A Study of Diabetes Classification based on NHANES Data with Different Machine Learning Methods

Xueting Tao, Jinhao Wang, Yili Wang, Dongyang Zhao

1. Background and Objectives

With a high prevalence of overweight individuals growing in the US, there is a trend of increasing prevalence in diabetes as well. To better control and prevent the development of diabetes, we aim to discover the most significant covariates and find the best machine learning algorithm for predicting diabetes, thus could give individual prevention ideas and help with earlier diagnosis of diabetes.

2. Method

2.1 Study Population

The Dataset we will be using is the National Health and Nutrition Examination Survey (NHANES), a program of studies designed to assess the health and nutritional status of adults and children in the United States,provided by the Centers for Disease Control and Prevention (CDC). The data range from 1999-2018, each with two years of a cross-sectional study. Different individuals were enrolled every two years. Data includes demographic, dietary, examination, laboratory, and questionnaire data.

2.2 Variables of selection

2.2.1 Outcome of Prediction

Diabetes was defined as “Doctor told you have diabetes”(named as DIQ010 in the NHANES dataset). 1 refers to Yes, 2 refers to No, 3 refers to Borderline, 7 refers to Refused, 9 refers to don’t know. Yes and Borderline were combined as “Yes”, 7, 9 and NA was excluded from the analysis.

2.2.2 Inclusion And Exclusion Criteria for variable selection

- Drop datasets without Sequence ID information
- Select datasets have information in more or equals to 10 year period.
- Use easy-to-obtain dataset: Demographic, Questionnaires and easy examination like Weight, Height, Oral, Vision and Audiometry.
- Variables in the Diabetes questionnaire was dropped, only keep DIQ010 as outcome.
- survy weights related variables were excluded.
- Variables with missingness more than 20% each year period was dropped.
- Remove all levels(factor) == 1 variable/constant variable for each year period
- After selection by year, we first select that variables that have more than 30% of overlap between years, then exclude the variables that have overall missingness more than 10,000.
- Complete cases was kept in the final model.

2.3 Analysis Approach

Following analysis approach was taken for the overall analysis

Insert flowchart here:

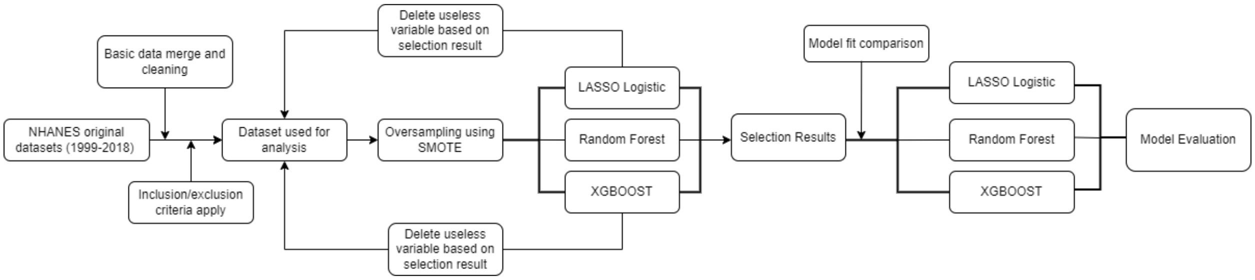

Delete useless variable based on selection result
Basic data merge and cleaning
Model fit comparison
LASSO Logistic
LASSO Logistic
NHANES original datasets (1999-2018)
Dataset used for analysis
Oversampling using SMOTE
Random Forest
Selection Results
Random Forest
Model Evaluation
XGBOOST
XGBOOST
Inclusion/exclusion criteria apply
Delete useless variable based on selection result

Figure 1: Analysis Approach

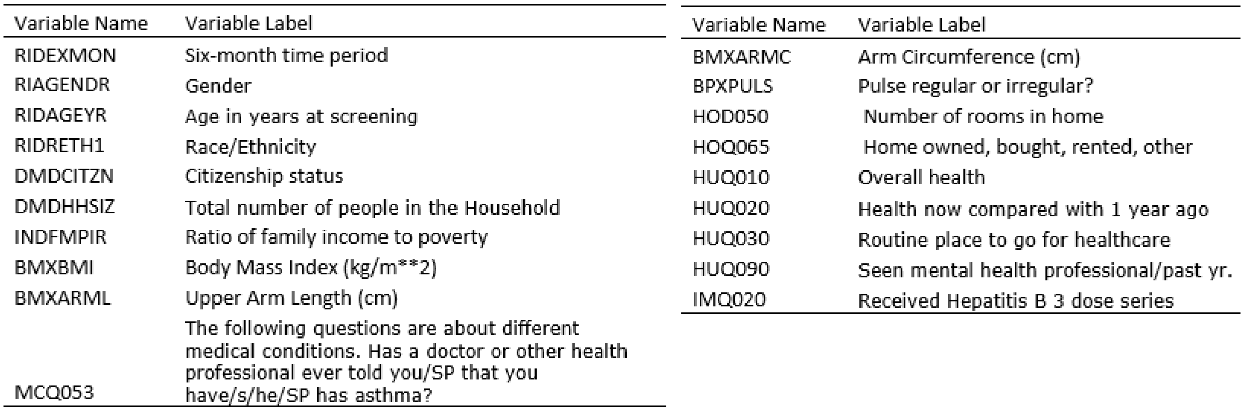
Key points: * Oversampling using SMOTE * Only train data has been SMOTE

3. Results

3.1 Feature Selection

Following is the sensitivity and specificity results for each year using different method in the variable selection process.

Based on the results, we found that XGBoost and Random Forests have low sensitivity, thus not reasonable to include variabels using these two methods. As a result, for the overall selection, we only include the variables results in LASSO. For LASSO, we chose variables that have at least 50% coverage of the whole year period (1999-2018), also with less than 10,000 missingness. Following is the plot showing the overall importance of variables selected and their definition.



Variable Name	Variable Label
RIDEXMON	Six-month time period
RIAGENDR	Gender
RIDAGEYR	Age in years at screening
RIDRETH1	Race/Ethnicity
DMDCITZN	Citizenship status
DMDHHSIZ	Total number of people in the Household
INDFMPIR	Ratio of family income to poverty
BMXBMI	Body Mass Index (kg/m**2)
BMXARML	Upper Arm Length (cm)
MCQ053	The following questions are about different medical conditions. Has a doctor or other health professional ever told you/SP that you have/s/he/SP has asthma?

Variable Name	Variable Label
BMXARMC	Arm Circumference (cm)
BPXPULS	Pulse regular or irregular?
HOD050	Number of rooms in home
HOQ065	Home owned, bought, rented, other
HUQ010	Overall health
HUQ020	Health now compared with 1 year ago
HUQ030	Routine place to go for healthcare
HUQ090	Seen mental health professional/past yr.
IMQ020	Received Hepatitis B 3 dose series

Figure 2: Variables selected

Based on the results, we found that _____ is the most importance ..(more interpretation)

3.2 Model comparision

Following table is the comparision between different model fits using the variables selected above. Overall, LASSO and XGBoost have similar results. Since random forest took longer than expected(), thus we stopped the process.

4. Conclusion

- SVM, DT not ideal in our situation
- LASSO, XGBOOST works great
- Feature Selection using LASSO/XGBOOST/RF:
  - Lasso: features that's important in 9 or 10 years
  - Xgboost/RF: Sensitivity Result not ideal, the selection result was not considered in the final model
- In order to make the result meaningful, meaningless variables need to be excluded based on selection. The process has been rerun multiple times. − Data is messy with over 3000+ variables without cleaning, thus, difficult to select by hand. − Thus use this process.

4.1 Computational chanllege and solutions:

- High missingess and low overlap on the variables for different year of survey, thus, there is no complete cases if we work on the overall dataset.
  - we chose to separate the analysis for different year and then combine to chose the highly overlapped variables
  - Drawback: Tried PCA/LDA, dimension reduction cannot be applied when we have data separated by year
- Figured the imbalance data problem ### 1:11
  - Talk about difference between class weights and SMOTE
  - Oversampling using SMOTE
  - Since we don't want to have overlap in test and train data, we smote the data after separating the train/test dataset. Only train data has been smote
  - After SMOTE, all the sensitivity from different method increase, however, the results from Random Forest and XGBoost still didn't ideal(most have sensitivity less than 0.5). Thus, we choose to only use the LASSO result as our selection results.
- Separate Data by year:no complete case in the overall dataset, some of the variables in the NHANES have different names throughout the years.
  - so we only chose those variables with missingness less than 10%.
  - Drawback: give us even larger datasets
- Feature selection is slow:
  - pack feature selection as separate functions and run on cluster
- Overall around 3,000 variables needed to be selected, difficult to clean by hand.
  - Thus, we first select, then based on the selection results, clean out meaningless variables and possible colinearities(like age in month & age in year)
  - After cleaning, we run the selection again until the selected variables are all meaningful.
  - Drawback: Saves labor but cost more time on running.

4.2 Future works

- Try the same process on less-missingness data
- Try imputation on the missing variables