :	7294 (76.2%)
10	bp_systolic [numeric]	Mean (sd) : 117.9 (18) min < med < max: 66 < 116 < 228 IQR (CV) : 20 (0.2)	71 distinct values	: : : : : . : : . : : : : .	2571 (26.8%)
11	bp_diastolic [numeric]	Mean (sd) : 65.7 (15) min < med < max: 0 < 66 < 122 IQR (CV) : 16 (0.2)	59 distinct values	: : . : : : : : : : : : : :	2571 (26.8%)
12	waist [numeric]	Mean (sd) : 86.9 (22.5) min < med < max: 40.2 < 87.4 < 177.9 IQR (CV) : 31.6 (0.3)	1030 distinct values	: . : : : . : : : : : : : : : : : : : : : : .	1091 (11.4%)
13	lifestyle [numeric]	Mean (sd) : 478.5 (642.1) min < med < max: 0 < 480 < 9999 IQR (CV) : 300 (1.3)	36 distinct values	: : : : :	2625 (27.4%)
14	education [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 7 7. 9	442 (7.9%) 761 (13.6%) 1261 (22.6%) 1715 (30.7%) 1406 (25.1%) 2 (0.0%) 5 (0.1%)	I II IIII IIIIII IIIII	3986 (41.6%)
15	married [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 77 8. 99	2866 (51.3%) 419 (7.5%) 637 (11.4%) 170 (3.0%) 1096 (19.6%) 401 (7.2%) 2 (0.0%) 1 (0.0%)	IIIIIIIII I II III I	3986 (41.6%)
16	depression [numeric]	Mean (sd) : 0.4 (0.8) min < med < max: 0 < 0 < 9 IQR (CV) : 0 (2.1)	0 : 3955 (75.5%) 1 : 876 (16.7%) 2 : 205 (3.9%) 3 : 194 (3.7%) 7 : 2 (0.0%) 9 : 3 (0.1%)	IIIIIIIIIIIIII III	4343 (45.3%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
17	sleep [numeric]	Mean (sd) : 7 (3.2) min < med < max: 2 < 7 < 99 IQR (CV) : 2 (0.5)	12 distinct values	: : : : :	3300 (34.5%)
18	diabetes [factor]	1. yes 2. no	737 (7.7%) 8841 (92.3%)	I IIIIIIIIIIIIIIII	0 (0.0%)

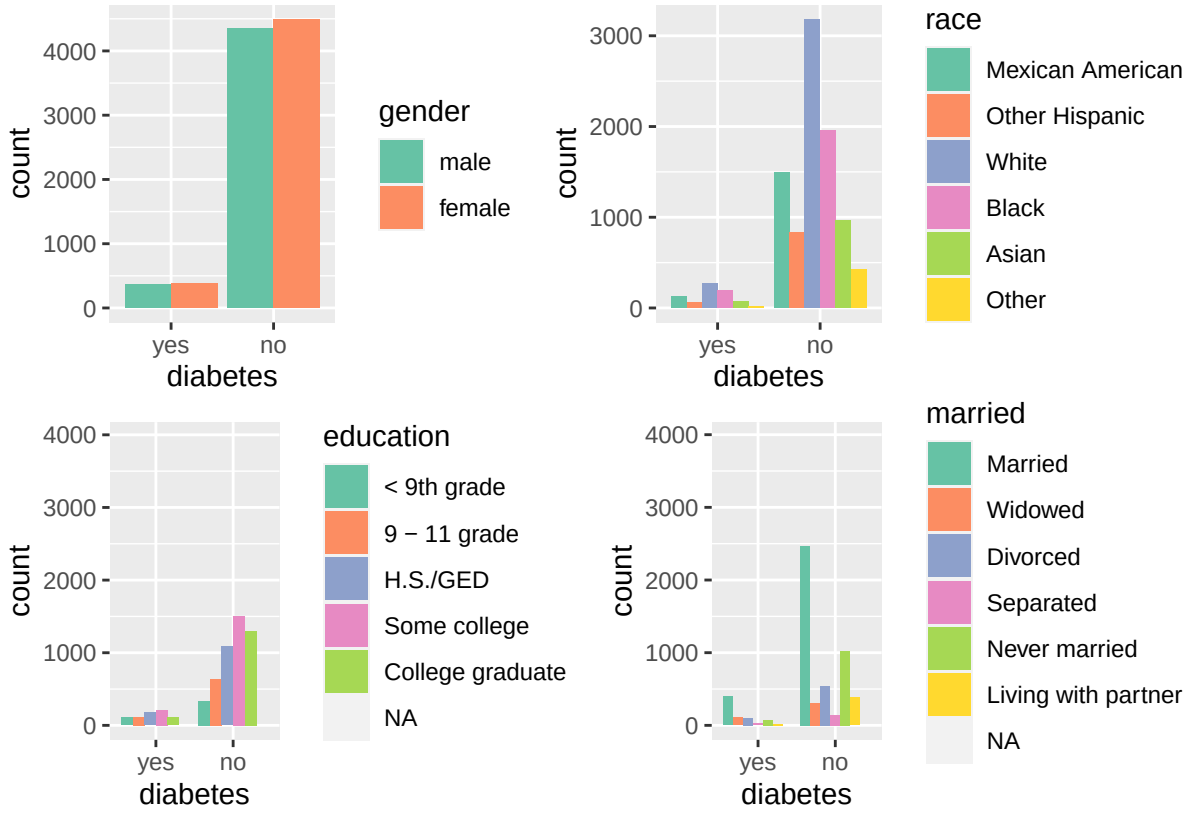
As noted above in the summary table, several of the variables for laboratory data had high missing values. The variables ldl, triglyceride, insulin and 2hr glucose-test had close to 70% of the data missing. In an effort to retain a large enough sample size, the four variables were not retained for further analysis in this project.

Density plots



Density plots of several numeric covariates showed differences in distributions of the two classes. Plots of systolic blood pressure, waist circumference measurement, age, and body mass index seemed significantly different between those with diabetes and those without. Most significant difference seemed to be among different age groups, with density curve of responses with no diabetes showing right-skewness, and those with diabetes skewed to the left along with a strong shift towards higher age.

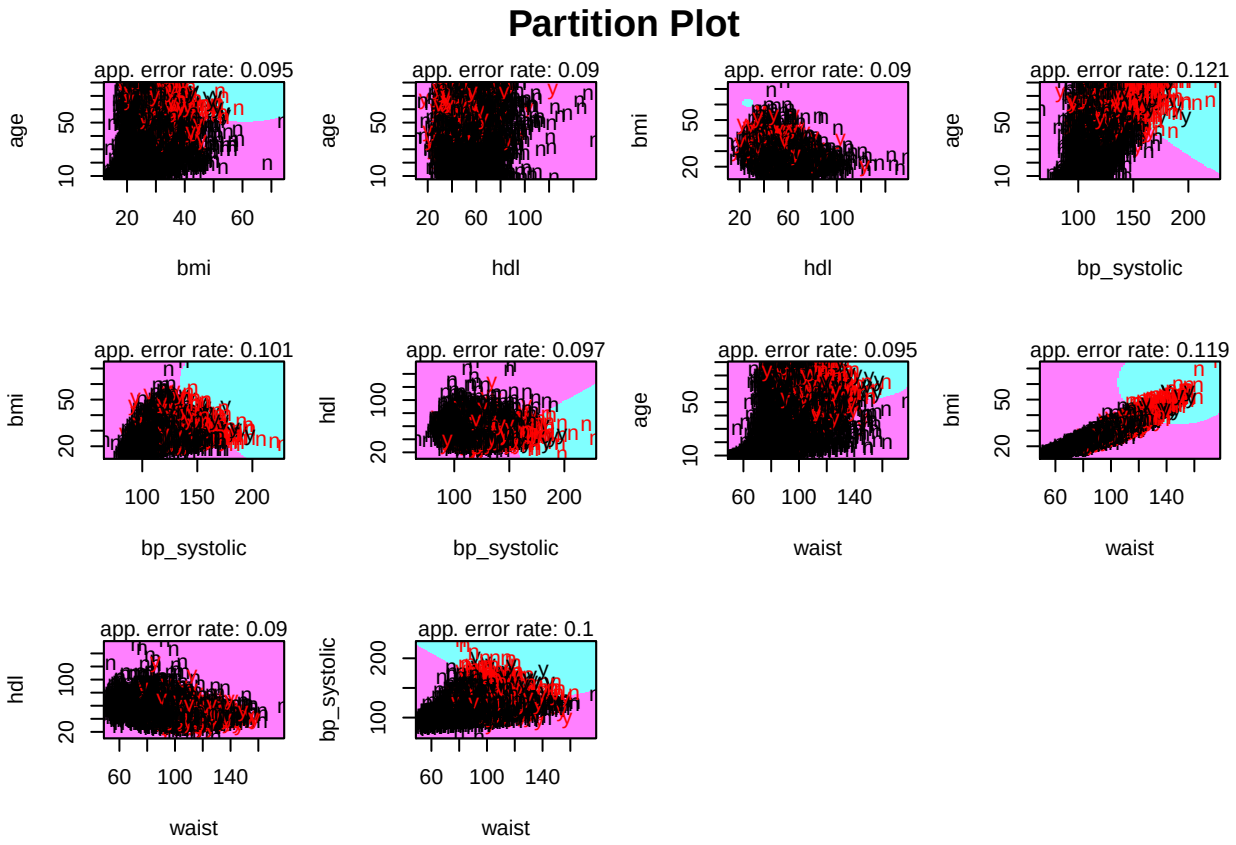
Bar plots



Bar plots were analyzed for categorical variables gender, race ,education level, and marital status. Presence of diabetes did not seem to be gender-dependent, and with slight differences based on education level. Proportion of positive diabetes cases did seem to vary among different races, and based on marital status. There seemed to be significantly higher proportion of non-diabetics among individuals who were never-married or divorced.

Finally, paired partition plots were also examined for several variables using training data, to analyze misclassification rate. Methods using linear discriminant analysis, and non-linear boundaries such as quadratic discriminant analysis and Naive Bayes method were used, giving similar results in terms of error rates. Shown below is partition plots using Naive Bayes method for paired visuals on age, bmi, hdl, systolic blood pressure, and waist.

Partition plots



Missing data

Variables with high proportion of missing data, most clsoe to 70% were removed. The dataset had close to 60% of missing values for the 4 variables which were considered significant for the given response. Assuming that the data was missing at random, and that single imputation might lead to bias and might not preserve relationships between variables; for those reasons imputation was not considered and the missing values were removed. The final sample consisted of 1,564 participants.

Models

Linear models

Non-linear models

Ensemble methods

Support vector machines

Model comparison

Final model

Model prediction performance

Limitations

Conclusion