2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM) | 978-1-6654-3260-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISAIAM53259.2021.00041

Estimation of Obesity Levels Based on Decision Trees

Tong Cui [1]
Vehicle Engineering, Tongji University,
Shanghai, China,
tedvincent23@gmail.com

Yuxuan Chen [2]
Computer Science dual Mathematics, Rensselaer Polytechnic Institute,
Troy, United States,
chenyuxuan18@outlook.com

Jiahao Wang [3]
Computer Science, Michigan State University,
East Lansing, United States,
wangj153@msu.edu

Haoran Deng[4]
Banner Christian School,
Chesterfield, United States,
dhr20040322@gmail.com

Yuchen Huang [5]
Nanjing Foreign Language School,
Nanjing, China,
13851834436@163.com

***Abstract*— In recent decades, there has been increasing concern about obesity in adolescents and adults. Obesity can cause many physical health problems and affect people's quality of life. So people are starting to look at the factors that lead to obesity and predict the emergence of obesity. This research presents an estimation of obesity levels based on eating habits, physical condition, and other factors, using a dataset found on UCI Machine Learning Repository. This dataset contains 17 attributes and 2111 records. The labels of this dataset are classified as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. In this research, three major methods are chosen for prediction: Decision Trees, Logistic Regression, and K Nearest Neighbor. Finally, the result obtained by Decision Trees has the best accuracy.**

Keywords — Obesity; Data Mining; Data Visualization; Decision Trees; XGBoost

I. INTRODUCTION

The World Health Organization (2020) states, "Overweight and obesity are defined as abnormal or excessive fat accumulation that may impair health." The WHO (2020) describes that worldwide obesity has nearly tripled since 1975, more than 1.9 billion adults, 18 years and older, were overweight in 2016, of these over 650 million were obese. Furthermore, over 340 million children and adolescents aged 5-19 were overweight or obese in 2016[1].

Obesity can cause many medical problems. According to (Kopelman P, 2000), the author states that, obesity causes or exacerbates many health problems, both independently, and in association with other diseases, it is associated with the development of type 2 diabetes mellitus, coronary heart disease (CHD), and an increased incidence of certain forms of cancer. Moreover, the adverse effects of excess weight tend to be delayed, sometimes for ten years or longer[2].

From above, obesity has become a major problem worldwide. Currently, the most common way to calculate obesity is to calculate BMI. The WHO (2020) describes, body mass index (BMI) is a simple index of weight-for-height that is commonly used to classify overweight and obesity in adults[1]. Several obesity calculators can be found on the internet, and they are all using BMI to calculate obesity. According to (Kopelman P, 2000), genetic susceptibility, the increasing availability of high-energy foods and the decreasing need for physical activity in modern society have a significant impact on obesity[2]. So, the dataset, which contains the BMI and family history, physical activities, and other daily habits, is used to predict obesity.

The main purpose of this study is to use 14 attributes other than BMI to predict obesity and compare the result with previous works. By solving the problem, we could know why people are suffering from obesity and predict whether they would suffer from obesity, and give them some suggestions.

II. BACKGROUND AND RELATED WORKS

In determining the causes of obesity, we read articles concerning all possible factors and selected the most important and decisive ones: gender, age, family history, drinking alcohol, smoking, high-calorie food intake, water consumption, and food intake between main meals. Most studies suggest that age is not closely related to obesity since people of all ages are equally overweight. On the other hand, family history plays an important role in certain diseases (Corica et al., 2018)[3]. Garawi revealed in his article that the prevalence of obesity is greater in women in most countries, but the magnitude of the difference varies in different countries due to gender inequality (Garawi et al., 2014)[4]. Additionally, drinking alcohol is proposed to have a dual impact on one's weight level depending on alcohol consumption. It is accepted that drinking a moderate amount of alcohol prevents people from weight gain, while too much consumption eventually leads to obesity (Traversy & Chaput, 2015)[5].

978-1-6654-3260-3/21/$31.00 ©2021 IEEE
DOI 10.1109/ISAIAM53259.2021.00041

Similarly, moderate water consumption also helps to control weight gain since drinking half a liter of water increases the activity of the sympathetic nervous system (Vij and Joshi, 2014)[6]. High-calorie food intake is directly connected with weight gain, while meal pattern is also decisive in one's risk of being obese. Aránzazu Aparicio investigated the effect of eating patterns on weight level and found that eating snacks between meals actually decreases the likelihood of obesity (Aparicio et al., 2017)[7].

All these studies analyze the effect of a single factor on obesity level using logistic regression models and multivariable fractional polynomials to predict the outcome. In order to find out the impact of mixing these elements, we built a machine learning model to make more comprehensive and convincing predictions.

III. Dataset

The dataset is based on the life routine and body condition of different volunteers. Data was collected using a web platform with a survey (Palechor & de la Hoz Manotas, 2019)[8].

The dataset consists of 2111 columns and 17 rows. There are different types of features in this dataset, including nominal and ordinal types. Moreover, fortunately, this dataset has been preprocessed by scientists, such as deleting missing and atypical data, data normalization, and solving the problem of imbalance, using the tool Weka and the filter SMOTE.

We divided the features into five categories: basic information, eating habits, physical conditions, family history, and harmful behaviors. The meaning of features in the dataset is shown in **Fig. 1**. Using the dataset, "Height" and "Weight" are not considered in our prediction model. On top of it, no new column or row is added to our dataset.

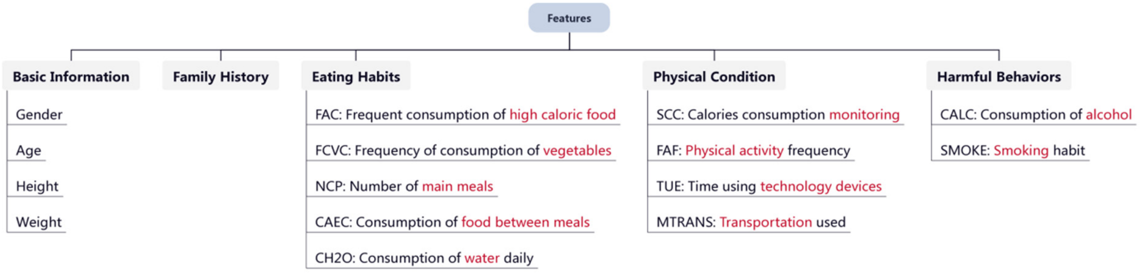
Features
Basic Information
Gender
Age
Height
Weight
Family History
Eating Habits
FAC: Frequent consumption of high caloric food
FCVC: Frequency of consumption of vegetables
NCP: Number of main meals
CAEC: Consumption of food between meals
CH2O: Consumption of water daily
Physical Condition
SCC: Calories consumption monitoring
FAF: Physical activity frequency
TUE: Time using technology devices
MTRANS: Transportation used
Harmful Behaviors
CALC: Consumption of alcohol
SMOKE: Smoking habit

Fig. 1: Meaning of Features in the Dataset

IV. Methods

To predict the obesity level of people based on eating habits and physical condition, we first discarded the Weight and Height of people because the BMI index has already been a critical factor of labeling the data (Palechor & de la Hoz Manotas, 2019)[8].

In the baseline training, we selected five models: Logistic Regression, Support Vector Machine, K Nearest Neighbor, Decision Tree, and Random Forest (accuracy1 in Table 1).

Then we tried visualizing the data mainly by histograms and dropped Gender and Smoke, which were of low importance, but after the baseline training, the performance of the models did not improve (accuracy2 in table1), and most models even had a much worse performance than keeping all the features. Thus, in all of the following methods, we kept all the features except Weight and Height.

The next method we tried was to reduce the dimensionality of data. Since we have 14 features for Decision Tree and Random Forest and 23 features for other algorithms while the size of the data set is only 2111, we tried compressing the features to avoid overfitting. We used the techniques of Linear Discriminant Analysis and Principle Component Analysis, but there was still no improvement, as is shown in accuracy (LDA, n=6) and accuracy (PCA, n=14) in table1.

Table 1 Accuracy Results of Different Methods

method	**accuracy1**	**accuracy2**	**accuracy (LDA, n=6)**	**accuracy (PCA, n=14)**
Logistic Regression	57.21%	49.17%	56.50%	56.74%
SVM	77.54%	73.76%	76.83%	82.51%
KNN	79.67%	77.54%	77.07%	79.20%
Decision Tree	74.94%	72.58%	64.77%	68.79%
Random Forest	85.34%	85.58%	77.78%	81.80%

Then we tried to aggregate the predictions made by multiple classifiers to improve the performance, especially to correctly classify the data that were often misclassified to be the neighboring labels. In the selection of classifiers in this ensemble method, we only selected the model with test accuracy higher than 75%. With this method, we had a slight improvement on Random Forest (86.29%), and the confusion matrix is shown inTable2.

TABLE 2 CONFUSION MATRIX OF RANDOM FOREST

	Insufficient	**Normal**	**Overweight I**	**Overweight II**	**Obesity I**	**Obesity II**	**Obesity III**
Insufficient	47	4	1	0	1	0	0
Normal	5	32	3	2	4	1	0
Overweight I	0	5	47	3	4	1	0
Overweight II	1	4	2	55	4	4	0
Obesity I	0	1	1	2	64	2	0
Obesity II	0	3	0	0	0	67	0
Obesity III	0	0	0	0	0	0	53

The last method we tried was a two-stage classifier. Since the labels of the data set we used differed from the traditional classification of obesity level (Overweight Level I and Overweight Level II for our labels while only Overweight for traditional ones), we tried merging the labels of Overweight Level I and Overweight Level II into one label Overweight in the first stage. Here, we also tried two different two-stage classifiers.

- In the first two-stage classifier, we only merged Overweight Level I and Overweight Level II in the first stage, doing the six-class classification and further classified into Overweight Level I or Overweight Level II in the second stage if necessary. The performance of the first stage (86.05%) was already lower than the baseline training, so we did not do the second-stage training.
- In the second two-stage classifier, we merged Overweight Level I and Overweight Level II into one label (Label "1") and all the other labels into another one (Label "0") doing the binary classification. Then in the second stage, we classified the data into the exact label. The results are shown in **Table 3**.

TABLE 3 ACCURACY RESULTS OF TWO-STAGE CLASSIFIERS

stage		**accuracy**
First stage		88.89%
Second stage	Label “1”	91.54%
	Label “0”	93.86%
combined		84.16%

V. EXPERIMENTS AND RESULTS

As mentioned earlier, this dataset has a total of 16 features, and we divided these features into five categories. And then, we completed data visualization and analysis by category. The results of EDA (Exploratory Data Analysis) are as follows.

A. Basic Information

Here, the obesity index is 0 for insufficient weight, 1 for normal weight, and so on. We can see from

Fig. 2 that the average age of obese people is higher than that of normal people and underweight people. According to the scatter plot, for people over 25 years old, most of them are overweight or obese. However, people of Obesity III, the most serious obesity, only exist under the age of 26. There may be sample bias here.

As is shown in **Fig. 3**, the average index is nearly the same for males and females, but the distribution is quite different. The variance among females is larger than that among males.

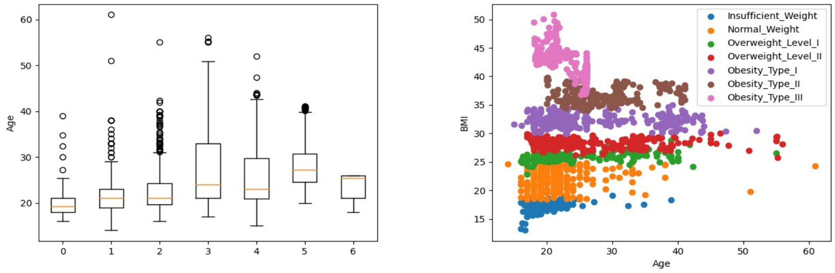

Fig. 2: Age

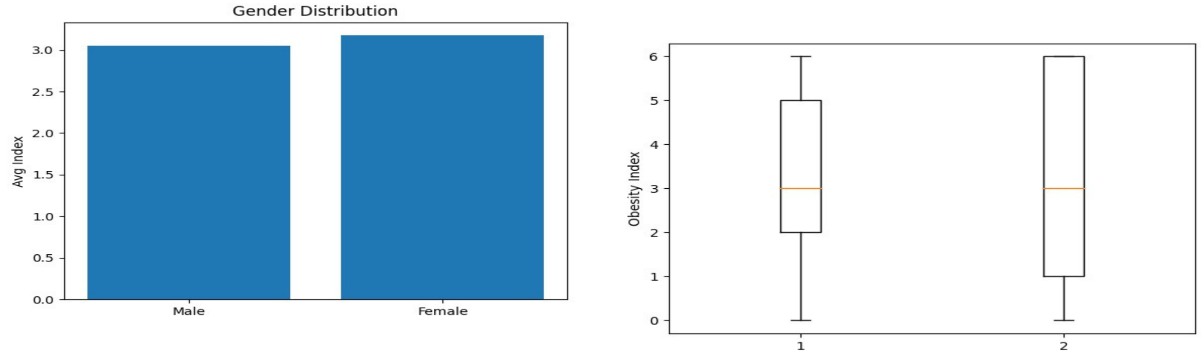

Fig. 3: Gender

B. Eating Habits

According to **Fig. 4**, when people eat high-calorie food frequently, they tend to have overweight problems, just as we expected. But we also found something interesting. Those who often or always eat vegetables have a high chance of being overweight, which is different from our usual perception. So we looked for related papers, and we learned that this result might be related to a balanced diet. Interestingly, in this dataset, when people eat more than three meals a day, they are more likely to maintain normal weight or underweight.

It is shown in **Fig. 5** that when people eat food between meals frequently, they tend to avoid overweight and obesity. Probably it is the same reason as above. But for water consumption, we found little evident pattern in the distribution of obesity. This is different from the papers we reviewed. Studies have shown that moderate water consumption can help maintain a normal weight.

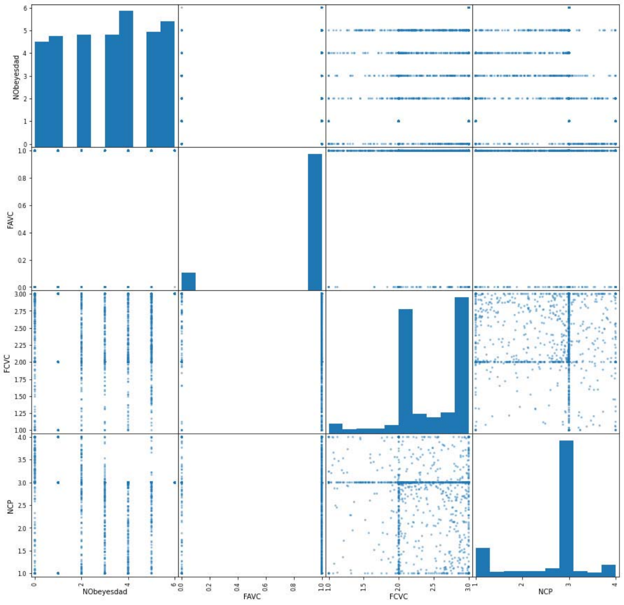

Fig. 4: Eating Habits (FAVC, FCVC, NCP)

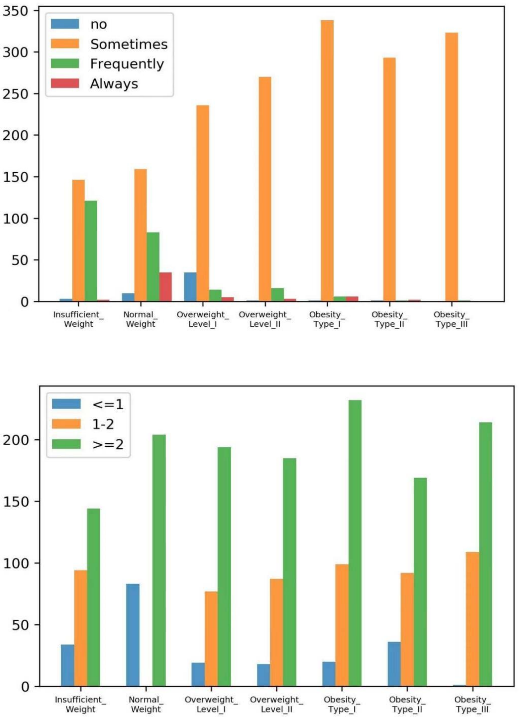

Fig. 5: Eating Habits (CAEC, CH2O)

C. Physical Condition

We found that 95% of people do not monitor calories. However, for those who monitor calories, most people have a BMI below 25, which means that if people have a habit of monitoring calorie intake, they will be more likely to maintain a normal weight. According to **Fig. 6**, in this dataset, most people do not have a physical activity or not often. As the frequency of physical activity increases, the number of people who are overweight is lower. We found that, in this dataset, when people do physical activity more than four days a week, nobody belongs to Obesity II or Obesity III. In this dataset, most people spend fewer than 3 hours using technological devices, and there is no obvious pattern.

According to **Fig. 7**, three quarters of people choose Public transportation, and the distribution of obesity among people is relatively even. However, people who choose Automobile are relatively more obese. There is also sample bias here. The number of people who choose Motorbike, Bike, and Walking is too small, making it hard to determine an obvious pattern. According to current observations, most people who like walking or biking maintain a normal weight.

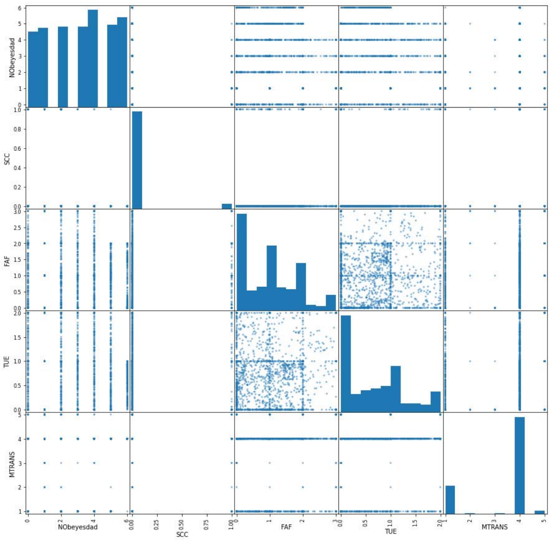

Fig. 6: Physical Condition

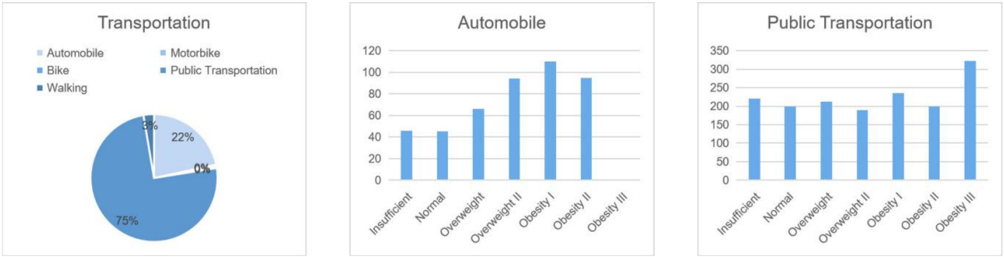

Fig. 7: Means of Transportation

D. Family History and Harmful Behaviors

Results of these features are shown in **Fig. 8**. We found that family history plays an important role in obesity. For those who are obese, family history is not a negligible factor. The relationship between smoking and obesity is negligible. However, for Alcohol Consumption, the relationship is quite obvious. No matter how often they drink, people who drink alcohol are more likely to be obese. However, this is just a rough finding. According to the previous literature review, moderate alcohol drinking is helpful for our health.

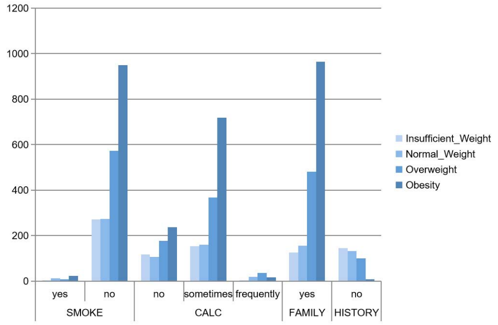

Fig. 8: Family History and Harmful Behaviors

E. Prediction Model

Overall we used a heat map (**Fig. 9**) to represent the correlation between features and found that most of them are weakly correlated. Among these features, family history shows a relatively strong correlation. Based on this result, we must comprehensively consider the relationship between each feature and obesity. This is also consistent with the fact in daily life that obesity is a complex problem.

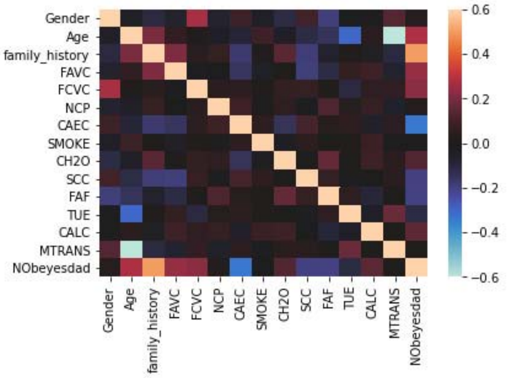

Fig. 9: Heat Map

After completing the EDA, we started to build the prediction model. We made a series of attempts, as mentioned in the Methods section. After that, we tried other new methods, such as XGBoost and LightGBM. When using different classification algorithms, it is important to tune the parameters, and we decided to use GridSearchCV from sklearn to finish our task. The accuracy results of different methods are shown in **Table 4**.

TABLE 4 ACCURACY RESULTS OF DIFFERENT METHODS AFTER TUNING PARAMETERS

method	accuracy
Logistic Regression	57.21%
Support Vector Machine	82.51%
K Nearest Neighbor	79.67%
Decision Tree (ID3)	79.08%
Decision Tree (CART)	79.22%
Random Forest	85.58%
XGBoost	85.99%
GBDT	85.58%
LightGBM	84.54%
Aggregate Prediction	86.29%

According to this result, aggregate predictions made by multiple classifiers indeed improved the performance with the best accuracy.

VI. CONCLUSION AND FUTURE WORK

Overall, the result we accomplished in this research is promising. In this project, we wanted to build a model that can help us to determine why people are overweight and predict whether a person will become overweight or not. At first, when we used all the given features, we did not get the result we were looking for. After looking into the relation of our features, we decided to eliminate some features. This new route we have is more promising. Then, we use different machine learning algorithms to build models such as Random Forest, Decision Trees. After about a week of tuning parameters and trying out other methods, we achieved a higher score in accuracy (refer to **Table 4** for more detailed information). After trying many different classification algorithms, we found obvious differences in the results of different algorithms related to their analytical principles and scope of application. For this dataset, it is obvious from the results that Decision Trees are more suitable. Especially when we used the integration algorithms of the Boosting framework, such as XGBoost and LightGBM, we achieved higher accuracy, with 85.99% and 84.54%, respectively. Besides, aggregate predictions made by multiple classifiers improved the performance a lot.

But the result is still not perfect enough. Despite efforts to tune parameters, we did not increase the accuracy to more than 90%. The result could be because the dataset we have is based only on individuals from Colombia, Peru, and Mexico. Since different nations have different diets and means of transportation, some of our features are imbalanced. In the future we are planning on using methods to balance out our dataset, and look into if a nation's affluence could affect the result of prediction.

AUTHOR'S CONTRIBUTIONS

Tong Cui (Vincent): Lead research, data visualization, adapting analysis, building prediction model of decision tree, XGBoost and writing the manuscript.

Yuxuan Chen (Jackie): data visualization, adapting analysis, building prediction model of logistic regression, SVM, KNN, GBDT and writing the manuscript.

Jiahao Wang (Johnny): data visualization, adapting analysis, building prediction model of logistic regression, KNN, LightGBM and writing the manuscript.

Haoran Deng (Paul): data visualization, adapting analysis, building prediction model of random forest, XGBoost and writing the manuscript.

Yuchen Huang (Sylvia): data visualization, adapting analysis, literature review and writing the manuscript.

REFERENCES

[1] WHO. "Obesity and Overweight." World Health Organization, World Health Organization, 1 Apr. 2020, www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] Kopelman, P. Obesity as a medical problem. Nature 404, 635–643 (2000). https://doi.org/10.1038/35007508

[3] Corica, D., Aversa, T., Valenzise, M., Messina, M., Alibrandi, A., De Luca, F., & Wasniewska, M. (2018). Does Family History of Obesity, Cardiovascular, and Metabolic Diseases Influence Onset and Severity of Childhood Obesity?. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5941161/.Retrieved 6 April 2021, from.

[4] Garawi, F., Devries, K., Thorogood, N., & Uauy, R. (2014). Global differences between women and men in the prevalence of obesity: is there an association with gender inequality?. https://www.nature.com/articles/ejcn201486.Retrieved 6 April 2021, from.

[5] Traversy, G., & Chaput, J. (2015). Alcohol Consumption and Obesity: An Update. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338356/.Retrieved 6 April 2021, from.

[6] Vij, V. and Joshi, A.(2014). Effect of excessive water intake on body weight, body mass index, body fat, and appetite of overweight female participants. https://pubmed.ncbi.nlm.nih.gov/25097411/.Retrieved 6 April 2021, from.

[7] Aparicio, A., Rodríguez-Rodríguez, E., Aranceta-Bartrina, J., Gil, Á., González-Gross, M., Serra-Majem, L., Varela-Moreiras, G. and Ortega, R., 2017. Differences in meal patterns and timing with regard to central obesity in the ANIBES ('Anthropometric data, macronutrients and micronutrients intake, practice of physical activity, socioeconomic data and lifestyles in Spain') Study. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5582404/.Retrieved 6 April 2021, from.

[8] 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, 104344.