Final Project

Obesity Levels Based On Eating Habits and Physical Condition

3/18/2024

Prepared by
Alex Camilleri
Anthony Montano
Fabian Gutierrez
Jacob Korte

Table of Contents pg

1. Introduction

1.1 Subject
Obesity levels based on eating habits and physical activity.

1.2 Purpose
By using data analysis techniques we can investigate the factors that contribute to obesity levels and obtain the ability to predict if an individual may be at risk of obesity in the future.

1.3 Data Analysis Techniques

- Supervised learning: KNN
- Unsupervised learning: K-Means Clustering
- Feature Selection

Some of the data analysis techniques we plan to use are K-Nearest Neighbors classification(KNN), K-Means Clustering, Feature Selection. The KNN algorithm will allow us to classify a new respondents' levels of obesity based on their answers to the survey and other respondent's data. We will also plan to use K-Means Clustering to predict a participant's level of obesity based on their assigned cluster. Feature Selection is used to see what major attributes impact a level of obesity.

2. Data Set

2.1 Description

The dataset is obtained from “Estimation of Obesity Levels Based On Eating Habits and Physical Condition” by Fabio Palechor and Alexis Manotas from UC Irvine Machine Learning Repository. This dataset contains 17 features and 2111 responses from individuals from Colombia, Peru and Mexico about their personal attributes, dietary habits and their physical health.

2.2 Cleaning

Some features in the dataset were categorical and in order to use the data for analysis, features were changed to numerical in the range of 0-n depending on the number of options available.

2.3 Features

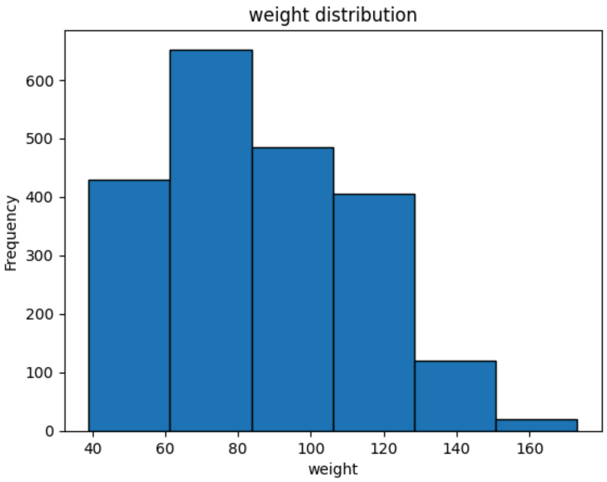
1. Gender
2. Age
3. Height
4. Weight
5. Has a family member suffered or suffers from being overweight?
6. Do you eat high caloric food frequently?
7. Do you usually eat vegetables in your meals?
8. How many main meals do you have daily?
9. Do you eat any food between meals?
10. Do you smoke?
11. How much water do you drink daily?
12. Do you monitor the calories you eat daily?
13. How often do you have physical activity?
14. How much time do you use technological devices such as cell phones, video games, television, computer and others?
15. How often do you drink alcohol?
16. Which transportation do you usually use?
17. Obesity level

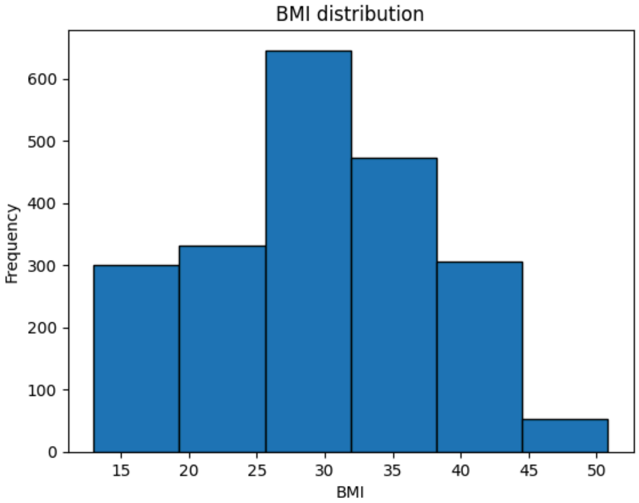
2.4 Reference

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

3. Exploratory Data Analysis (EDA)

Based on the data set, we first explored the sample's weight distribution and BMI using histograms. The histogram of weights shows that the majority of the sample falls within the 60 to 80 kg range. Similarly, the BMI histogram reveals that the average BMI of the sample falls between 25 and 30. This range indicates that the average BMI of the sample is in the overweight to obese category, according to the World Health Organization's classification. These findings suggest that the sample population may have a higher prevalence of overweight and obesity.



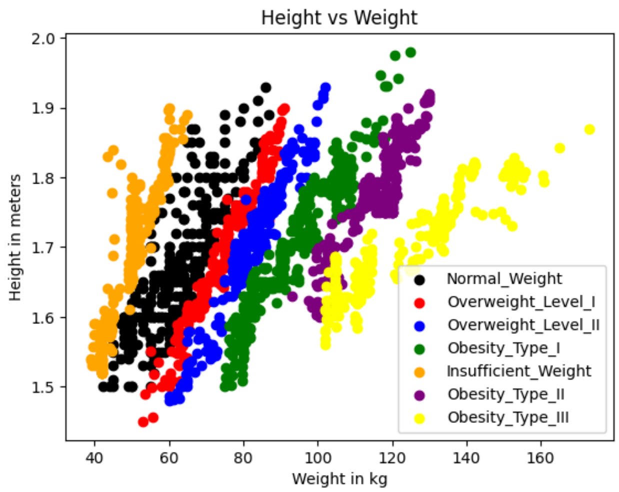


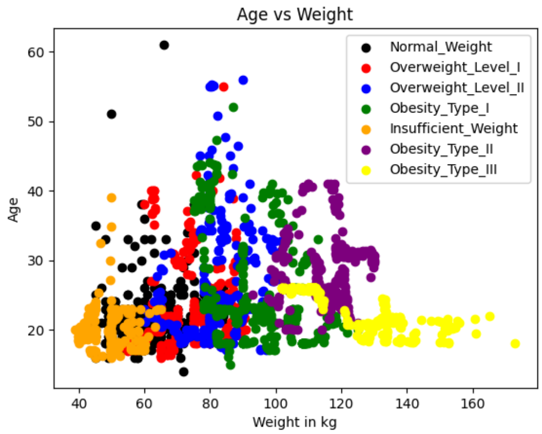
Based on the feature selection algorithm, we found that what correlated the most predicting obesity was weight, age, gender, and family history. For weight, a scatter plot of weight versus height is used, with weight classes being categorized by colors. We see a very clear separation for obesity class and weight, which is fairly self-evident, since obese people tend to weigh more.

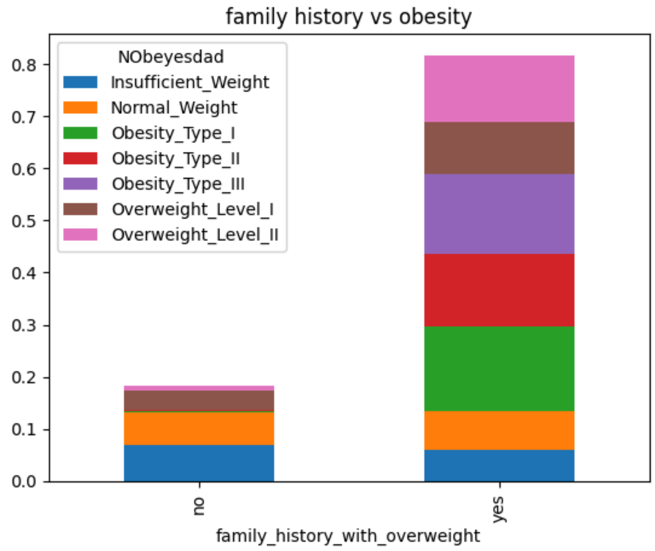
The other feature I showcased was age vs weight in a scatter plot. The separation between the different weight classifiers is not as significant, but it still shows a general trend that most obese people are younger. Surprisingly, Obese type 3 people are strictly in the range of 20–30 years old.

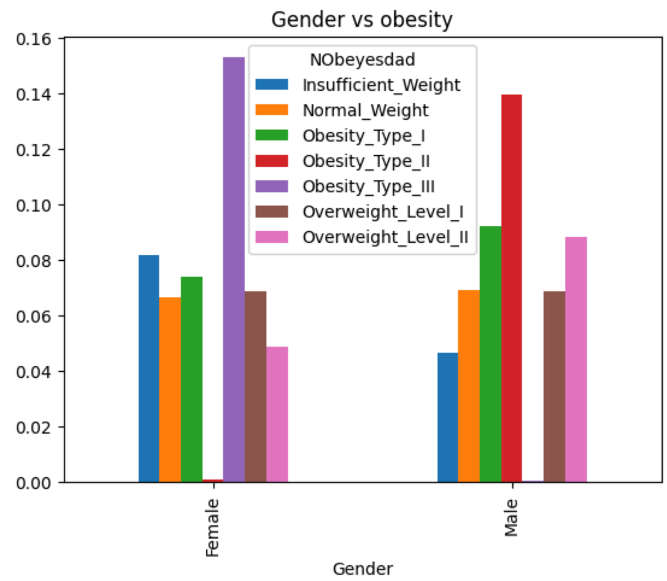
The stacked bar chart of family history provided some insightful results. There are two categories, people who have a family history with obesity and people who do not. In the group who answered no, we see that the majority of the group is either, insufficient weight, normal weight, or overweight 1. For the people who answered yes, we see that a majority of the bar chart is made up of people who are over weight to obese. This bar chart shows how a genetic component me be key in obesity.

For the gender feature, it shows how the majority of obese_type 3 people being female and obese_type 2 being male. Other than that the distributions are mostly similar, It seems that in general though, for females there is a more polarized wight decision with most of the frequencies in obese_type 3 and insufficient weight.


Height vs Weight
Height in meters
2.0
1.9
1.8
1.7
1.6
1.5
40
60
80
100
120
140
160
Weight in kg
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III


Age vs Weight
Age
60
50
40
30
20
40
60
80
100
120
140
160
Weight in kg
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III

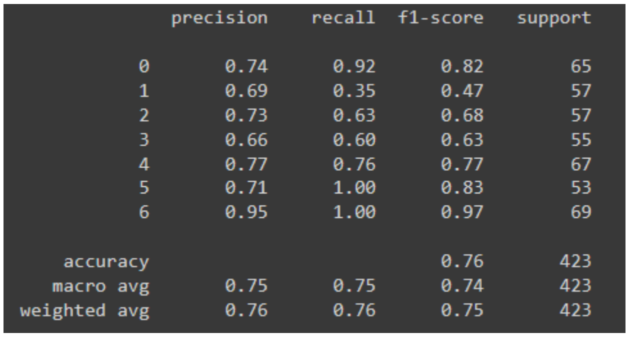

family history vs obesity
NObeyesdad
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
no
yes
family_history_with_overweight



4. Supervised Learning Technique

4.1 Knn

We used the Knn algorithm to classify our data into 7 different categories (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III). For our K we chose an odd number(19) since we have an even number of data points. We evaluated the performance using a classification matrix and calculating the precision, recall, etc.



```
              precision    recall  f1-score   support

           0       0.74      0.92      0.82        65
           1       0.69      0.35      0.47        57
           2       0.73      0.63      0.68        57
           3       0.66      0.60      0.63        55
           4       0.77      0.76      0.77        67
           5       0.71      1.00      0.83        53
           6       0.95      1.00      0.97        69

    accuracy                           0.76       423
   macro avg       0.75      0.75      0.74       423
weighted avg       0.76      0.76      0.75       423
```

5. Unsupervised Learning Technique

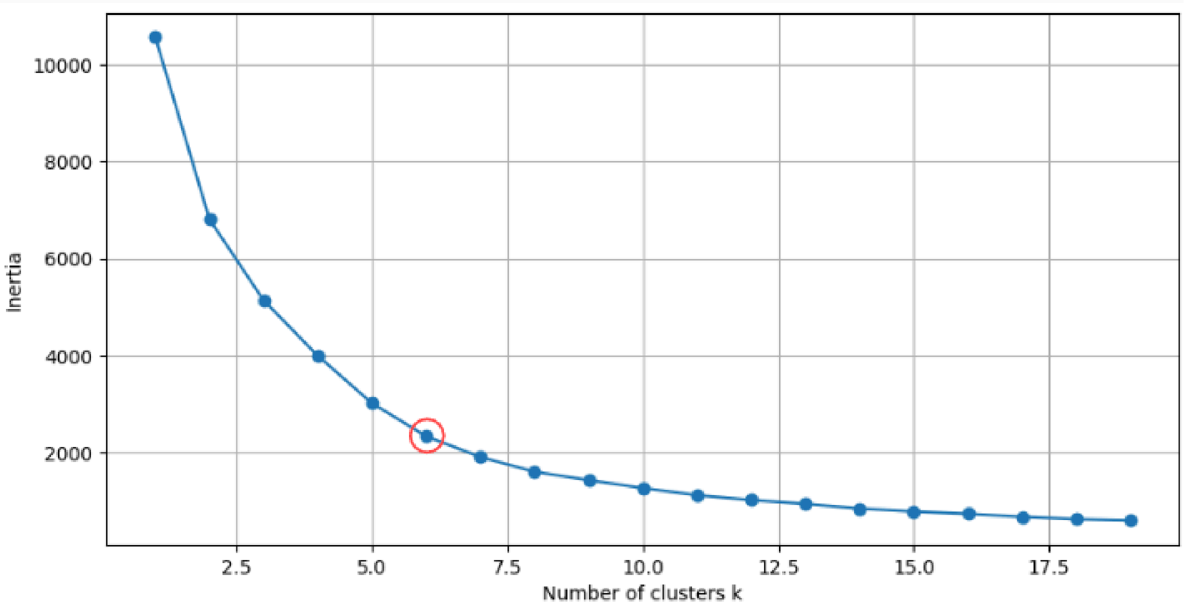
5.1 Feature selection

To figure out which features out of the 16 contributed most to the classification they were given, we used the feature selection algorithm shown in class. The algorithm returned the top 4 features (Gender, age, family history with obesity, and weight) that had the most significance.

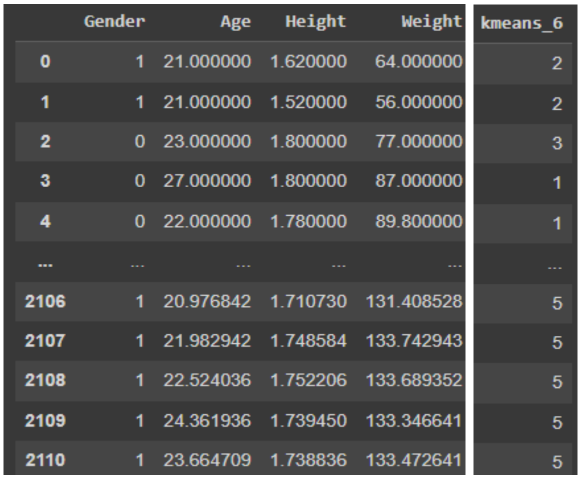
5.2 K-Means

K-means as an unsupervised learning technique allows for the clustering of values based on dimensional similarity between features. Given an amount of cluster centers, similar values are grouped together and organized. The mean is then calculated for each cluster, and its center moved to the mean location, followed by regrouping all values. The process is repeated until there is no difference in the groups between iterations.

K-means is used in this project to clump together similar categories to determine the level of obesity in each group. The dataframe is applied as a scalar for determining inertias, which are used to map out the elbow plot for determining the k-value. Using the four selected features, the graph showed the elbow curving around k=6 and slowing the change in inertia.



The k-value for the amount of clusters was applied to the dataframe for gender, age, weight, and family history of obesity. The grouped clusters were then appended to the dataframe as its own column, showing which values in those four groups fit into which cluster, grouping them based on predicted obesity level.



	Gender	Age	Height	Weight	kmeans_6
0	1	21.000000	1.620000	64.000000	2
1	1	21.000000	1.520000	56.000000	2
2	0	23.000000	1.800000	77.000000	3
3	0	27.000000	1.800000	87.000000	1
4	0	22.000000	1.780000	89.800000	1
...	...	...	...	...	...
2106	1	20.976842	1.710730	131.408528	5
2107	1	21.982942	1.748584	133.742943	5
2108	1	22.524036	1.752206	133.689352	5
2109	1	24.361936	1.739450	133.346641	5
2110	1	23.664709	1.738836	133.472641	5

6. Conclusion

We were successful in predicting the obesity levels of the data points. After analyzing the data further we found that Gender, age, family history with obesity, and weight contribute the most out of the 16 features to a person's obesity level.

Jacob - EDA, slides
Alex - K mean clustering, slides
Anthony - feature selection, KNN, slides
Fabian - reports, slides