What Makes You Obese?

Introduction to Data Science Capstone Project

Introduction

Overweight and obesity are major public health concerns in the United States, with a significant portion of adults affected. According to the 2017-2018 National Health and Nutrition Examination Survey, 30.7% of adults are classified as overweight and 42.4% as obese, with 9.2% affected by severe obesity. These conditions can lead to an increased risk of non-communicable diseases such as cardiovascular disease and diabetes. This report examines the contributing factors of overweight and obesity, which can be broadly categorized as diet and lifestyle. The report will analyze data using hypothesis tests, linear regression, and machine learning models to identify trends and make predictions on the Body Mass Index (BMI) and classify obesity.

Data and codebook

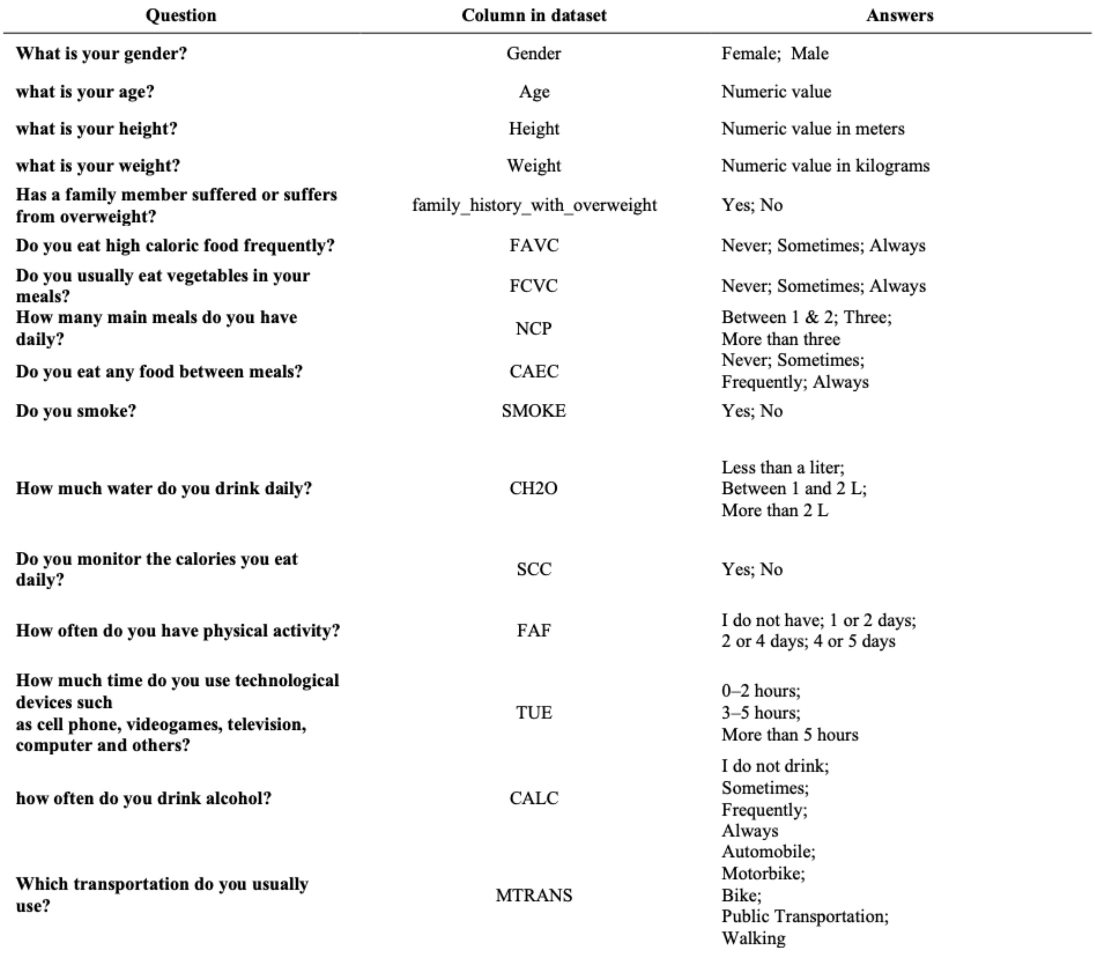
The data utilized in this project was collected from previous research on the topic of overweight and obesity. The dataset contains 17 attributes and 2111 records. A detailed description of each column and its corresponding values can be found in Table 1. The data is complete and contains no missing values. The data collector employed methods to address any missing data, resulting in some cells that are supposed to be integers being represented as float values in the dataset.

We calculated the Body Mass Index (BMI) for each record using the equation:

$$BMI = \frac{Weight}{Height^2}$$

Using the calculated BMI, each record was labeled with the class variable NObesity (Obesity Level) according to the criteria outlined in Table 2. To facilitate hypothesis testing, we further divided the NObesity variable into two categories – “is_obesity”, where respondents with Obesity I, Obesity II, or Obesity III are labeled as "1" and those with normal or overweight are labeled as "0".

Table 1: Questions of the survey used for initial recollection of information



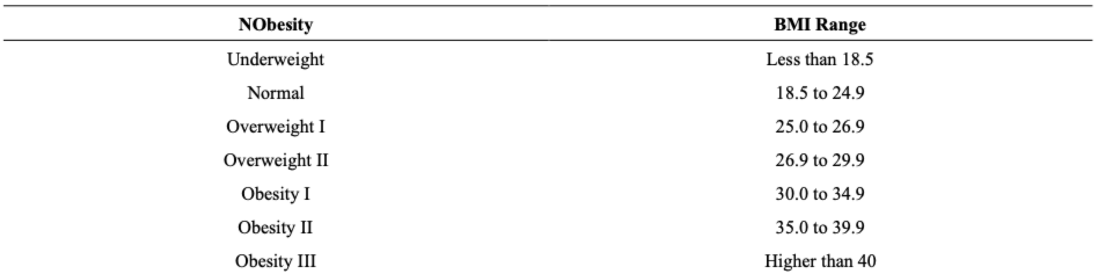
Question	Column in dataset	Answers
What is your gender?	Gender	Female; Male
what is your age?	Age	Numeric value
what is your height?	Height	Numeric value in meters
what is your weight?	Weight	Numeric value in kilograms
Has a family member suffered or suffers from overweight?	family_history_with_overweight	Yes; No
Do you eat high caloric food frequently?	FAVC	Never; Sometimes; Always
Do you usually eat vegetables in your meals?	FCVC	Never; Sometimes; Always
How many main meals do you have daily?	NCP	Between 1 & 2; Three; More than three
Do you eat any food between meals?	CAEC	Never; Sometimes; Frequently; Always
Do you smoke?	SMOKE	Yes; No
How much water do you drink daily?	CH2O	Less than a liter; Between 1 and 2 L; More than 2 L
Do you monitor the calories you eat daily?	SCC	Yes; No
How often do you have physical activity?	FAF	I do not have; 1 or 2 days; 2 or 4 days; 4 or 5 days

How much time do you use technological devices such as cell phone, videogames, television, computer and others?	TUE	0–2 hours;
		3–5 hours;
		More than 5 hours

how often do you drink alcohol?	CALC	I do not drink;
		Sometimes;
		Frequently;
		Always

Which transportation do you usually use?	MTRANS	Automobile;
		Motorbike;
		Bike;
		Public Transportation;
		Walking

Table 2: Body Weight Types by BMI Range



NObesity	BMI Range
Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight I	25.0 to 26.9
Overweight II	26.9 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

Part I: Inference Statistics

In this analysis, we sought to determine if there were differences in diet habits and lifestyles between individuals who are obese and those who are not. To answer this question, we conducted two types of hypothesis tests: comparisons of categorical (is_obesity) and numeric (FCVC, NCP, CH2O, FAF, TUE) variables, and comparisons of categorical (is_obesity) and categorical (family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC) variables. The significance level was set at 0.05.

For the first set of comparisons, we found that the data was not normally distributed and therefore used the Mann-Whitney U test (see Figure 1). The null hypothesis for this test was that there is no difference in diet habits and lifestyles between people who are obese and people who are not, and the alternative hypothesis was that there is a significant difference.

For the second set of comparisons, we used the Chi-Square Test to compare the observed and expected frequencies of diet habits among people who are obese and those who are not. The null hypothesis for this test was that there is no significant association between obesity and diet habits, meaning that the frequency distribution of diet habits among people who are obese is the same as the expected frequency distribution. The alternative hypothesis for this test was that there is a significant association between obesity and diet habits.

The results of our analysis are presented in Tables 3 (Mann-Whitney U test) and 4 (Chi-square Test). Table 3 shows that FCVC, CH2O, FAF, and TUE were statistically significant with small-medium effect sizes as measured by rank-biserial correlation. This suggests that these factors may have a moderate influence on the body weight of individuals. However, the p-value for NCP was 0.096, indicating that there is no statistically significant difference in the number of meals taken per day between obese and non-obese individuals.

Table 4 shows that every variable except for smoking habits had a significant result with a p-value close to 0, and small-medium effect sizes as measured by Cramer's V. This suggests that these factors may have a small-medium influence on the body weight of individuals.

In summary, our analysis suggests that some diet habits and lifestyles are statistically different between obese and non-obese people. Further research is needed to explore the extent to which these factors can influence a person's body weight.

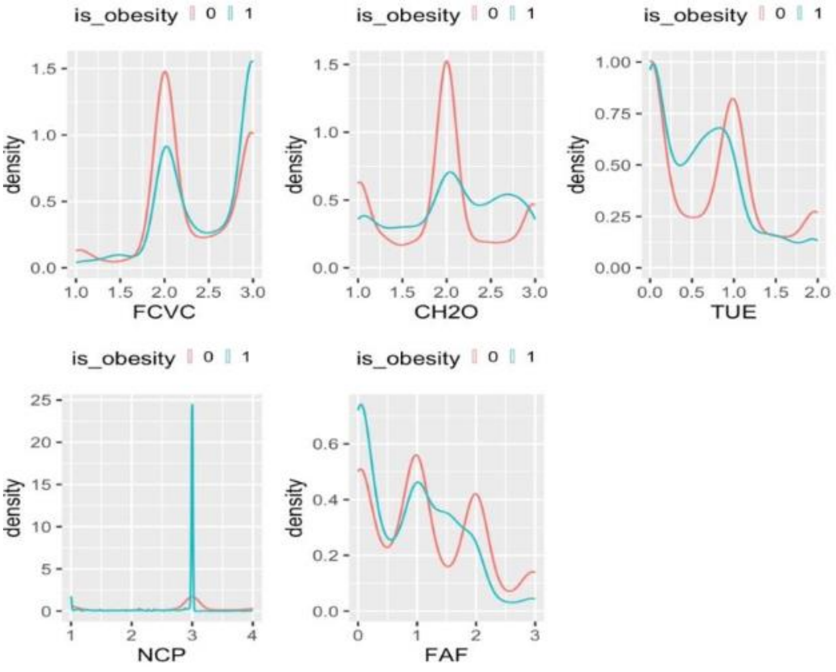
Table 3: Mann-Whitney U Test and Effect Size

group	group1	group2	n1	n2	p	Effect size
FCVC	0	1	1139	972	0.000	-0.218
NCP	0	1	1139	972	0.096	0.096
CH2O	0	1	1139	972	0.000	-0.139
FAF	0	1	1139	972	0.000	0.146
TUE	0	1	1139	972	0.004	0.071

Table 4: Chi-square Test Results and Effect Size

group	group1	group2	n1	n2	p	Effect size
family_history	0	1	1139	972	0.000	0.417
FAVC	0	1	1139	972	0.000	0.278
CAEC	0	1	1139	972	0.000	0.364
SMOKE	0	1	1139	972	0.705	0.012
SCC	0	1	1139	972	0.000	0.188
CALC	0	1	1139	972	0.000	0.157

Figure 1: Distribution of Numeric Variables



Part II: Prediction

To investigate if personal eating habits and physical conditions can predict health conditions, we used linear regression analysis. The dependent variable (Y) in the analysis was the Body Mass Index (BMI) of each person, which is a widely accepted and reliable indicator of healthy weight. The independent variables (X) in the analysis were 14 other variables related to eating habits and physical conditions. To consider the effect of gender and age, we controlled for these variables in the analysis.

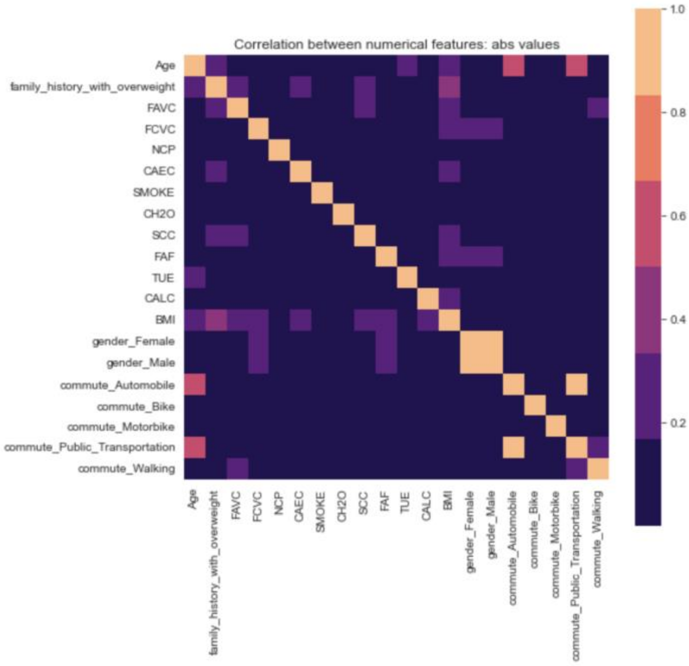
Before we can make predictions and classify the data, we performed some feature engineering. Specifically, we converted the following columns from character to categorical variables:

- Family_history_with_overweight
- FAVC (frequent consumption of high-caloric food)
- CAEC (consumption of food between meals)
- SMOKE (smoking habit)
- SCC (calories consumption monitoring)
- CALC (consumption of alcohol)

If the variable is binary (i.e., it has only two categories), we converted "yes" to 1 and "no" to 0. If the variable is a frequency measure (such as "no," "sometimes," "frequently," and "always"), we converted it to a 4-point Likert scale. This will allow us to better analyze the relationships between these variables and the target variable. Also, we used one-hot encoding to convert the gender and transportation used variables into numerical form for the model. By this approach, the model would learn and make predictions based on the individual categories, rather than assuming a linear relationship between the categories.

After feature engineering process, we examined the patterns of all predictors using a correlation matrix. The results are shown in Figure 2. We found that there were a few correlated features, such as age and using public transportation (r = 0.60), and age and using an automobile (r = 0.55). However, these two features are simply two categories within the transportation used variable, and the correlations were not particularly strong. As a result, we decided not to use dimension reduction methods before training our predictive models.

Figure 2: Correlation Heat Map



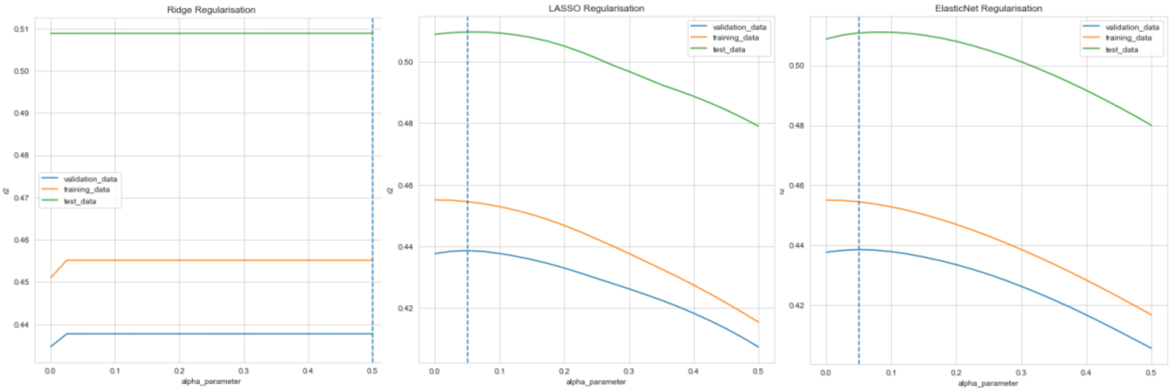
We performed multiple linear regression on our data, using variables related to eating habits and physical conditions as predictors and body mass index (BMI) as the target. We first scaled the features matrix and the BMI vector, then divided the data into a training set (80%) and a test set (20%). We trained the model using the training set and used it to make predictions on the test set. To evaluate the model's accuracy, we calculated the root-mean-square error (RMSE = 4.325) and the coefficient of determination (R^2 = 0.5096). We also compared the performance of the model with ridge regression, LASSO regression, and elastic net regression using hyperparameter tuning. The results of the models are summarized in Table 5. The elastic net model had the lowest RMSE of 5.690 and the highest R^2 of 0.5109, indicating it was the most effective.

Table 5: Model Comparisons

Method	alpha	R^2	RMSE
Ridge	0.5	0.5089	5.701
LASSO	0.05	0.5097	5.697
ElasticNet	0.05	0.5109	5.690

The best parameters using grid search methods are visualized as the following plots.

Figure 3: Hyperparameter Tuning



Part III: Classification

The goal of our analysis was to determine if personal eating habits and physical conditions can predict obesity. To accomplish this, we used machine learning models.

The original data had 7 categories: insufficient weight, normal weight, overweight (levels I and II), and obesity (types I, II, and III). Our goal was to classify whether an individual had obesity, so we labeled our data into two categories: 1 if the individual had obesity and 0 if they did not. This column was used as the outcome variable (Y), and the rest of 16 variables related to eating habits and physical conditions was used as the predictors (X). The data was relatively balanced, with 972 individuals having obesity and 1139 individuals not having obesity.

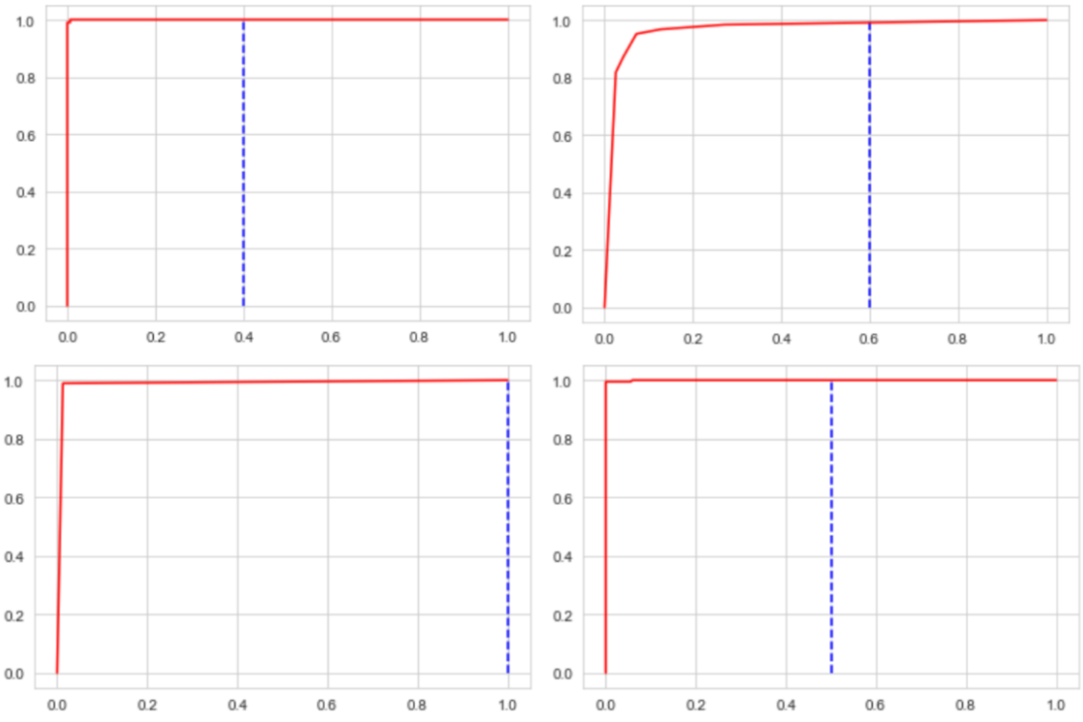
After scaling the eating habits and physical conditions features, we divided the data into a training set (80%) and a test set (20%). We then used logistic regression, K-Nearest-Neighbors (KNN), decision tree, and random forest to predict obesity. We calculated the area under the curve (AUC) values to assess the classification results, which are summarized in Table 6. Overall, the four models performed very well, with the logistic regression model performing the best with an AUC value of 0.9999.

Table 6: AUC Values

Method	AUC value
Logistic regression	0.9999
KNN	0.9726
Decision tree	0.9885
Random forest	0.9995

The ROC curve for the four models are shown below:

Figure 4: ROC Curves for Logistic regression (top-left), KNN (top-right), Decision tree (bottom-left) and Random forest (bottom- right)



Conclusion

Obesity and being overweight are major public health concerns that can have significant impacts on both physical and mental health, leading to various health problems such as heart disease, diabetes, and cancer. As students of Introduction to Data Science, we were curious to see if we could gain any insight into the factors contributing to obesity. In our inference phase, we tested whether obese people have different eating and lifestyle habits than non-obese individuals and found that, in most aspects considered, the answer was indeed yes. We also attempted to identify the best model for predicting obesity and find the most suitable parameters.

However, our project has some limitations, such as the inclusion of imputed data in our dataset, which may affect its representativeness. We hope to continue exploring this topic and gain a deeper understanding of what contributes to obesity.

Reference

- National Center for Health Statistics, Centers for Disease Control and Prevention. United States National Health and Nutrition Examination Survey 2017-2018. Hyattsville, United States of America: National Center for Health Statistics, Centers for Disease Control and Prevention.
- 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.
- World Health Organization. (n.d.). Body mass index (BMI). World Health Organization. Retrieved January 24, 2023, from https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index.