Determining Risk Factors That Lead to Obesity

Chikaze Mori

City, University of London
Student ID: 200038013

***Abstract*—Obesity is a global pandemic, and it is necessary to determine risk factors that lead to obesity. In this paper, an empirical research was conducted which used the dataset whose attributes are physical conditions and eating habits. We applied Random Forest to the dataset and inspected the results by looking at the feature importance, the confusion matrices, and the performance measures. It was discovered that family history of obesity is the physical condition that has the highest risk for obesity and that eating habits affect obesity more than physical conditions. Yet it was not confirmed that frequency of consumption of vegetables was the eating habit that has the highest risk for obesity, although it was found that the habit had the highest feature importance in the model and has the biggest effect on obesity.**

INTRODUCTION

Obesity is associated with a host of cardiovascular risk factors [1], and the prevalence of obesity has been rising steadily over the last several decades and is currently at unprecedented levels, and this increase has occurred across every age, sex, race, and smoking status in many countries [2]. Although the most common cause of obesity is excess energy consumption, obesity can be considered as a disease with multiple factors including biological hazard factors such as hereditary background [3]. Thus, there is an urgent need for the development of highly effective interventions that aim to reverse the risk factors for obesity, including both government policies as well as health education and promotion programs [2].

The conducted analysis in this paper provides insights into risk factors that lead to obesity. Our motivation of the analysis is to help governments and health organisations to make a discipline for individuals that informs them what risks they have and what habits they should change, e.g., stopping smoking.

1. ANALYTICAL QUESTIONS AND DATA

1.1. Analytical Questions

The scope of the analysis in this paper is the questions listed below:

I. What physical conditions are risk factors that leads to obesity?
II. What eating habits are risk factors that leads to obesity?
III. Which more likely lead to obesity, eating habits or physical conditions?

The first and second questions are to indicate what specific physical condition or eating habit is a risk for obesity. The last question is to show what leads to obesity in a large scale.

These questions are to warn people who have such physical conditions or eating habits of their risk for obesity and possibly prevent obesity.

1.2. Data

The dataset used in the analysis is from UCI Machine Learning Repository, and was collected by Palechor and Manotas [4] specifically for estimation of obesity levels in individuals. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, while 23% of the data was retrieved from online survey in countries of Mexico, Peru and Colombia, based on their eating habits and physical conditions. The dataset contains 17 variables and 2111 records including obesity levels as a class variable, and already pre-processed including missing data imputation, and data normalisation. The attributes related with eating habits 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), Consumption of alcohol (CALC), and Consumption of cigarettes (SMOKE), while the attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), and Family member who suffered or suffers from overweight (family_history) [4]. Considering all the attributes could be a risk factor that leads to obesity, the dataset is suitable for our analysis.

2. ANALYSIS

2.1. Data preparation

Each obesity level was labelled by mass body index calculated for each individual using the equation shown below (Equation 1) [4].

$$mass\ body\ index = \frac{\text{weight}}{\text{height} * \text{height}}$$

Equation 1

The dataset includes the attributes, height and weight, which were directly used for calculating the obesity levels. Since our target label is obesity levels, we eliminated these attributes to avoid problems that would occur due to their correlation. The attributes Gender and Age were also eliminated from the dataset as they are not physical conditions nor eating habits.

The dataset was split 70% and 30% as train and test sets respectively, including 14 attributes that represent eating habits and physical conditions and 1 label which has 7 classes that represent obesity levels.

2.2. Data derivation

The dataset has already been normalised and its missing data was imputed. However, it still contains categorical and ordinal attributes which are not ideal for applying Random Forest in Python. Thus, we converted all the categorical attributes into numerical, e.g., Automobile in MTRANS attribute into 0, and Motorbike into 1.

2.3. Construction of models

To answer the first and second analytical questions, we applied Random Forest to the dataset, which is suitable for finding non-linear prediction rules that are also interpretable [5]. First, we conducted a hyperparameter tuning on a Random Forest model using 10-fold Cross Validation to prevent overfitting and validate the model. Even though the two important parameters for Random Forest are the number of trees and the number of the attributes [6], the only hyperparameter tuned in the process was the number of trees in the forest as we needed to use all the attributes to see their feature importance and interpret the result. The model with the lowest classification error was to be chosen as the best model, and we generated the confusion matrix of the best model applied to the test set to see its performance (Figure 1).

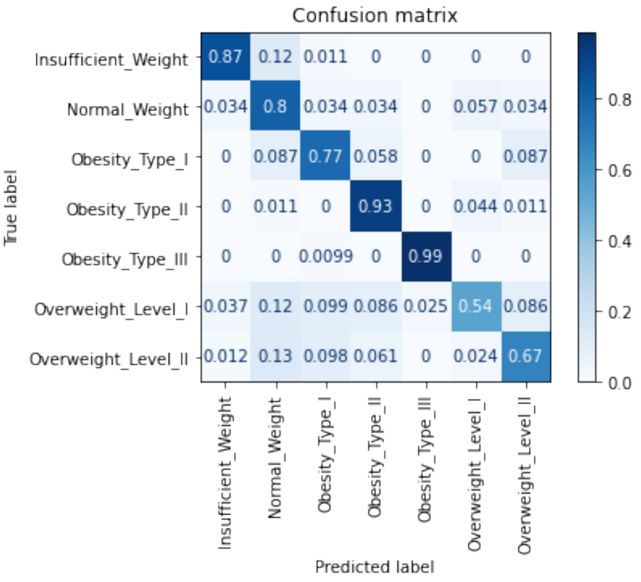

Figure 1

The feature importance calculated was Permutation Importance (PIMP), which is a heuristic for normalizing feature importance measures that can correct the feature importance bias of Random Forest models [5]. In our analysis, PIMP was used to decide the importance of attributes and interpret the result for analysing which attribute leads obesity (Figure 2).

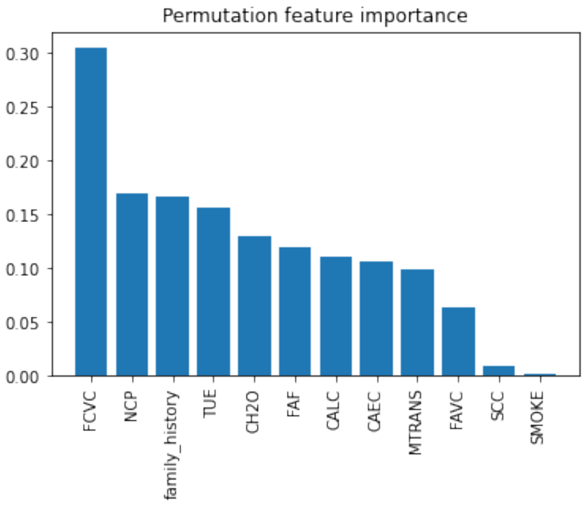

Figure 2

To answer the third analytical question, we generated the 2 new datasets whose attributes are only physical conditions or eating habits. To each of the datasets, we applied Random Forest with the hyperparameter that was used for the best model trained for the entire dataset. By applying Random Forest to the datasets, we are to compare the performance of the models and its effect to obesity. The confusion matrices of the models applied to the test set were generated and shown below as Figure 3 and 4.

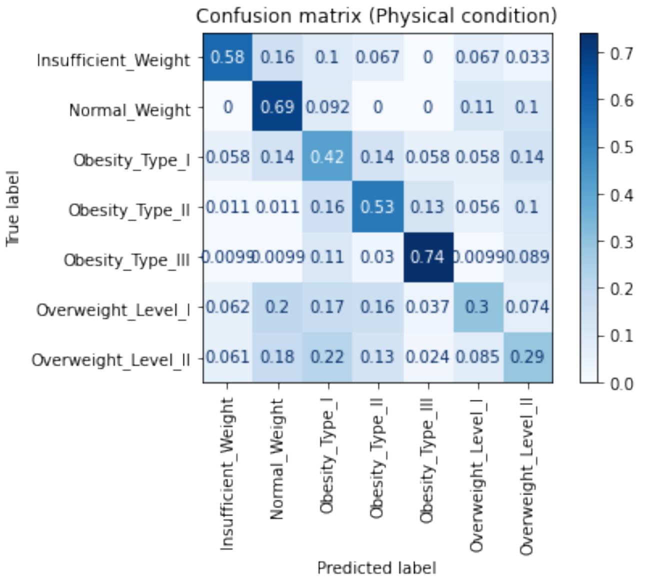

Figure 3

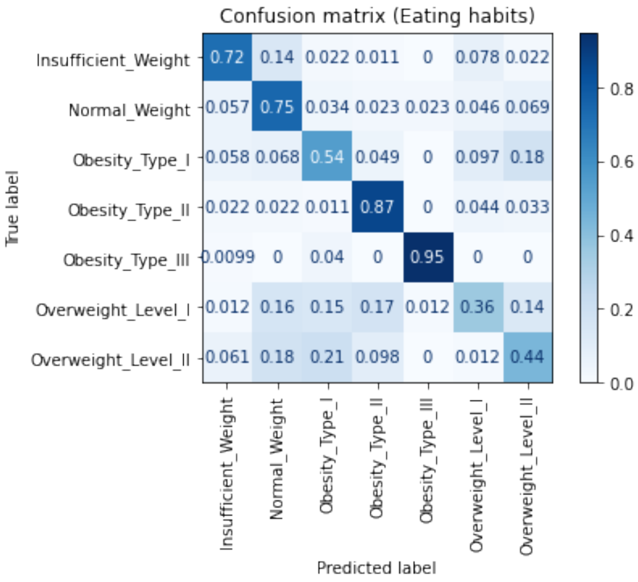

Figure 4

2.4. Validation of results

As confusion matrix is not a single value that explains performance of models, classification error and Matthews Correlation Coefficient (MCC) were also calculated and shown below as Table 1.

	all	Physical condition	Eating habits
Classification error	0.195584	0.485804	0.329653
MCC	0.773215	0.433466	0.616547

Figure 5

As accuracy is by far the simplest and widespread measure in the scientific literature, we calculated the accuracy of each model and convert it to classification error to show the performance of each model. However, accuracy poorly copes with unbalanced classes and it cannot distinguish among different misclassification distributions [7]. Thus, we also calculated MCC which is a performance measure of multi-class classification models that is a good compromise among discriminant, consistency and coherent behaviours with varying number of classes, unbalanced datasets, and randomisation. [7].

3. Findings, reflections, and further work

3.1. Findings and reflections

From Figure 1, it is clear that the best model predicts obesity levels very well, considering 77%, 93% and 99% of accuracy in obesity type 1, 2, and 3 respectively. However, the accuracies in overweight were relatively low, while the accuracies in insufficient and normal weight were high. Based on the result, it is thought that some attributes lead to obesity and others prevent it.

The confusion matrix, Figure 2, shows that Frequency of consumption of vegetables (FCVC) is the eating habit with the highest feature importance and Family member who suffered or suffers from overweight (family_history) is the physical condition with the highest feature importance. The importance of FCVC is outstanding and almost twice as big as the second highest. This might mean FCVC is the risk factor that leads to obesity the most, however, it is possible that FCVC is an eating habit that reduces the risk of obesity the most. Considering the nature of the attribute, the latter is more likely. On the other hand, family_history has to be considered as the physical condition with the highest risk for obesity, considering the nature of the attribute, genetics.

Comparing Figure 3 and 4, it seems that the model based on physical conditions predicted worse than the model based on eating habits. It is noticeable that the model based on eating habits still predicted obesity levels well, which suggests eating habits are more associated with obesity levels than physical conditions.

Figure 5 shows the supremacy of the model based on eating habits over the one based on physical conditions. From this result and the fact FCVC had outstanding feature importance, it is concluded that eating habits have a bigger effect on obesity than physical conditions. However, it was not confirmed that eating habits more likely led obesity, as it was not clear that the attributes were risk factors or risk reducers.

3.2. Further work

By the analysis, the important factors were determined, and it was confirmed that eating habits have more effect on obesity than physical conditions. Our conclusion of the analysis is to help individuals to realise eating habits are important for their health. However, it is necessary to confirm if the factors are risk factors or reducers, so the governments and health organisations are able to publish a detailed guideline based on scientific facts. For example, researching more on the most important factor, FCVC, would be helpful to know if eating vegetables reduces the risk of obesity more than not eating vegetables increases obesity. To do so, one can create a new dataset that contains dummy variables of each attribute, apply a classification algorithm such as logistic regression, and see how the variables contribute to the model.

References

[1] A. De Schutter, C. J. Lavie, and R. V. Milani, ‘The Impact of Obesity on Risk Factors and Prevalence and Prognosis of Coronary Heart Disease—The Obesity Paradox’, *Progress in Cardiovascular Diseases*, vol. 56, no. 4, pp. 401–408, Jan. 2014, doi: 10.1016/j.pcad.2013.08.003.

[2] S. M. Wright and L. J. Aronne, ‘Causes of obesity’, *Abdom Radiol*, vol. 37, no. 5, pp. 730–732, Oct. 2012, doi: 10.1007/s00261-012-9862-x.

[3] A. De la Hoz Manotas, E. De la Hoz Correa, F. Mendoza, R. Morales, and B. Sanchez, ‘Obesity Level Estimation Software based on Decision Trees’, *Journal of Computer Science*, vol. 15, p. 10, Jan. 2019, doi: 10.3844/jcssp.2019.67.77.

[4] F. M. Palechor and A. de la H. Manotas, ‘Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico’, *Data in Brief*, vol. 25, p. 104344, Aug. 2019, doi: 10.1016/j.dib.2019.104344.

[5] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, ‘Permutation importance: a corrected feature importance measure’, *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010, doi: 10.1093/bioinformatics/btq134.

[6] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, ‘An Empirical Study of Learning from Imbalanced Data Using Random Forest’, in *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, Oct. 2007, vol. 2, pp. 310–317, doi: 10.1109/ICTAI.2007.46.

[7] G. Jurman, S. Riccadonna, and C. Furlanello, ‘A Comparison of MCC and CEN Error Measures in Multi-Class Prediction’, *PLOS ONE*, vol. 7, no. 8, p. e41882, Aug. 2012, doi: 10.1371/journal.pone.0041882.

WORD COUNTS

Section	Word counts
Abstract	134
Introduction	162
Analytical questions and data	298
Analysis	556
Findings, reflections and further work	447