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 for all 18 initial variables, including the outcome of response, 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

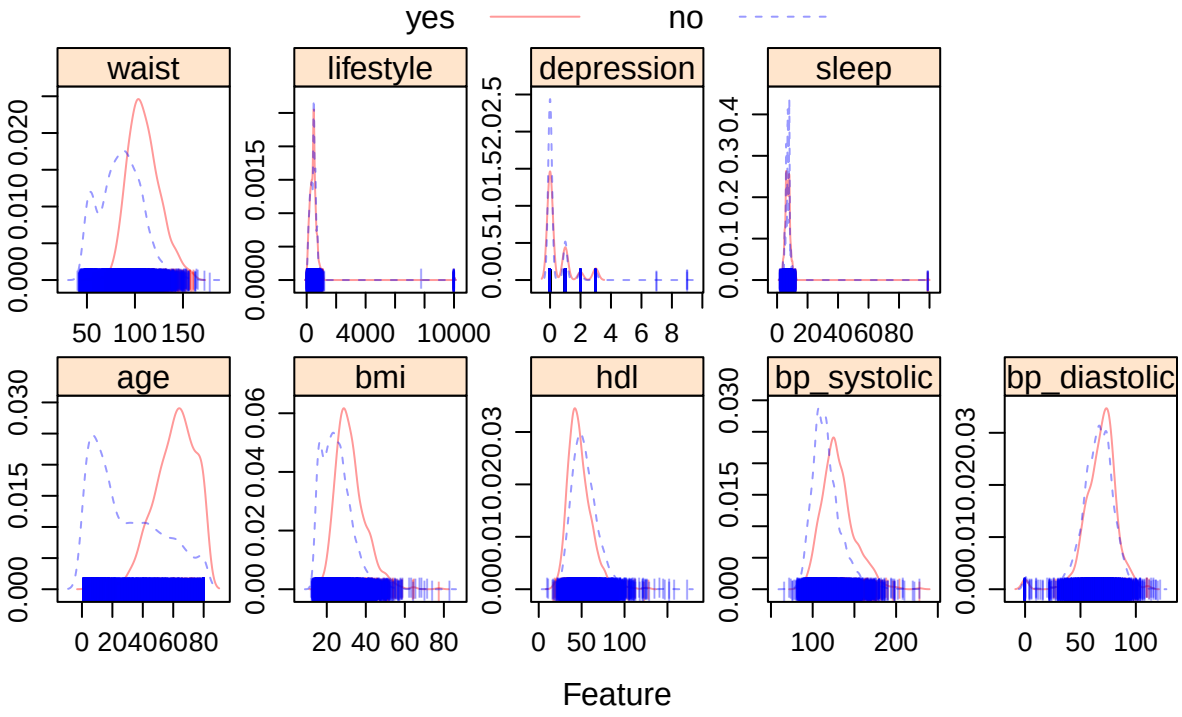
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. 1 2. 2	4706 (49.1%) 4872 (50.9%)	IIIIIIIII IIIIIIIII	0 (0.0%)
2	age [numeric]	Mean (sd) : 32.4 (23.9) min < med < max: 1 < 28 < 80 IQR (CV) : 41 (0.7)	80 distinct values	: : : : : . : : : : : : : : . . : : : : : : : : : :	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%)

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 a 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.

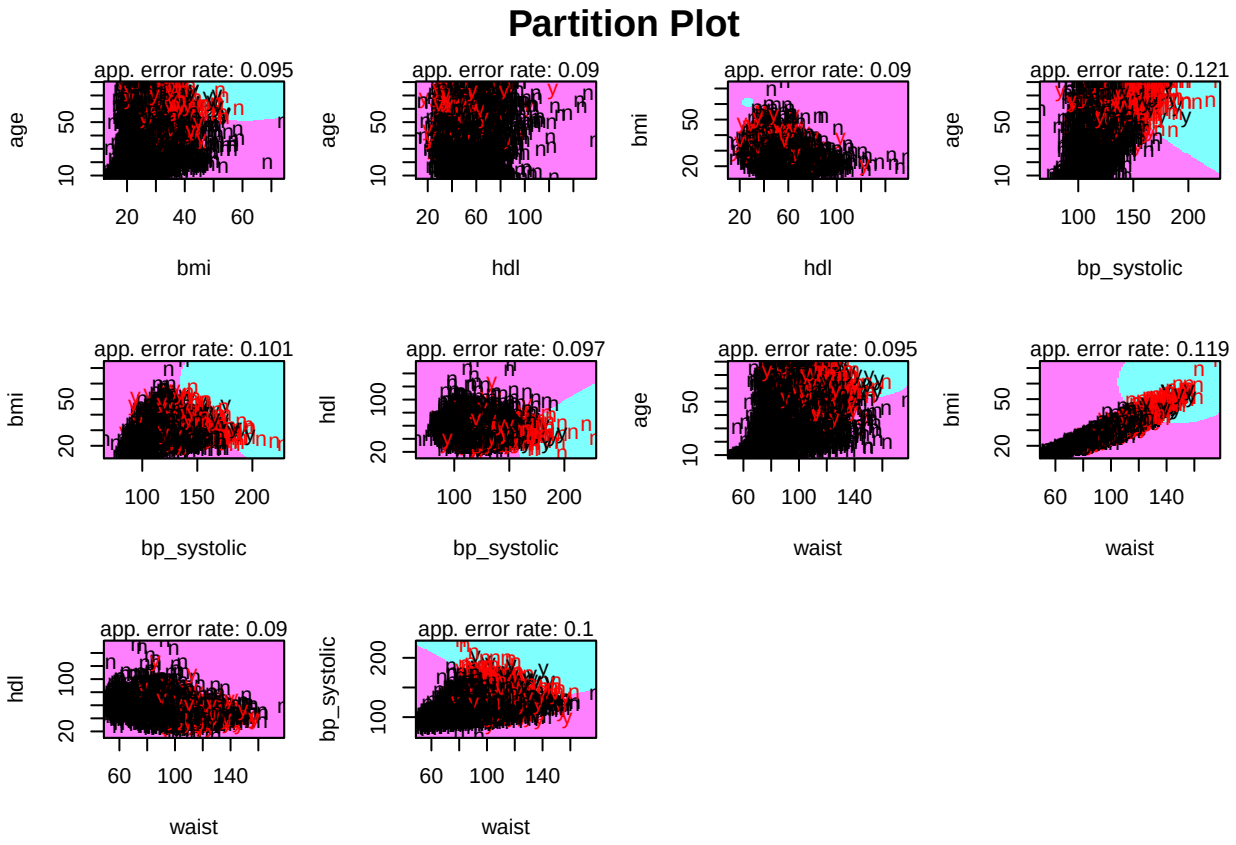
Bar plots

Bar plots, not shown, 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 a 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

The four Variables with high proportions of missing data, most of which were close to 70%, were removed. For the remaining variables with missing values, most were close to 25%, except marital status, education and depression level, which were approximately 40% missing. 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 4,246 participants, still a fairly large dataset.

Models

Methods

Several model-building techniques were used to predict the risk of diabetes using the 15 variables associated with it. Gender, age, race, bmi, hdl, systolic and diastolic blood pressure, waist circumference, lifestyle, education, marital status, depression and sleep were used as the predictor variables in the model building process. Given that the outcome was binary, generalized linear and non-linear methods such as logistic regression and discriminant analysis were tried, along with ensemble methods for trees, and the SVM (support vector machines) linear and non-linear boundary methods.

Due to ease of interpretability and the binary nature of the outcome, glm (logistic regression) and penalized logistic regression models were fitted.

———— GLM/Penalized model building ————————

Generalized additive model (GAM) was considered but not used due to the length of time for its execution. Non-linear MARS model, with similar performance to GAM, was retained. Various combinations for tuning grid were tried to get its optimal performance.

While no assumptions are needed for the predictors in linear logistic model, for discriminant analysis it is assumed that the predictors follow a normal distribution within each group of response, and that the variance-covariance matrix for response classes are the same for linear model (LDA), or could be different for non-linear QDA. Since four of the covariates in the dataset were categorical, these models were tried but excluded from analysis, also for the fact that these models work well for well-separated classes which was not apparent in the EDA. Under the assumption that features are independent within each class, Naives Bayes method (NB) was used due to its ability to handle mixed covariates, and due to its slightly better performance than LDA and QDA in terms of area under the ROC (receiver operating characteristic curve).

Models using non-parametric approach such as trees were also fitted using ensemble methods, which help improve prediction accuracy. Wisdom of crowds (bagging and random forest), and wisdom of weighted crowds of experts (Boosting) were the ensemble methods used.

————————- ensemble model details ——————————

Lastly, support vector classifiers were also fitted, with linear and non-linear decision boundaries.

————————- about SVM modeling ——————————

Neural networks were not considered due to their blackbox approach, which would result in non-transparency.

Model comparison

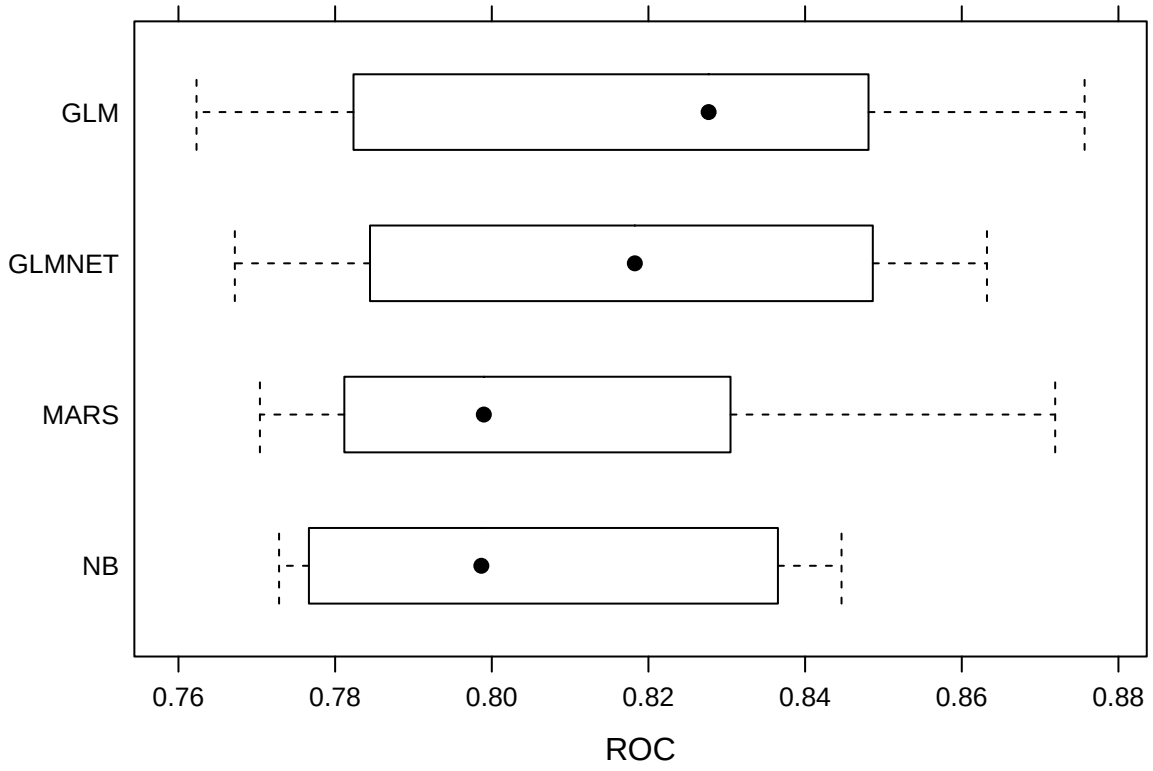
Call: summary.resamples(object = res)

Models: GLM, GLMNET, MARS, NB Number of resamples: 10

ROC Min. 1st Qu. Median Mean 3rd Qu. Max. NA's GLM 0.7623056 0.7861968 0.8276816 0.8196973 0.8465606 0.8756791 0 GLMNET 0.7671958 0.7860153 0.8182620 0.8155202 0.8457133 0.8632151 0 MARS 0.7704024 0.7818502 0.7989819 0.8084961 0.8296717 0.8719239 0 NB 0.7728447 0.7796216 0.7986685 0.8061999 0.8354096 0.8446480 0

Sens Min. 1st Qu. Median Mean 3rd Qu. Max. NA's GLM 0.07142857 0.11904762 0.1292913 0.14147287 0.1607143 0.2325581 0 GLMNET 0.04761905 0.07142857 0.1057586 0.09905869 0.1183555 0.1627907 0 MARS 0.04761905 0.11904762 0.1428571 0.13909192 0.1578073 0.2325581 0 NB 0.16666667 0.22245293 0.3095238 0.28117386 0.3313953 0.3720930 0

Spec Min. 1st Qu. Median Mean 3rd Qu. Max. NA's GLM 0.9730640 0.9772727 0.9815154 0.9831842 0.9898990 0.9966330 0 GLMNET 0.9697987 0.9806906 0.9898990 0.9858778 0.9898990 0.9932886 0 MARS 0.9697987 0.9747644 0.9831650 0.9828509 0.9882296 1.0000000 0 NB 0.9259259 0.9351852 0.9478114 0.9471855 0.9587542 0.9664430 0



-analysis

Final model

-visual -equation (?if exists) -explanation

Model prediction performance

-training/test output -variables of importance in prediction

Limitations

The models were limited in their accuracy due to the imbalance of the dataset.A larger dataset would have been needed to correct the skewness of class distribution, by including a greater range of data time-periods. Additionally, including greater set of covariates of potential influence, such as work conditions, exposure to environmental pollutants, etc. would also have helped to formulate a more precise model. Another limitation of models were that only the complete cases were used in the building process, under the assumption that data was missing at random. That might not have been the case for all missing data, for example the body weight data for participants who had limb amputations were set to missing. This factor was not a part of this data. Finally, extremely small sample sizes in several subcategories, which had to be excluded, were not handled well by the models.

Conclusion

- findings
- are findings what was expected?
- The data included both type and type 2 diabetes records without segregating the two types. It was unfortunate to realize that the age distinctions between the two forms of the disease are disappearing; what was known as adult-onset diabetes can begin during childhood.