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 Data cleaning

Data was download from NHANES website. All the datasets were downloaded from website and then arranged by year and data type(Demographic, Dietary, Examination, Laboratory and Questionnaire). A indicator table was made to check the coverage of each dataset (see "Reference/Codebook for datatables V2.xlsx"). Based on the indicator table, we decided to use the following inclusion/exclusion criteria to choose the datasets to include.

- Drop datasets without Sequence ID information
- Select the datasets with information appearing in more or equals to 10 year period.
- Use easy-to-obtain variables: Demographic, Questionnaires and easy examination like Weight, Height, Oral, Vision and Audiometry.

After basic exclusion, we merge the selected datasets and further clean the data exclude the variables based on the following criteria:

- Variables in the Diabetes questionnaire was dropped, only keep DIQ010 as outcome.
- survey 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

Our outcome variable is defined as:

- 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 were excluded from the analysis. The variable was then releveled to 0 and 1: 1 being Yes and 0 being No.

2.3 Analysis Approach

In the whole dataset with over 3000 variables, each having different missingness, there was no complete case in our data. Moreover, there were less than 200 variables with less than 20% missing values. To prevent excluding potentially useful variables and keep as many observations as possible, instead of modelling on the whole dataset, we predicted diabetes in each group. After conducting feature selection by every two years, we combined the variables which contributed the most in each group.

Another problem we met was that our data was imbalanced. The ratio of positive to negative class in response was 1:11. Synthetic Minority Oversampling Technique (SMOTE) was introduced to deal with the imbalanced data issue. SMOTE is a commonly used oversampling method to re-balance the response variable for better performance on predictive models. To avoid over-fitting, we partitioned data into training and testing data and applied SMOTE to re-balance the response variable.

The feature selection methods in our project include LASSO, Xgboost, and Random Forest. By applying those methods to our data of 10 groups, we not only obtained sets of variables selected by the methods, we were also able to compare the sensitivities, specificity, accrancies, and time elapsed among the three models. Based on the performances of the three methods, we determined which variables should be further selected. Furthermore, we also checked the meaning of those variables selected to prevent problem of multicollinearity. After determine the final variable set, we wanted to know if those variables would perform well in our overall data using the three different fitting methods. Thus we fit the three models again using our finally selected variables.

Complete cases was used in the final model. R version 4.1.12 was used for the analysis.

Following analysis approach was taken for the overall analysis

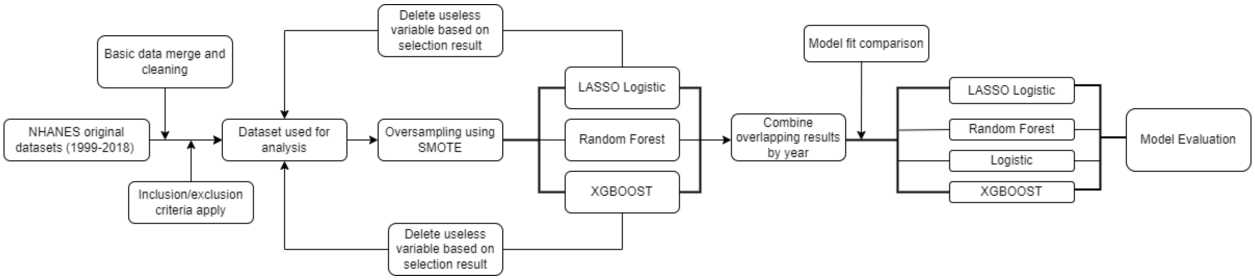

Delete useless variable based on selection result
Basic data merge and cleaning
Model fit comparison
LASSO Logistic
NHANES original datasets (1999-2018)
Dataset used for analysis
Oversampling using SMOTE
Random Forest
Combine overlapping results by year
LASSO Logistic
Random Forest
Logistic
XGBOOST
Model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

Figure 1: Analysis Approach

3. Results

3.1 Feature Selection

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

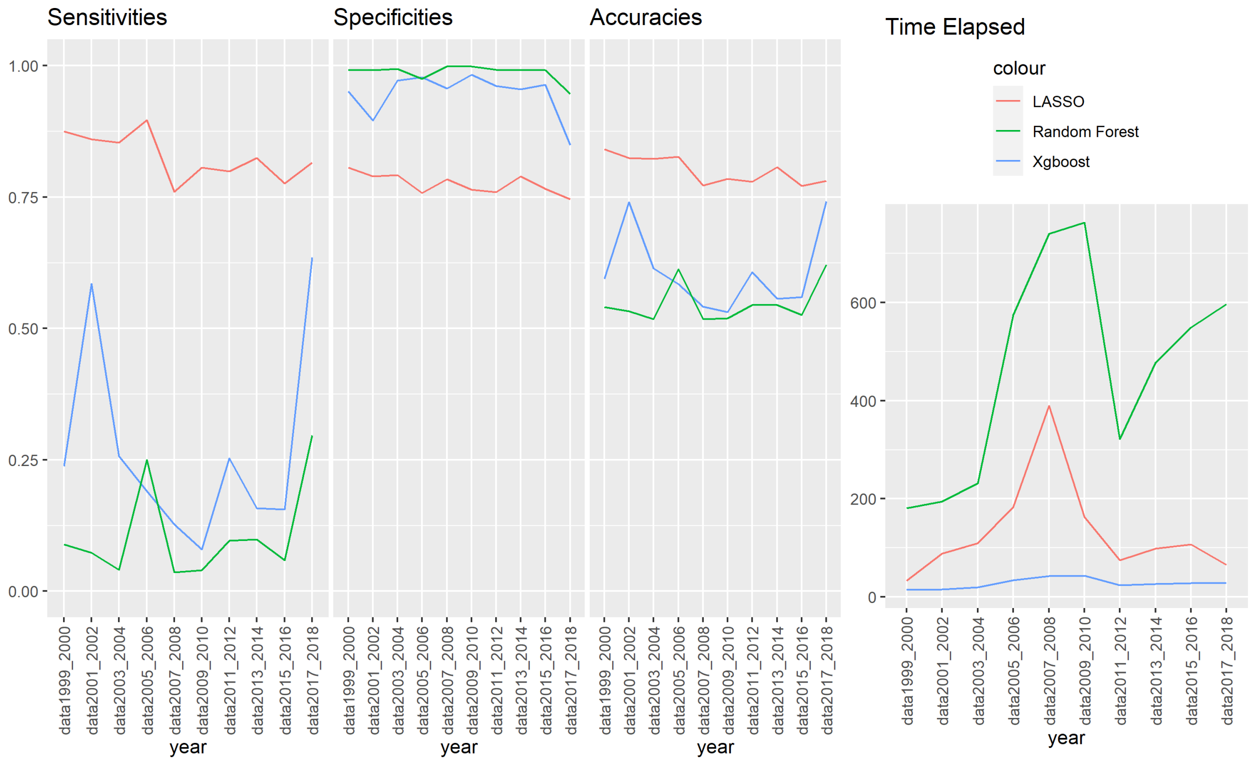


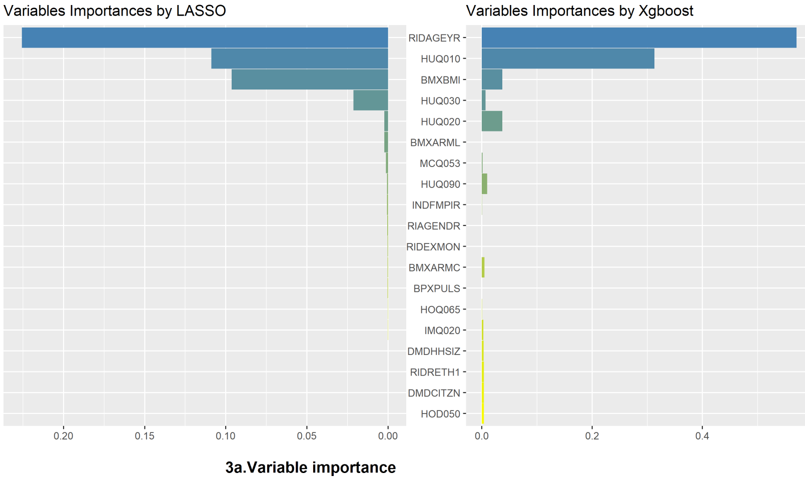
Figure 2: Sensitivity, Specificity and Accurancy by year

Based on the results, we found that XGBoost and Random Forests have unsatisfiable sensitivity, which gave no reason to include variables selected by the above two methods. As a result, we only kept the variables selected by LASSO.

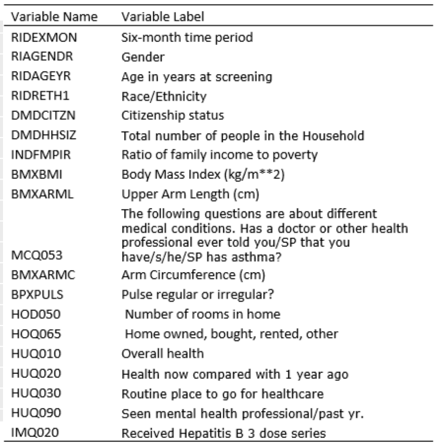
We further reduced the number of variables according to at least a 50 percent selection rate in the years they appeared and less than 10,000 missing values.

Here we present the descriptions of each selected variable.

Following is the plot showing the overall importance of variables selected.



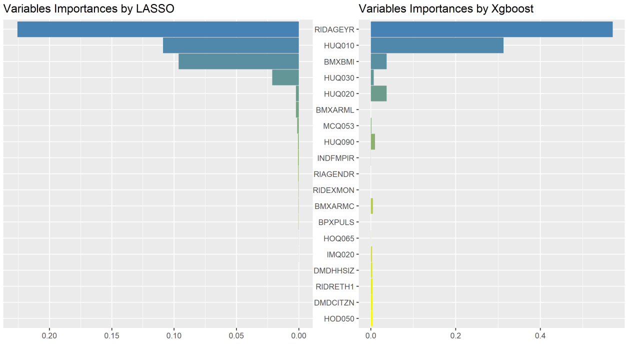
3a.Variable importance



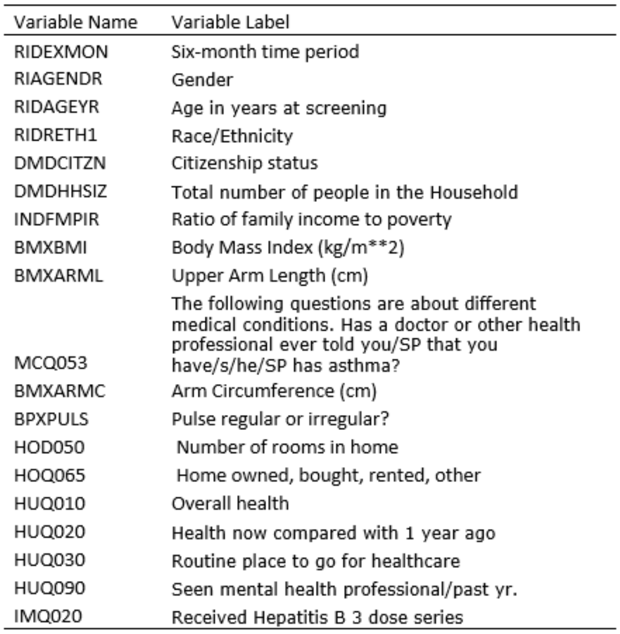
Variable Name	Variable Label
RIDEXMON	Six-month time period
RIAGENDR	Gender
RIDAGEYR	Age in years at screening
RIDRETH1	Race/Ethnicity
DMDCITZN	Citizenship status
DMDHHSIZ	Total number of people in the Household
INDFMPIR	Ratio of family income to poverty
BMXBMI	Body Mass Index (kg/m**2)
BMXARML	Upper 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?
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

3b. Variable defination

Figure 3: Variables of selection



3a.Variable importance



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?
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

3b. Variable defination

Figure 4: Variables of selection

Based on the results, we found that Age(RIDAGEYR),Overall Health(HUQ010),BMI(BMXBMI),Routine place to go for healthcare(HUQ030) is the most important variables used to predict Diabetes.

3.2 Model comparision

The following confusion matrices compare our final models fitted using the selected variables shown above. Overall, Logistic regression, LASSO Logistic Regression, Random Forest and XGBoost all produced similar results. Both logistic regression types had a slight advantage in sensitivity, where Random Forest and XGBoost showed higher specificity.

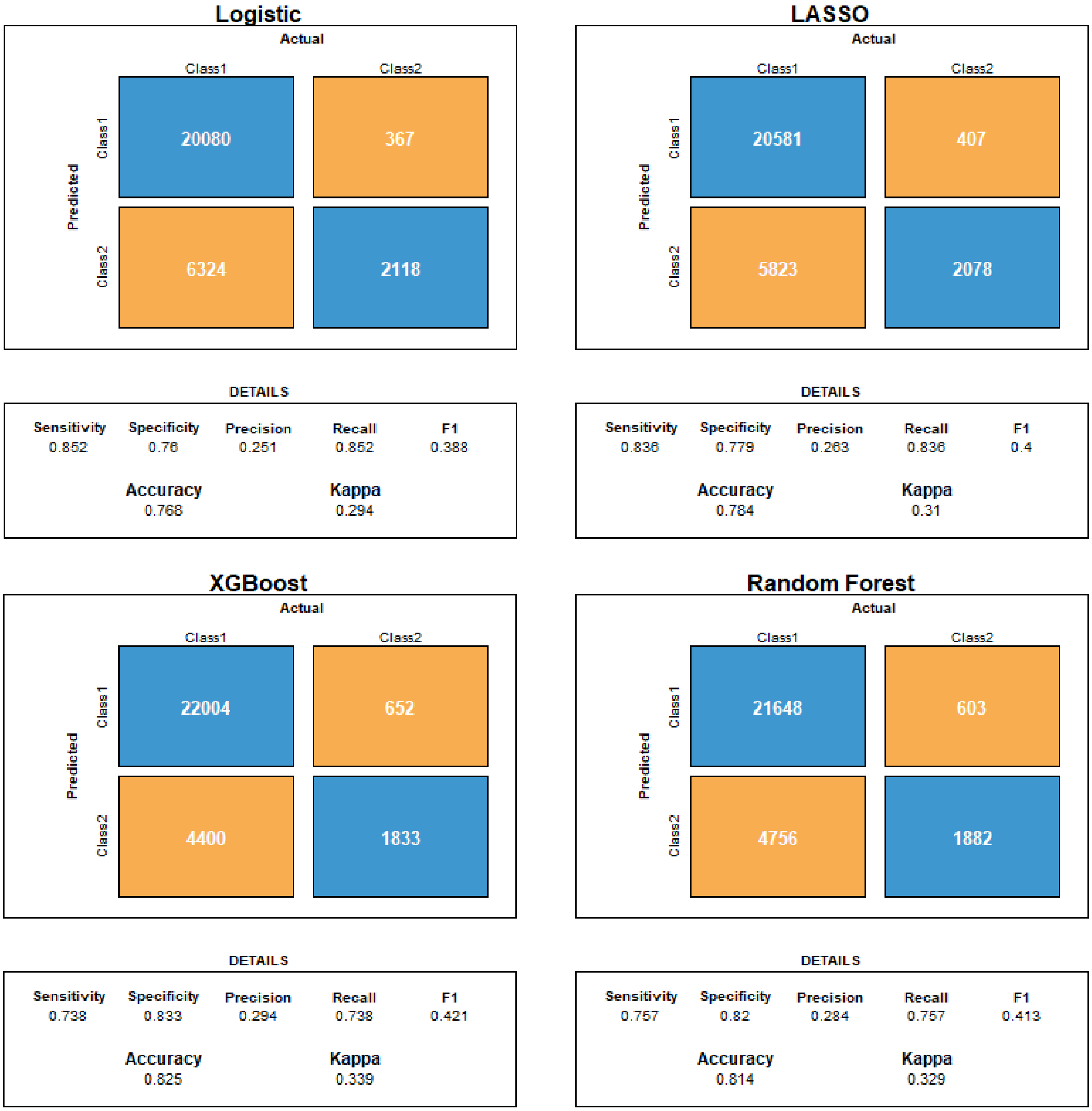


Figure 5: Confusion Matrix

4. Conclusion

4.1 Feature Selection

Based on our result, it is obviously that LASSO works great in feature selection while the sensitivities of Xgboost and Random Forest are not consistent and not ideal. Thus, selection result was only considered from the output of LASSO. Features were selected according to at least 50% rate of being selected by the model to the number of years the variables appeared in the data. To make sense of the variables selected from the model and prevent multi-collinearity, meaningless variables or variables with high correlation with the others were excluded from the selection. Since overall,there is around 3,000 variables needs to be selected, it is difficult to check if the definition is meaningful by hand one by one. Thus, we decided to first run the whole selection process, and then check the variables selected in the process. If the variable is not meaningful, we remove that from the data cleaning process and run the whole selection again. This process has saved us lots of labor.

4.2 Computational challenge and solutions:

Since there is 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. Thus, we decided to separate the analysis for different year and then combine to chose the highly overlapped variables. However, by using this approach, we found that PCA/LDA cannot be applied.

As we have mentioned above, the outcome of interests in our dataset is highly unbalanced, which results in low sensitivity in the feature selection process. After literature research, we decided to apply SMOTE for oversampling. 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.This have solve our imbalance data problem but it gives us an even larger datasets

During the process, we found that our feature selection is slow. Thus, we decided to pack the feature selection as separate functions and run the feature selection on the cluster. This saves us a lot of time and allow us to work separately together.

4.2 Future works

One of the biggest problem in our data is imbalances. Although we tried oversampling in the minority group using SMOTE, the result in feature selection process is still not ideal for Xgboost and Random forest. It would be of great interest for us if there is any better way to handle the unbalanced data or if we could have less-missing data and apply Xgboost and Random forest again to see if the features selected by these two models could also work well. On the other hand, as Random Forest and Xgboost are not good at handling imbalanced data, we would also be interested in improving the algorithm of these two methods to deal with imbalanced data better in the future.

Another limitation in our data is large missingness. Due to lack of knowledge in missing data, we only used complete cases in our project, which may cause to loosing some information. We also hope to find a way to deal with the large missingness in our data and apply our models again.

5 Group contribution

Xueting Tao: Downloading the data, data cleaning and merging, finalize the final report. Jinhao Wang: Yili Wang: Dongyang Zhao: