Obesity Prediction Using Random Forest

Shuai Huang

2022-12-06

Introduction

Obesity is a chronic metabolic disease related to genetic, environmental and social factors. It is a risk factor for many diseases such as hypertension, diabetes, cardiovascular diseases and respiratory diseases. With the modernization of lifestyles, poor diets and reductions in physical activity, the prevalence of obesity has increased at an alarming rate, in both developed and developing countries. According to data from the CDC, the obesity prevalence in the US has increased from 30.5% in 1999 to 42.4% in 2018. During the same period, the prevalence of severe obesity increased from 4.7% to 9.2%. According to a 2014 report by Health Economics, obesity is highly correlated with diabetes, hypertension, coronary heart disease and stroke. The annual medical cost of obesity in the United States was estimated at $147 billion in 2008, with the average cost for an obese person being $1,429 higher than a non-obese person. The causes and prevention for obesity is of great significance. However, the causes and mechanisms of obesity are not well understood. Obesity is related to multiple causes. Environmental factors, lifestyle preferences, and cultural environment may play key roles in the growing global obesity. There are also no effective sustainable obesity interventions. Hence, identifying important risk factors may lead to developing effective obesity prevention strategies and programs.

Method:

Data:

Part of the data was collected from people from the countries of Mexico, Peru and Colombia using a survey in a web platform. The age of the participants is between 14 and 61. In order to make the data, the authors first searched for literatures to find out the most possible factors that may induce obesity. The covariates include diverse eating habits and physical condition. The questionnaire was conducted anonymously, so the researchers could ensure that participants' privacy was not violated. The total sample size of the original data was 485, and the outcome was not very balanced, with about 300 normal participants, and the total number of other weight level just a small percent. For example, both of the number of overweight I and overweight II participants are only about 50. To make the data more balanced, the original data was

processed to obtain a sample size of 2111. After the balancing class problem was identified, synthetic data was generated, up to 77% of the data, using the toolWeka and the filter SMOTE. The final data set has a total of 17 features and 2111 records. Then all the participants were labeled as obesity and not obesity based on height and weight using the equation for calculating the BMI and the criteria for classifying obesity (BMI larger than or equal to 25.0 will be classified as obesity and below 25.0 will be classified as non-obesity). In the analysis, the response variable is obesity, which is a binary variable, and there are 14 covariates used for evaluation, including eating habits, physical condition and other features. The eating habits features are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The physical condition features are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other covariates obtained were: Gender, Age, Smoke, Family overweight history. The covariate age is a continuous variable, while other covariates are all categorical variables.

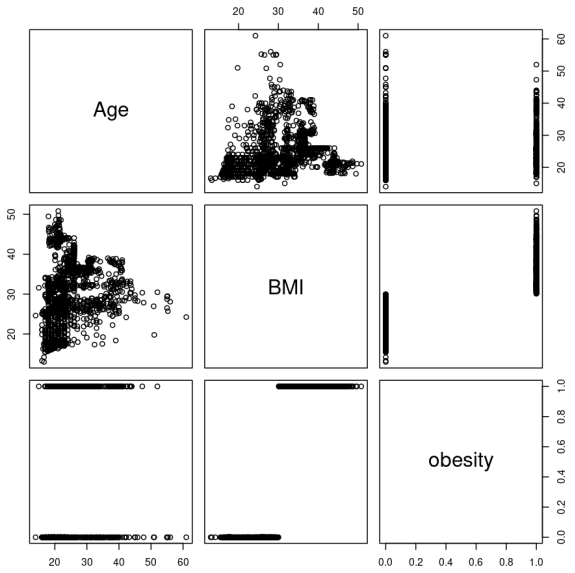


Figure 1: Figure1: Correlation plot

Statistical analysis

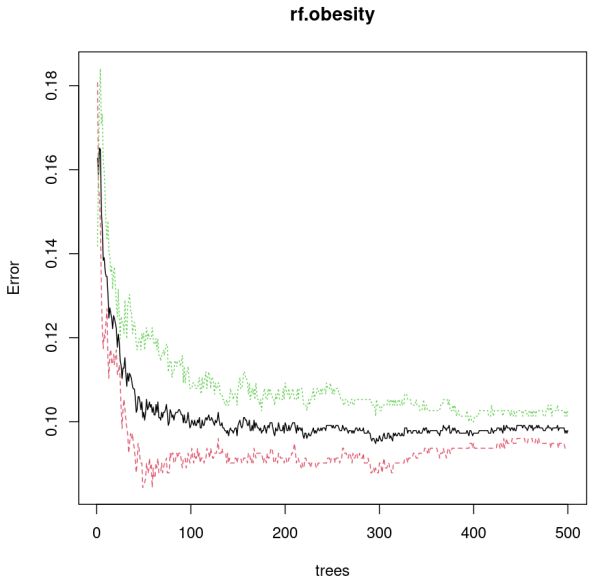
Since the sample size is relatively large (2111 records), I split the data set into three parts: 75% train data set, 25% test data set. The train data set was used to do cross validation to tune the best parameters in the random forest. The test data set was used to do prediction on the obesity level based on the final models trained on the train data set. The binary variable obesity was used as the response variable. The rest 14 variables (excluding height and weight) are used as the covariates. The ROC curves were plotted, and the AUC was calculated to measure the performance. For the random forest model, the best number of features

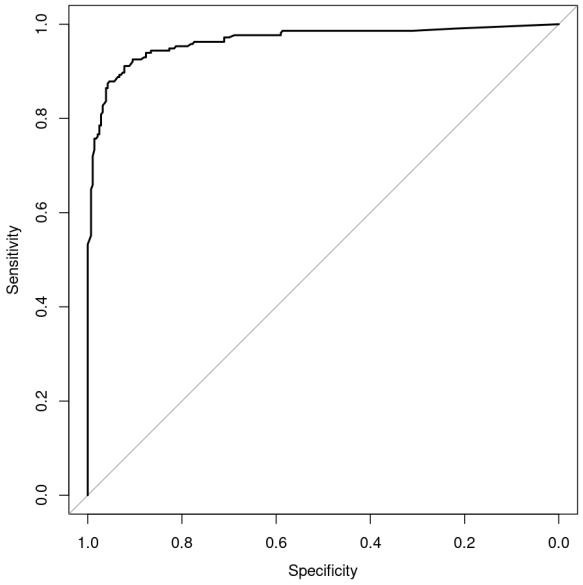
at each split point (called mtry) was tuned using the 3-time repeated 5-fold cross validation. The default number of trees to grow when tuning the mtry parameter was 1000. After choosing the best mtry, to balance the prediction performance against costs, the relationship between the error and the number of trees was evaluated to choose an appropriate number of trees. I then did a feature selection using GINI impurity (feature importance will be calculated based on mean decrease GINI impurity).

Results

Model and the performance

For random forest, the best number of features at each split point selected is 14. Then use this number to fit the random forest model, and use the classification error as the response, and the number of trees as the covariate, we can find that when the number is greater than 300, the large number does not do good to the error, then we will choose the number of trees as 300. Then we use the final model with the parameters stated above to predict the obesity status in the test data set. The ROC curve is demonstrated in Figure 3 as an example.

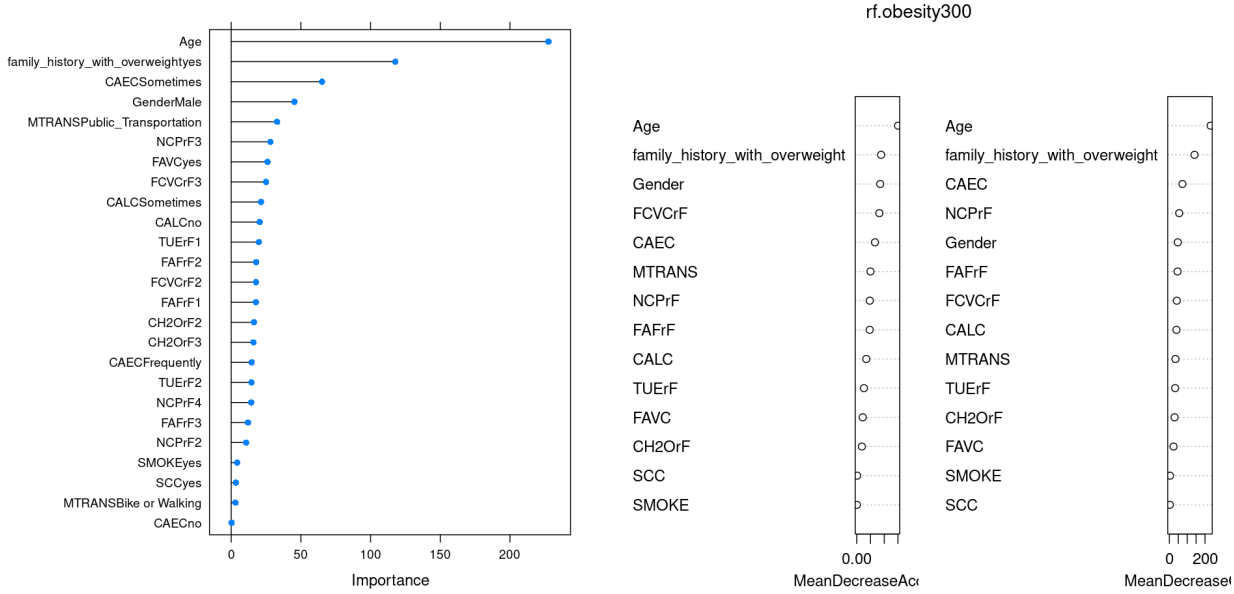




Feature importance

For the random forest, the feature importance was calculated using the mean decreased GINI impurity, and the most important four features are Age, Family history with overweight, Consumption of food between meals, Number of main meals, and Gender. Smoke and Calories consumption monitoring (SCC) are the least important, the mean decreased GINI impurity of which are about 0. Use the 3-repeated 5-fold cross validation, I found that when the number of features used is larger than 12, the accuracy will decrease,

then we may choose 12 variables as the most important variables to predict the obesity (Smoke and SCC omitted). The Smoke variable is one of the least important features, so that it will be removed from the prediction models. For obesity prevention, we can tell people to focus on the most important things instead of expending too much energy on other less important tasks, since people usually don't want to focus on too many things.



Limitations

The response variable obesity is binary. We can have more categories, like ranking weight into underweight, normal, different levels of overweight and obesity, which helps to make more targeted prediction. The features are all categorical variables, although it may save some resources, but it may also reduce the accuracy for the prediction. If the resources permit, we can further collect more detailed information. Besides, some features are not balanced enough, for example, only 96 people (4.55%) monitored their calories consumption, then we may not make accurate comparison between those who monitored their calories consumption and those who did not.