**Obesity levels based on eating habits and physical condition**

**Content**

List of Figures iii

List of Tables iv

1. Introduction 1
2. Dataset for estimation of obesity levels 2

2.1 Objectives 2

2.2 Source 2

2.3 Analysis 3

2.3.1 Numerical variables 3

2.3.2 Categorical variables 4

2.3.3 Relationships between obesity and potential risk factors 8

Appendix A Environment 13

Bibliography 14

**List of Figures**

Figure 2.1: Number of people withing each weight category. 3

Figure 2.2: Histograms of weight, BMI, height, and age by gender. 4

Figure 2.3: Bar charts of family history, high-caloric food intake, food consumption between main meals, and monitoring calories. 5

Figure 2.4: Bar charts of gender, smoking, alcohol consumption and main means of transportation. 6

Figure 2.5: Histograms of vegetable intake, number of main meals, water intake, technology use and physical activity. 8

Figure 2.6: Scatterplot of age and BMI. 9

Figure 2.7: Boxplots of BMI by family history, high-caloric food intake, food consumption between main meals, and monitoring calories. 10

Figure 2.8: Boxplots of BMI by gender, smoking, alcohol intake and transportation. 11

Figure 3.1: Boxplot of BMI and clusters from K means analysis with 7 clusters. 13

Figure 3.2: Sum of Squared Errors for different numbers of clusters. 14

Figure 3.3: Silhouette score for different numbers of clusters. 14

Figure 3.4: Boxplot of BMI and clusters from K means analysis with 8 clusters. 15

**List of Tables**

Table 2.1: Selected questionnaire questions and multiple answer options. 7

Table 3.1: Odds ratio and 95% Confidence Intervals. 18

Table 3.2: The final model predictions. 19

**Chapter 1**

**Introduction**

Obesity is a growing public health challenge (Boutari and Mantzoros, 2022). Approximately 59% of the adult population was either overweight or obese in the European region in 2022 (Boutari and Mantzoros, 2022). Obesity prevalence is increasing every year due to unhealthy diets and sedentary lifestyles (Zhang et al., 2023). The number of people with this disease tripled from 1975 to 2022 (Dominguez et al., 2023).

Obesity and overweight are major risk factors for multiple chronic diseases, including but not limited to diabetes mellitus, cardiovascular diseases, musculoskeletal diseases, various types of malignancies, and mental disorders (Dominguez et al., 2023). Obesity decreases life expectancy, negatively affects the quality of life, and raises costs for health systems (Boutari and Mantzoros, 2022).

According to the World Health Organization, BMI is used to classify overweight and obesity. Underweight is defined as BMI<18.5 kg/m2. Normal BMI for adults is 18.5 – 24.9 kg/m2. An individual with a BMI between 25.0 and 29.9 kg/m2 is considered overweight. Obesity is diagnosed when a person’s BMI is ≥30 kg/m2. Individuals with class 1 obesity have a BMI ranging from 30.0 to 34.9 kg/m2. Individuals with a BMI from 35.0 to 39.9 kg/m2 are considered to have Class 2 obesity. Class 3 obesity is diagnosed if the BMI is 40 kg/m2 or higher (Aronne, 2002).

Weight loss can help decrease the risk of many chronic diseases associated with overweight and obesity. There are multiple weight loss treatment options, including diet and lifestyle modification, medications, and surgery (Bray et al., 2018). Obesity management is only possible with intensive research on the topic and informative interventions for the most vulnerable groups. Although a lot of data on obesity is available now, there are still data gaps (Allman-Farinelli, 2023). The objective of this research is to provide more information on obesity and potential risk factors. Logistic regression has been used to predict overweight/obesity and evaluate risk factors. The study identifies clusters of patients based on eating habits and physical activity. Clustering analysis may be used for determining target groups for lifestyle interventions.

**Chapter 2**

**Dataset for estimation of obesity levels**

*2.1 Objectives*

Potential obesity-related eating habits and physical activity were examined to provide more information on the topic for future obesity prevention and intervention. In addition, the aim of this chapter is to identify dataset characteristics that might be relevant to future unsupervised and supervised analysis.

*2.2 Source*

This study used data on obesity, eating habits, and physical condition from 485 people in countries of Mexico, Peru, and Colombia. Data was collected via a web platform survey. 77% of the data was synthetically generated to make the number of individuals in each weight category similar. This dataset does not represent the population, because the data was synthetically generated. For this analysis, the full data set is not considered to be synthetically created. The data was preprocessed (deleted missing data, atypical data, etc.) before the balancing process. The final dataset contains 17 variables and 2111 records and does not contain missing values.

Data was obtained on responders’ family history of obesity, gender, age, height, and weight. In addition, information was collected on eating habits, like high-calorie food consumption, vegetable consumption, number of main meals, food consumption between meals, water intake, smoking, and alcohol consumption. Data obtained on physical condition were calorie monitoring, amount of physical activity, technology use, and main means of transportation (Palechor and Manotas, 2019).

*2.3 Analysis*

A total of 1552 individuals in the dataset are overweight or obese. Most weight categories are similar in size. The smallest number of people have insufficient weight (272 survey participants). The largest category containing 351 responders is obesity type I (Figure 2.1).

A graph of weight categories

Description automatically generatedFigure 2.1: Number of people withing each weight category.

*2.3.1 Numerical variables*

Figure 2.2 illustrates distributions of four numerical variables namely weight, BMI, height, and age. Women have lower median weight of 78.0 kg compared to men's median weight of 89.9 kg. However, there is a subgroup of women with very high weight. Only 0.1% of men have a BMI equal to or higher than 40 kg/m2, compared to 25.6% of women with BMI ≥40 kg/m2. Height has bimodal distributions explained by gender. Women are shorter than men with a median height of 1.64 m and 1.76 m for women and men respectively. Age distribution is right-skewed. Age ranged from 14 to 61 years with median age of 22.8 years. We can hypothesize that the web platform used to conduct the survey was more popular or accessible to the younger population.

A group of graphs showing different sizes of weights

Description automatically generated with medium confidenceFigure 2.2: Histograms of weight, BMI, height, and age by gender.

*2.3.2 Categorical variables*

Figures 2.3 and 2.4 illustrate counts of categorical variables. Many categorical variables have unbalanced subgroups. For example, only 1 responder stated drinking alcohol always. The means of transportation variable is very unbalanced as well: only 7 people ride bikes and 11 people use motorbikes. Analysis associated with low-frequency subgroups should be evaluated with caution because less frequent outcomes will have low precision in the further analysis (Cramer, 1999).

A graph of food consumption

Description automatically generated with medium confidence

Figure 2.3: Bar charts of family history, high-caloric food intake, food consumption between main meals, and monitoring calories.

A graph of different types of people

Description automatically generated with medium confidence

Figure 2.4: Bar charts of gender, smoking, alcohol consumption and main means of transportation.

Data on vegetable intake, number of main meals, water intake, technology use, and physical activity were collected via a questionnaire with multiple-choice questions. The questions and multiple-answer options are presented in Table 2.1. These variables are ordinal categorical. However, the above-mentioned variables are coded as numerical. Moreover, these variables have unique values with decimal places. Figure 2.5 illustrates that the values with the highest counts are integers. Possibly the integers were used to code categorical variables. It was discussed above that this dataset was preprocessed and balanced. Likely these manipulations created values with decimal places for categorical variables. Results associated with vegetable intake, number of main meals, water intake, technology use, and physical activity should be interpreted with caution.

Table 2.1: Selected questionnaire questions and multiple answer options.

|  |  |
| --- | --- |
| Question | Answer options |
| Do you usually eat vegetables in your meals? | Never  Sometimes  Always |
| How many main meals do you have daily? | Between 1 y 2  Three  More than three |
| How much water do you drink daily? | Less than a liter (L)  Between 1 and 2 L  More than 2 L |
| How often do you have physical activity? | 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? | 0–2 hours  3–5 hours  More than 5 hours |

A graph of different types of data

Description automatically generated with medium confidence

Figure 2.5: Histograms of vegetable intake, number of main meals, water intake, technology use and physical activity.

*2.3.3 Relationships between obesity and potential risk factors*

Figure 2.6 illustrates the positive linear relationship between BMI and age. It has been shown in previous research that BMI increases until around 60 years (Elia, 2001). It is hypothesized that a decrease in physical activity and basal metabolic rate might contribute to BMI increase with age.

A diagram of a body mass index

Description automatically generatedFigure 2.6: Scatterplot of age and BMI.

A median BMI for responders without a family history of obesity was 21.0 kg/m2 with the i**nterquartile range** (IQR) from 17.6 to 25.2 kg/m2 (Figure 2.7). A median BMI of 31.6 kg/m2 (IQR, 26.2 – 36.9 kg/m2) was calculated for individuals with family history of obesity. This value was higher compared to participants with no family history.

Responders not consuming high-calorie food regularly had lower median BMI than individuals who eat high-calorie food regularly(Figure 2.7). The median BMI was 24.4 kg/m2 (IQR, 19.7 – 28.0 kg/m2) for the group not consuming high-caloric food regularly. The median BMI of 30.7 kg/m2 (IQR, 25.4 – 36.4 kg/m2) was higher for patients eating high-calorie food.

Responders frequently consuming food between main meals had the lowest medium BMI of 18.9 kg/m2 (IQR, 17.6 – 23.5 kg/m2) followed by individuals always consuming food between meals with a medium BMI of 23.8 kg/m2 (IQR, 21.0 – 26.1 kg/m2) and participants never eating between main meals with medium BMI of 26.4 kg/m2 (IQR, 25.3 – 26.7 kg/m2) (Figure 2.7). The highest medium BMI of 31.2 kg/m2 (IQR, 26.0 – 36.7 kg/m2) was calculated for responders sometimes consuming food between meals. These results might not be representative, because only 53 individuals responded always, and 51 participants responded no.

Figure 2.7 illustrates that Individuals who count calories had a lower BMI of 24.2 kg/m2 (IQR, 19.3 – 25.4 kg/m2). The median BMI for responders not monitoring calories was 29.3 kg/m2 (IQR, 24.8 – 36.2 kg/m2).

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure 2.7: Boxplots of BMI by family history, high-caloric food intake, food consumption between main meals, and monitoring calories.

A median BMI for female responders was 28.5 kg/m2 with the **interquartile range** from 22.4 to 40.1 kg/m2 (Figure 2.8). Median BMI men was 28.9 kg/m2 (IQR, 25.6 – 35.2 kg/m2).

Figure 2.8 shows that smokers and non-smokers had comparable median BMI of 29.7 kg/m2 (IQR, 24.2 – 36.1 kg/m2) and 28.7 kg/m2 (IQR, 24.3 – 36.0 kg/m2) respectively. Only 44 individuals smoke, more data should be collected to represent the group of patients who smoke.

The highest median BMI of 30.9 kg/m2 (IQR, 25.4 – 38.0 kg/m2) was calculated for individuals who sometimes drink alcohol(Figure 2.8). Responders who never drink alcohol had the second highest median BMI of 28.0 kg/m2 (IQR, 21.2 – 32.1 kg/m2). 70 individuals consume alcohol frequently and had a median of BMI 26.9 kg/m2 (IQR, 24.7 – 28.9 kg/m2). Only one person with a BMI of 22.5 kg/m2 reported always consuming alcohol.

Figure 2.8 shows that the highest median BMI of 29.0 had responders that usually use automobiles (IQR, 25.8 – 33.3 kg/m2) or public transportation (IQR, 24.2 – 36.6 kg/m2). A group of 11 responders who usually use motorbikes had median BMI of 24.4 kg/m2 (IQR, 22.8 – 29.7 kg/m2). Only 7 people usually ride bikes and their median BMI was 24.2 kg/m2 (IQR, 21.6 – 25.7 kg/m2). The lowest BMI of 23.6 kg/m2 (IQR, 21.2 – 25.4 kg/m2) was calculated for a group of 56 responders that usually walk. In conclusion, people using physically demanding means of transportation (like biking or walking) have lower BMI compared to individuals who drive.

A group of boxes with different colored squares

Description automatically generatedFigure 2.8: Boxplots of BMI by gender, smoking, alcohol intake and transportation.

CONCLUSION

The results of this study add to the body of literature on overweight and obesity. These results can be used to promote healthy lifestyle and develop weight management programs. Future research on the topic could be important for decreasing the prevalence of overweight and obesity.

**Appendix A Environment**

Language: Python version: 3.11.5 | packaged by Anaconda, Inc.

**Bibliography**

Allman-Farinelli, M. (2023) ‘Nutritional Strategies to Prevent Weight Gain and Obesity’, *Nutrients*;15(19),4180.

Aronne, L. J. (2002) ‘Classification of obesity and assessment of obesity-related health risks’, *Obesity research*, 10 (2), 105S–115S.

Boutari, C., & Mantzoros, C. S. (2022) ‘A 2022 update on the epidemiology of obesity and a call to action: as its twin COVID-19 pandemic appears to be receding, the obesity and dysmetabolism pandemic continues to rage on’, *Metabolism: clinical and experimental*, 133, 155–217.

Bray, G. A., et al. (2018) ‘The Science of Obesity Management: An Endocrine Society Scientific Statement’, *Endocrine reviews*, *39*(2), 79–132.

Cramer, J. S. (1999) ‘Predictive Performance of the Binary Logit Model in Unbalanced Samples’, Journal of the Royal Statistical Society Series D: The Statistician, 48(1), 85–94.

Dominguez, L. J. et al. (2023) ‘Mediterranean diet in the management and prevention of obesity’, *Experimental gerontology*, 174, 112–121.

Elia, M. (2001) ‘Obesity in the elderly’, *Obesity Research*, 9(S11), 244S–248S.

Hwalla, N., & Jaafar, Z. (2020) ‘Dietary Management of Obesity: A Review of the Evidence’, *Diagnostics (Basel, Switzerland)*, 11(1), 24.

Palechor, F. M., & Manotas, A. H. (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.

Robertson, C, Browne, A., Stewart, R (2023) “MM924: Statistical Modelling and Analysis” University of Strathclyde, Glasgow.

Zhang, X., Ha, S., Lau, H. C., & Yu, J. (2023) ‘Excess body weight: Novel insights into its roles in obesity comorbidities’, *Seminars in cancer biology*, 92, 16–27.