# Obesity Prediction Based on Logistic Regression, Random Forest and Support Vector Machine

Shuai Huang

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:
Qaqish Bahjat

# 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 [1]. 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 [2]. 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%. [3]

According to a 2014 report by Health Economics, obesity is highly correlated with diabetes, hypertension, coronary heart disease and stroke. [4] 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. [3]

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.

# Methods:

## Data:

Part of the data was collected from people from the countries of Mexico, Peru and Colombia using a survey in a web platform [5]. 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 [6]. 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.

Table 1: Descriptive statistics of the data set. The sample sizes are 2111 individuals.

| | No. (%) |
|---|---|
| Age (years) | |
| Mean (SD) | 24 (6) |
| Gender | |
| Male | 1068 (50.59%) |
| Female | 1043 (49.41%) |
| Obesity | |
| Obesity | 974 (46.14%) |
| Non-obesity | 1137 (53.86%) |
| Family overweight history | |
| Yes | 1726 (81.76%) |
| No | 385 (18.24%) |
| Frequent consumption of high caloric food | |
| Yes | 1866 (88.39%) |
| No | 245 (11.61%) |
| Frequency of consumption of vegetables | |
| Never | 102 (4.83%) |
| Sometimes | 1013 (47.99%) |
| Always | 996 (47.18%) |
| Number of main meals per day | |
| 1 | 316 (14.97%) |
| 2 | 176 (8.34%) |
| 3 | 1470 (69.64%) |
| 4 | 149 (7.06%) |
| Consumption of food between meals | |
| Always | 53 (2.51%) |
| Frequently | 242 (11.46%) |
| Sometimes | 1765 (83.61%) |
| No | 51 (2.42%) |
| Consumption of water daily | |
| Less than a liter | 485 (22.97%) |
| Between 1 and 2 L | 1110 (52.58%) |
| More than 2 L | 516 (24.44%) |
| Consumption of alcohol | |
| Frequently | 71 (3.36%) |
| Sometimes | 1401 (66.37%) |
| No | 639 (30.27%) |
| Calories consumption monitoring | |
| Yes | 96 (4.55%) |
| No | 2015 (95.45%) |
| Physical activity frequency per week | |
| I do not have | 720 (34.11%) |

| | |
|---|---|
| 1 or 2 days | 776 (36.76%) |
| 2 or 4 days | 496 (23.50%) |
| 4 or 5 days | 119 (5.64%) |
| Time using technology devices | |
| 0-2 hours | 952 (45.10%) |
| 3-5 hours | 915 (43.34%) |
| More than 5 hours | 244 (11.56%) |
| Transportation used | |
| Automobile or Motorbike | 468 (22.17%) |
| Bike or Walking | 63 (2.98%) |
| Public Transportation | 1580 (74.85%) |
| Smoke | |
| Yes | 44 (2.08%) |
| No | 2067 (97.93%) |

**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 and support vector machine, and was used to do backward model selection in the logistic regression. The test data set was used to do prediction on the obesity level based on the final models trained on the train data set.

Three methods are used to fit the prediction models: Random Forest, Logistic Regression and Linear Support vector machine [7]. 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, sensitivity and specificity and the corresponding 95% confidence intervals are calculated to measure and compare the performance of the three different models.

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).

For the Logistic Regression, the full model was first fitted, and then, using the backward selection method and AIC as the criteria, a reduced model was fitted. Then compared between full model and the reduced model to see if there was significant change.

For the support vector machine, the 3-time repeated 5-fold cross validation was used to tune the best cost for the support vector in the wrong side of the margin. Then used the tuned parameter to fit the final model and calculate the three metrics.

Then, I further studied the robustness when some kind of record error occurred. In real life, we may mismatch the record with participants, for example, the BMI score for person A may be recorded as the BMI score for person B. To simulate that kind of record error, I shuffled the response variable of the first 10% of the data. Then use the shuffled data set to fit the three models and do predictions on the test data set to see the performance.

# Results

## Model comparisons

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 1 as an example.

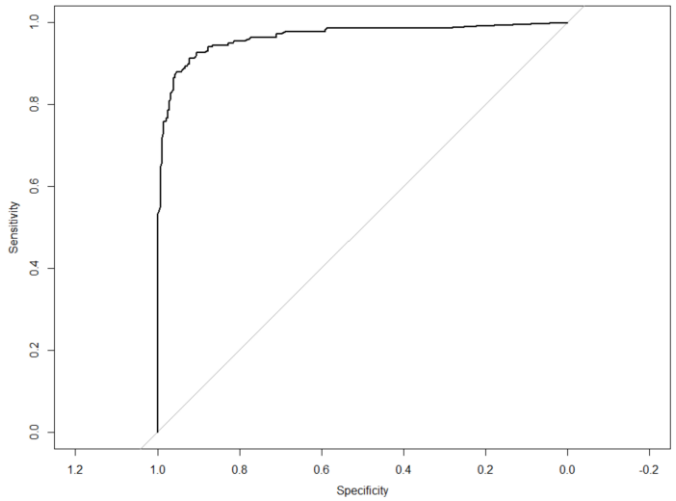


Figure 1: ROC curve (random forest, mtry=14, number of trees=300)

For logistic regression, the AUC for the full model was 0.87 and the corresponding 95% confidence interval is (0.83, 0.90).

For the reduced model after using the backward selection, 12 features were left. The variables Smoke and Time using technology devices are removed. The AUC is 0.87, with the 95% confidence interval (0.83, 0.90). The AUC using the full model and the reduced model are not of a big difference, since the 95% confidence intervals tangle with each other, which means there is no significant reduce in the performance of the reduced model. But if we can delete one or two variables, we may save our money and time to conduct the analysis. Then I chose the reduced model for the comparison between three models.

For the support vector machine, the best cost for the support vector in the wrong side of the margin selected is 0.5, then this cost was used to fit the linear support vector machine.

The AUC, sensitivity and specificity are compared in the table 2.

The AUC using random forest is significantly better than other models. The AUC using the

support vector machine is the lowest, but the 95% confidence interval of AUC for the support vector machine and logistic regression tangle with each other, so there is no big difference between the two models.

Table 2: AUC, sensitivity and specificity and the 95% CI for three final fitted models

| | Random Forest | Logistic Regression | Support Vector Machine |
|---|---|---|---|
| AUC (95% CI) | 0.96 | 0.87 | 0.84 |
| | (0.95, 0.98) | (0.83, 0.90) | (0.81, 0.88) |
| Sensitivity (95% CI) | 0.93 (0.90, 0.96) | 0.82 (0.77, 0.86) | 0.92 (0.87, 0.95) |
| Specificity (95% CI) | 0.89 (0.84, 0.93) | 0.77 (0.70, 0.82) | 0.69 (0.64, 0.75) |

## Record error simulation

I further study the robustness of the three methods when some record errors occur.

As mentioned above, to simulate the record error, I shuffled the response variable of the first 10% of the data. After shuffling, only 34 records are different. For the reduced logistic regression model, only the Smoke variable was removed. The AUCs before and after shuffling are listed in the table 3. We can see that for all the three models, the AUC before and after shuffling the data are not significantly different, which means all the three models are robust with little record errors.

Table 3: AUC and the corresponding 95% CI before and after shuffling for three models

| | Random Forest | Logistic Regression | Support Vector Machine |
|---|---|---|---|
| AUC (95% CI) | 0.96 | 0.87 | 0.84 |
| Before shuffling | (0.95, 0.98) | (0.83, 0.90) | (0.81, 0.88) |

| | | | |
|---|---|---|---|
| AUC (95% CI) | 0.96 | 0.86 | 0.84 |
| After shuffling | (0.95, 0.98) | (0.83, 0.90) | (0.81, 0.88) |

## Feature importance

Feature importance was evaluated before shuffling the data, and we can see the similarities and differences among the three methods.

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).

For the reduced logistic regression, Smoke and Time using technology devices are removed.

The least important measured using the support vector machine is Smoke and Gender, while the most important features are family history with overweight, Age, Consumption of food between meals, Number of main meals, Frequent consumption of high caloric food.

For the three models, the random forest is significantly better than the other two models for the AUC. So that for better prediction on the level of obesity, we may use random forest. However, for better interpretation, we may use logistic regression. Five coefficient estimates for the most significant covariates are listed in the table 4. The p-values are less than 0.00001.

For example, in this cross-sectional study, controlling for other factors, individuals who had

habit of frequent consumption of high caloric food had 5.7 times the (prevalence) odds to have obesity compared to those who did not have that habit. The 95% CI is (3.2, 10.3), which did not include 1 (the null), and therefore was statistically significant.

Controlling for other factors, with 1 year older, an individual would have 1.09 times the odds to have obesity compared to a year ago. The 95% CI is (1.06, 1.12), which means aging is also one of the factors that cause obesity.

Table 4: Estimated coefficients of the 5 covariates out of 12 covariates fitted in the reduced logistic regression that have p-value less than 0.00001

| | Coefficient (95% CI) |
|---|---|
| Age | 0.087 (0.059, 0.114) |
| Family overweight history | 3.560 (2.708, 4.411) |
| Frequent consumption of high caloric food: Yes | 1.746 (1.159, 2.334) |
| Public transportation | 1.221 (0.803, 1.638) |
| Number of main meals: 1 | -2.368 (-3.356, -1.381) |

For all the three methods, the Smoke variable is one of the least important features, so that it will be removed from the prediction models. The most important features selected by the three models are not very different, therefore, if the resources, like money and time, are not adequate, we can choose the most important features selected by any model to conduct further research. 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.

## Conclusion and Discussion

In conclusion, the random forest method showed the best performance in the prediction of

obesity. However, logistic regression and the support vector machine had good performance as well.

Logistic regression has the advantage of easily interpreted regression coefficients. All the three methods were robust against 10% random errors in the data.

## 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.

Also, it would be more meaningful if the outcome has more categories when we are simulating the record error situation. If we have more categories of outcome, more incorrect records would be generated after shuffling the data.

The decision boundary for the support vector machine used is the linear boundary, we can further try non-linear kernel to see the performance.

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.

## References

[1] World Health Organization. (2020). *Obesity: Preventing and Managing the Global Epidemic*. Geneva: World Health Organization.

[2] Zhang, J., Shi, X. M., & Liang, X. F. (2013). Economic costs of both overweight and obesity among Chinese urban and rural residents, in 2010. *Zhonghua Liuxingbingxue Zazhi*, *34*(6), 598-600.

[3] CDC. (2015). Adult obesity facts. Available from: https://www.cdc.gov/obesity/data/adult.html

[4] Winter, J., & Wuppermann, A. (2014). Do they know what is at risk? Health risk perception among the obese. *Health Economics*, *23*(5), 564-585.

[5] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief*, *25*, 104344.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

[7] Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.