Can Unsupervised Machine Learning be used to identify groups in lifestyle behavior that correspond to the different levels of obesity?

By: Darien Chiang, Paul Adutwum, Peter Lee

Introduction:

Obesity and being overweight in our current day and age is a growing global public health concern which, according to the WHO, caused over 3.7 million deaths in 2021 (World Health Organization 2023). Linked to cardiovascular disease, diabetes, and other chronic conditions these two classifications are not to be taken lightly. To the unperceptive onlooker, one might think that the condition of being obese and overweight is a simple matter to handle where one affects just exercising more and eating healthier; however, upon closer look there is much more to unpack. Beyond just simple body weight, obesity and overweightness are shaped by combinations of dietary habits, physical activity, daily routine, genetic history, and environment. Therefore understanding how these behaviors cluster together across individuals can provide insights into different so-called types of lifestyles that may carry different levels of health risks on this front. To clarify our usage of the terms "obesity and overweight", we point to the WHO's definitions where obesity is "a chronic complex disease defined by excessive fat deposits that can impair health" (World Health Organization 2023) and overweight is "a condition of excessive fat deposits (World Health Organization 2023).

Although metrics such as "BMI" exist to indicate the healthiness of a person's weight relative to their height, it actually reveals very little about the behaviors that give rise to obesity. "BMI" alone cannot capture the combinations and effects of diet, hydration, physical activity, and lifestyle behaviors. This makes unsupervised learning a natural choice for exploring whether behavioral and physiological patterns emerge directly from the data itself. Offering two key advantages unsupervised machine learning contains dimensionality reduction techniques such as Principal Component Analysis (PCA), and clustering methods such as k-means and Gaussian Mixture Models (GMM). By combining these techniques we shift from just exploring individual variables to analysing broader implications that may give us insight into groups at risk. This leads us to our research question:

Can unsupervised machine learning methods be used to identify groups in lifestyle behavior that correspond to different levels of obesity?

To answer this question, we will primarily work along the following lines: (1) Clean and Explore the dataset, (2) Reduce dimensionality using Principal Component Analysis, (3) Apply k-means and Gaussian Mixture Models to the PCA representation, (4) Validate cluster quality

using Silhouette Score for both clustering models, (5) Interpret visually the clusters with respect to the obesity levels, and (6) Discuss the strengths, weaknesses, and implications of our models.

In order to conduct our analysis and plan though, the first step is to determine whether such a dataset exists that contains all of the aforementioned metrics. On UC Irvine's Machine Learning Repository we were able to find such a dataset created by (Banos, Giménez, and Hervás-Martínez 2015), called "Estimation of Obesity Levels Based On Eating Habits and Physical Condition.". This is a dataset that contains observations of 2,111 individuals from Mexico, Peru, and Columbia with their corresponding demographics: lifestyle habits and family history relating to obesity and overweightness amounting to a total of 18 variables. For the sake of our study we will restrict our analysis to the following 5 ordinal and 1 categorical variables.

1. FCVC (frequency of vegetable consumption) - ordinal variable on a scale from 1 to 3, indicating how frequently the individual consumes vegetables.
2. NCP (number of main meals per day) - ordinal variable on a scale from 1 to 4, representing the typical number of main meals eaten per day.
3. CH2O (daily water consumption) - ordinal variable on a scale from 1 to 3, approximating increasing levels of liters consumed per day.
4. FAF (physical activity frequency) - ordinal variable on a scale from 0 to 3, representing increasing levels of weekly physical activity (ranges, not exact hours).
5. TUE (time spent using technological devices) - ordinal variable on a scale from 0 to 2, indicating approximate ranges of hours spent per day using technology.
6. NObeyesdad (obesity category) - categorical (nominal) variable with seven labeled categories: Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III.

We chose to use the 5 ordinal variables because other variables, even if they were ordinal, had categories like "Sometimes", or "yes / no", which were not straightforward in terms of quantifiability.

Data Visualization Introduction:

In an initial exploration of our dataset, we found that many of the variables that indicated lifestyle habits or family history were ordinal or categorical, while there were very few variables with continuous data. In case there were, these were variables tied directly to the calculation of obesity, such as height and weight. This hinted that we may benefit more from analysis that incorporates grouping with regard to some metric rather than other forms of analysis, such as regression. As a way to test if there were any recognizable patterns, even if vague or inconclusive, we generated the following two plots of two lifestyle habit variables colored by obesity level.

Although we could not make any solid claims here due to the presence of noise, the concentration of certain obesity levels for certain lifestyle habits confirmed our suspicion that there may indeed be some grouping happening with regard to lifestyle habits, and that those groups may correspond to specific obesity levels. In **Figure 1**, this is evident with the column of green dots (FCVC = 2.0) and blue dots (FCVC = 3.0), which shows a concentrated alignment of Obesity_Type_I and Obesity_Type_III for certain amounts of vegetable consumption frequency. In **Figure 2**, similarly, we see a noticeable concentration of Obesity_Type_III in the intersection of (CALC = Sometimes, Smoking Status = no). This indicates that perhaps, if we could find a way to incorporate more variables simultaneously while keeping the coloring by obesity levels, we could identify some meaningful correspondence between certain lifestyle habits and obesity levels. We will elaborate more on how we achieved this in the following section.

Methods:

Because our goal was to identify groups in lifestyle behavior that correspond to different levels of obesity we wanted to see if the clustering corresponding to variances in certain lifestyle habits could match individuals' obesity levels. In other words, we wanted to see whether the data itself would tell a story. Thus we choose to implement unsupervised machine learning techniques — using models to identify groups within data without defining labels ourselves. Below are the models that we used:

1. Principal Component Analysis
2. K-Means Clustering in PCA representation
3. Gaussian Mixture Modeling in PCA representation
4. Validation via the Evaluation of Silhouette Score for k-Means and GMM

Beginning with (1: PCA) we focused on 6 variables with 5 of them being lifestyle related ordinal: vegetable consumption (FCVC), number of meals (NCP), water intake (CH2O), physical activity frequency (FAF), technology use time (TUE). PCA was used in order to reduce these five variables into a smaller set of principal components that display the greatest variance in the data.

Procedural Steps (PCA):

1. Center each variable by subtracting its mean
2. Standardize variables to set variance of each variable to 1.
3. Compute eigenvalues and eigenvectors of the covariance matrix
4. Projecting the original data onto the principal components.

Moving onto (2: k-means), k- means is a hard assignment model (meaning that each point is assigned to one cluster) that partitions data into k-number of clusters through minimizing the

within cluster sum of square distances (WSS).

Procedural Steps (k-means):

1. Prepare the data to be used. In this case we are using the PC data from the PCA above.
2. Compute Total Within Cluster Sum of Squares for k = 1:10 given by:

$$tot.withinss = \sum_{k=1}^{k} W(C_k) \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

   whereby x_i is a data point in cluster C_k, and μ_k is the mean value of points assigned to cluster C_k.
3. Generate Elbow plot based on WSS values to determine the best value of k
4. Fit the final k-means plot using the best value of k and visualize the clusters

Next for (3: GMM), GMM is a soft assignment model (meaning that each point is assigned a probability of belonging to a cluster) that models the PCA data as a mixture of 2-D Gaussian distributions.

Procedural Steps (GMM):

1. Prepare the data to be used. Again we are using the PC data from the PCA above.
2. Using Mclust() we determined the best value of G.
3. Fit the GMM using the data and the optimal value of G determined by Mclust().

$$GMM = \sum_{i=1}^{G} \pi_i N(x|\, \mu,\, \sigma_i^2)$$

   whereby G is the number of Gaussians, of which each Gaussian *N* has its own mean μ_i, variance σ_i^2, and weight π_i. Further recall that GMM utilizes the idea of responsibility r_{ni}, or the responsibility of each Gaussian *i* for data point *n*.
4. Visualize the density plot, uncertainty, and classification plots.

Finally for (4: Silhouette Score), silhouette score determines how "good" the clustering from our k-means and GMM are for our dataset, which in our case is in 2-D PC space. The score ranges from -1 to 1, whereby a value closer to 1 indicates a correct clustering, 0 means it is between two clusters, and -1 reflects an incorrect clustering. All calculations regarding the Silhouette scores were done using the "Silhouette" package from Shrikrishna (2023).

Procedural Steps (Silhouette Score):

1. Compute the proximity matrix from the dataset in PC space. The "proximity" is the Euclidean distance between a point and the center of a cluster, which is repeated for all data points in the PCA space. The Euclidean distance formula in question is

$$Distance = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2}$$

whereby x_1, x_2 are the positions of data in PC space and c_1, c_2 are that for the cluster center.

2. Determine the silhouette score using the formula for each point in a cluster.

$$s(i) = \frac{b(i)-a(i)}{max(a(i), b(i))}$$

whereby for some *i*th point the score is determined by taking b (the distance from point *i* to the nearest other cluster center), subtracting a (the distance from point *i* to its own cluster center), and divide it by the larger between a(i), b(i).

3. Find the average silhouette score for each cluster.
4. Generate the silhouette plot.

Results:

PCA Results:

Utilizing R's built in function prcomp(), we applied PCA to five lifestyle variables- FCVC, NCP, CH2O, FAF, and TUE- after following the associated procedural steps. The resulting two dimensional space is shown in **Figure 3**: 2D PCA Representation with Obesity Level. Each point is an individual of 2,111 projected into the dimension-reduced 2-D subspace where the coloring correlates with the corresponding obesity level (NObeyesdad). The x-axis is the first principal component PC1 and the y-axis the second principal component PC2.

In order to quantify our principal components, we must examine the variable loadings shown in **Appendix B: Table 1**. PC1 by nature represents the pattern responsible for the most variation in our five lifestyle habits which by analyzing the table is characterized by variable loadings of (1) Frequency of Vegetable Consumption: 0.209, (2) Number of Meals: 0.477, (3) Water Intake: 0.557, (4) Physical Activity Frequency: 0.629, (5) Technology Use Time: 0.149. We observe that the highest contributions come from (2) Number of Meals (NCP): 0.477, (3) Water Intake(CH2O): 0.557, (4) Physical Activity Frequency(FAF): 0.629. This strongly suggests that PC1 is an axis that represents healthy lifestyle habits with high PC1 values correlating to a healthy lifestyle and low PC1 value to an unhealthy lifestyle.

On the other hand analyzing **Table 1** for PC2, a contrasting pattern emerges characterized by variable loadings of (1) Frequency of Vegetable Consumption: -0.684, (2) Number of Meals: 0.0567, (3) Water Intake: -0.125, (4) Physical Activity Frequency: 0.128, (5) Technology Use Time: 0.705. We observe that the highest contributions come from (1) Frequency of Vegetable Consumption(FCVC): -0.683, (5)Technology Use Time (TUE): 0.705. Therefore a higher PC2 value corresponds with higher technology use and lower vegetable intake, and a lower PC2 value corresponds with lower technology usage and higher vegetable consumption.

Furthermore, an initial look at the PCA plot reveals a few key observations. First, obesity type III (cyan) clusters the most in regions of lower PC1 (low meals, low water consumption, low physical activity), higher PC2 (low vegetable intake, higher tech usage) in the second quadrant. Second, apart from obesity type III there seems to be very little, at least visually, clustering for the different obesity levels.

K-means Clustering Results:

Following the procedural steps outlined for k-means, we generated the elbow plot in **Figure 4**, which shows an elbow-like reduction in the total within the cluster sum of squares as the number of k increases. This is expected as increasing the number of k clusters should lower the sum of the Total WSS, but at some threshold the amount of lowering should converge at some k. As evident in the Figure, this occurs around k = 6, 7. We chose to proceed with k = 6 clusters.

Doing so yielded the result in **Figure 5**, which shows six clusters indicated by different colors. We see that (1) cluster 2 is very compact around the center while clusters 4 and 6 are more concentrated as they approach the center but are more sparse at the opposite edges. The other clusters seem more scattered compared to clusters 2, 4, and 6. These clusters, which arguably are concentrated around some point (PC1, PC2) in the PC space, display the grouping pattern that we mentioned. As we established that PC1 and PC2 correspond to high variability of specific variables, the location of clusters indicate that we can form groups for some level of lifestyle habits in some range of coordinates (PC1, PC2).

Also in **Figure 5** is the original 2-D PCA Representation with Obesity Level. Although there is some visual matching between Obesity_Type_III and clusters 2 and 3, it is hard to confirm as there is too much noise within the Obesity Level plot. However, notwithstanding the noise, first recall that

1. PC1 represents lifestyle habits corresponding to high number of meals, water consumption, and physical activity, and
2. PC2 represents secondary lifestyle habits corresponding to low vegetable intake and high tech usage time.

For example, cluster 2, which is located at low PC1, low PC2, would indicate populations who typically display low number of meals, low water consumption, and low physical activity, which are paired with high vegetable intake, low tech usage time. For cluster 3, this becomes harder to analyze as the cluster ranges across the fourth quadrant. However, it is arguable that from the horizontally stretched shape of clusters 2 and 3, their certain lifestyle habits correspond to obesity levels like Obesity_Type_III.

While it may appear disappointing that we did not find any significant grouping from PCA representation for obesity levels, this is to be expected. There are so many predictors of obesity levels, including other lifestyle habits, environmental influences, and genetic factors, that make it difficult to form clustering given the five variables we examined.

GMM Results:

To complement k-means clustering we also applied a Gaussian Mixture Model to the PCA transformed data. This is because GMM provides us soft assignments which allow resulting clusters to overlap and adopt non spherical shapes whereas k-means does not. Following the procedural steps assigned to GMM, we first utilized the Mclust package to determine the best value of G. To achieve this Mclust evaluates a range of models defined by the number of Gaussian components, G, and the different covariance structures which it then computes and selects the model with the highest Bayesian Information Criterion (BIC). BIC is a metric that primarily balances two priorities: goodness of fit and model complexity.

$$BIC = \ -2\,ln(L) \ + k * ln(n)$$

* The likelihood (L) estimates how well our model fits the dataset, (n) represents the number of data points, and (k) represents the number of parameters used in the model. (Geeks for Geeks 2025)

Once we had the optimal G value, 6, we fitted the GMM and generated the associated classification, density, and uncertainty plots which are **Figures 8, 6, and 7** respectively.

First pointing to **Figure 6**: Density Plot, we can clearly see a higher concentration of points in the low-PC1, high-negative PC2 regions marked by a concentration of contour lines. Secondly looking at **Figure 7**: The Uncertainty Plot, we reference a paper from Carnegie Mellon, which tells us that larger filled symbols are points assigned to their respective Gaussian with lower uncertainty whereas smaller, lighter symbols are points with higher uncertainty (Fraley, Faftery, and Scrucca 2012). Furthermore, the close proximity of the Gaussians indicate that there is not a hard separation in the groups, which is also backed by the overlapping of points across Gaussians; however, we see that for certain Gaussians in the third quadrant there exists points that are assigned close to their centers, which indicate that there is low uncertainty in regard to their assignment. Finally for **Figure 8**: The Classification Plot, we can observe that certain Gaussians are more compact than their peers, which display a tighter clustering of groups in regards to the characteristics of PC1(lifestyle habits), and PC2(vegetable intake / tech usage) we discussed previously. This observation is particularly evident in low to negative PC1 regions. In contrast, elongations of Gaussians (e.g. red) indicate a loose correlation between the two PCs.

Although the overlapping of Gaussians is a trait of GMM, it makes it difficult to make definite and solid claims about how certain lifestyle habits can be grouped for obesity levels; however, this is not to negate that for Gaussians with compact, low-uncertainity assignments of points, we can still identify certain groups of lifestyle habits that correspond to obesity levels.

For example, we see that the elongated shape of the yellow and red Gaussian in **Figure 8**, correspond fairly closely with Obesity_Type_II in **Figure 3**. This shows that the variability in the habits of PC1(number of meals, water consumption, and physical activity) loosely correspond to the obesity level. On a similar note, we can also claim that the high density of blue points corresponding to the blue Gaussian in **Figure 8**, corresponds to Obesity_Type_III in **Figure 3.**

Silhouette Score Evaluation:

Applying the procedural steps for the Silhouette Plot, we came up with the results in **Figure 9** and **Figure 10**.

First looking at **Figure 9**, which is the silhouette plot for k-means, we see that the mean silhouette score is 0.5009, which indicates that there is no truly "bad" clustering that is taking place for k-means. This indicates that the lifestyle data that we clustered are able to form meaningful, and not random or wrong, groups. While this does not suggest anything about clusters corresponding to obesity levels, the fact that our clustering is forming meaningful groups is encouraging. Furthermore, the absence of especially narrow colors indicate that there is an acceptable level of separation going on between each plot. This is also encouraging to see as it means there is not some single, dominant cluster that is taking every point in the PC space.

Now looking at **Figure 10**, which is the silhouette plot for GMM, the mean score is 0.4341, which indicates an acceptable level of clustering. However, this is a poorer showing than k-means, as we can see in the plot that the results for Gaussians 3 and 6 are poor as indicated by their low Silhouette score (0.3792 and 0.2376, respectively). This implies that there is a lot of unclear clustering with considerable overlap occurring for the points in these two Gaussians with other Gaussians. However, the relatively stronger showing from the other Gaussians, as well as the lack of any negative scores, indicates that there is some meaningful clustering that is taking place. This observation is mirrored in the plot, which shows that Gaussians 3 and 6 are very narrow and short (thus have poor clustering) while the others are wide (except 5) and tall, with better clustering.

Discussion:

Before we proceed to discuss our results, to highlight the relevancy of our chosen unsupervised machine learning clustering models, that is k-means and GMM, we will briefly discuss why we decided against using other models.

Unlike unsupervised ones, supervised methods force the data to match the given labels, even when the underlying behavioral structure does not support them. Unsupervised methods, on the other hand, allow the data to “speak for itself.” By using clustering, we were able to investigate whether patterns in vegetable consumption, meal frequency, water intake, physical activity, and screen time, our five core behavioral variables, give rise to intrinsic groupings and

whether those groupings bear any resemblance to the obesity categories provided in the dataset. In short, unsupervised learning was not simply an alternative method; it was the only framework capable of answering our research question about the behavior patterns in the dataset. For instance, if we used a supervised machine learning approach like a support vector machine, which takes in both the reduced principal components from our analysis(PC1 and PC2) and the obesity levels, and yields estimation for parameters with the smallest error of classification (Marsland 2020). This fundamentally assumes that the obesity levels are true for the dataset, which for our research question of identifying groups in lifestyle behavior that correspond to obesity levels is a circular reasoning. That is, it would force classification for lifestyle habits by obesity levels, instead of us being able to see if lifestyle habits will naturally correspond to obesity levels.

The reason we consider both k-means and GMM is that both make completely different assumptions about the data. As evident in **Figure 5**, the clusters that k-means generated will be naturally concentric about its cluster center as it assigns points to clusters based on how close it is to a cluster versus another. On the other hand, in our context, GMM uses the G-number of 2-D Gaussian to assign points to, which means that it is better for identifying clusters that may be elongated or correlated along an axis. Thus in the context of our research question, k-means would help us uncover clusters that are hopefully grouped circularly around some location on the PC1-PC2 axes, although it would have neglected the possibility that certain lifestyle habits can be better groups if done with an elongated shape, say a Gaussian distribution. GMM, then, would seem more practical due to it being able to be both circular and concentric depending on the dataset, the fact that points can be shared among different Gaussians, the responsibility that each Gaussian then takes for each point must be considered. Furthermore, in cases whereby the Gaussians do a poor clustering of the data (e.g., say, a Gaussian within a Gaussian, as is the case in **Figure 8** with Gaussians 3 and 6) or have very low density, it becomes hard to claim whether there are ample amount of points clustered together to enable us to claim that they correspond to an obesity level. Thus by looking at them together, we are able to fill the gaps of k-means and GMM by supplementing each other.

Now onto a discussion of the results, we first reiterate that k-means is a hard assignment of points, whereby a point is assigned to a cluster, and one cluster only. Conversely, GMM is a soft assignment of points, whereby a point can be assigned to multiple Gaussians, with each assignment carrying some weight in the form of responsibility to each Gaussian to which a point is assigned. The reason why we decided to attempt both hard and soft clustering is because we hypothesized that if both clustering methods show very clear correspondence to the obesity levels, it may validate our connection of lifestyle habits and obesity levels even more. By bolstering how “good” our clustering is through the silhouette score and the plots, we were able to examine if each of the clusters from k-means of GMM were correct, and not random or wrong, clustering.

From our results, it is clear that our hypothesis for the hard and soft clustering did not yield similar-looking clustering results. As a way of discussing this, we turn to **Figures 5** and **8**. Whereas the k-means clustering shows a centralized cluster around the origin in PC1-PC2 axis with a concentric arrangement of five other clusters around the central cluster, GMM yields "stacks" of Gaussians with an extremely dense gathering of points, as evident in **Figure 7**, for Gaussians 3, 2, and 1.

Now considering the higher mean silhouette score (0.5009) for k-means compared to GMM (0.4341), k-means yielded better separation than GMM, which shows that we could put a little more weight on the conclusion from k-means. Further, as a point of nuance, clusters with larger silhouette scores should also be considered with more weight than those with lesser scores. With this established, looking at **Figure 5,** our previous observation on cluster 3 corresponding to Obesity_Type_III can be looked at with more emphasis, as individually it has a high silhouette (score = 0.5331). For cluster 4, we also see that there is a high correspondence to Obesity_Type_II. The lower silhouette score for cluster 4 does lower our confidence for this specific observation, but it nonetheless exhibits the correspondence behavior between lifestyle habits and obesity levels that we were searching for. Although we wish we could claim similar for other clusters, such as cluster 6 matching to some parts of Obesity_Type_III , the noise (i.e., presence of other obesity colors) in the PCA Representation with Obesity Level across the PC space makes it difficult to make any definitive claims.

For our observations on GMM, the lower silhouette score, again, reduces our confidence in the below discussion. Notwithstanding, looking at **Figures 3**, **8** and **10**, we still recognize that there is some weak correspondence between Gaussian 2, which is noticeably more elongated than others) that share a similar spot with Obesity_Type_III. Furthermore, the high-density Gaussian 3 also shares this spot with a high density of Obesity_Type_III, although this should be taken with caution due to Gaussian 3 having a noticeably lower silhouette score of 0.3792.

Again, we want to be clear that due to significant noise in **Figure 3**, it is difficult to draw any definitive conclusions about how much a cluster in either k-means or GMM corresponds to an obesity level, which is what we initially sought out to do in our research question. However, the fact that we see these weak links between obesity levels and certain clusters hint at the fact that perhaps with improvement in the number of variables or our method of quantifying this correspondence between clusters and obesity levels, we can draw more meaningful conclusions. Not to get ahead too much, we will now turn to a discussion of limitations and future work.

Limitations and Future Work:

Through the combination of PCA, GMM, and k-means we observe some lifestyle patterns within our dataset; however, there still lay many limitations within our work primarily in 2 areas: (1) Limitations of the dataset, and (2) Limitations of methods. Beginning with the first, we recognize that our dataset includes 2,111 individuals primarily from three countries- Peru, Mexico, and Columbia. Conclusions drawn from this dataset are not strong enough to generalize to the rest of the world as individual lifestyle habits vary largely across different cultures and continents. Furthermore many variables within our dataset, Frequency of Vegetable Consumption (FCVC), Number of Meals (NCP), Water Consumption (CH20), Physical Activity Frequency (FAF), and Screen Time Usage (TUE) are all self reported variables which introduces bias into our work. For example, recall bias may occur when participants of the study misreport / misremember past events and experiences. By creating a system in which people will strictly stick to a certain diet or lifestyle habits for a long period of time, we can better model how these lifestyle choices contribute to an individual being grouped to a certain obesity level. Another aspect to note is that the distribution of obesity levels in our dataset is imbalanced where normal weight and overweight individuals make up a larger share of the sample comparatively to severe obesity levels (Obesity_Type_I, II, III). Lastly, for the dataset a difficult aspect of the dataset directly hindering our ability to analyze our visualizations was simply just the amount of noise present. With 2,111 datapoints overlapping on a single graph it is easy to draw conclusions as it can just look like one large clump. Thus, if we had a technique in which we were able to visualize our graphs layer by layer we would be able better determine any patterns or clear groupings that arise.

This brings us into (2), where we will discuss the scope of PCA, k-means, GMM. Looking at **Figure 3**, what was most remarkable was how Obesity_Type_III had such a distinct, elongated clustering that stretched from the third to fourth quadrant. This makes us wonder whether this is an indicator of a lifestyle pattern underlying certain obesity levels or if it was solely due to pure chance. As such, in the future we need to add more variables (dimensions) to do PCA on, as it would introduce more nuance into the numerous factors that group with certain obesity levels. This could be done by introducing a method (that we are not aware of at the moment) that could somehow convert our categorical variables into some continuous (numerical) variables that could be used for PCA. On the subject of elongated clustering, k-means inherently assumes that all clusters are round and equally shaped, therefore the existence of our so-called “elongated Gaussians” is a violation of the assumption. Furthermore, k-means is a hard assignment model meaning that it assigns one point to one cluster which in the context of human behavioral data is not realistic. Although GMM (a soft assignment model) does answer some of these shortcomings , allowing elliptical clustering, it has its own battles to face as well. While in the context of our work overlapping boundaries is intuitive, strong overlapping clusters are

difficult to interpret which is supported by **Figure 7**: Uncertainty plot. High overlap leads to high uncertainty scores making clusters ambiguous to analyze. Furthermore, GMM is at risk of overfitting and if too much noise is present the model may fit to noise rather than to actual structure.

These limitations imply that with the consideration of more variables (e.g., other lifestyle habits, environmental influences, genetic factors), we can make more insightful conclusions. Furthermore, if we can incorporate a way to mathematically quantify how much correspondence is ongoing between obesity levels and the clusters / Gaussians, we can draw more objective, less visual-based conclusions.

Conclusion

Thus in our work to respond to our research question, "Can unsupervised machine learning methods be used to identify groups in lifestyle behavior that correspond to different levels of obesity?", we recall the following steps.

First, we dimension-reduced five variables, FCVC (frequency of vegetable consumption), NCP (number of main meals per day), CH2O (daily water consumption), FAF (physical activity frequency), and TUE (daily time spent using technological devices) into 2-D using PCA, which revealed the two highest directions of variability in the dataset with respect to the five variables. We found that PC1 was the most defined by NCP, CH20, and FAF and PC2 with FCVC and TUE.

Second, we then used two clustering methods, k-means clustering and Gaussian Mixture Model, in which we found the most optimal parameter for the former through the elbow plot and the latter with the built-in functionality within the function Mclust() that uses BIC.

Third, we evaluated the quality of the clustering by utilizing the mean silhouette score and silhouette plot, which revealed that both k-means and GMM had an acceptable, and not random nor incorrect, clustering. This, to a great extent, validates the use of clustering for our dataset.

Fourth, we visually checked to see if there were clear correspondences between certain clusters or Gaussians with obesity levels. This is mainly possible as both are in PC space, and we reasoned that groups of lifestyle habits corresponding clearly to obesity levels in PC space should indicate that certain lifestyle habits do indeed identify with certain obesity levels. Even though we observed some correspondences between certain clusters and Gaussians with some

obesity levels, it is still difficult to make definitive claims due to significant noise being present in the PCA representation with obesity levels.

Although it may seem disappointing that our conclusion was not as strong as we would have liked it to be, it is still valuable in that it allows for the possibility that by addressing our ideas in the aforementioned future works section, we can greatly improve the relevance of our work. Overall, the possibilities of improvement for this project affirms the idea that obesity is indeed a nuanced, complex topic that deems more study. It also opens new doors to explore in the realm of unsupervised machine learning, and perhaps with a deeper understanding of other existing techniques we can draw more meaningful and firmer conclusions in the future.

Bibliography

1. *Banos, Oscar, David Giménez, and Andrés Hervás-Martínez. 2015. "Estimation of Obesity Levels Based on Eating Habits and Physical Condition." UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.*
2. *Datanovia. 2024. "K-Means Clustering in R: Algorithm and Practical Examples." Datanovia. https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/.*
3. *Fraley, Chris, Adrian E. Raftery, and Luca Scrucca. 2012. "MCLUST: Software for Model-based Clustering, Classification, and Density Estimation for R." Journal of Statistical Software 61 (6): 1–36. https://www.stat.cmu.edu/~rnugent/PCMI2016/papers/fraleymclust.pdf.*
4. *Marsland, Stephen. 2020. Machine Learning: An Applied Mathematics Introduction. Cham, Switzerland: Springer. https://mml-book.github.io/book/mml-book_printed.pdf.*
5. *World Health Organization. 2023. "Obesity and Overweight." WHO. June 9, 2023. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.*
6. *GeeksforGeeks. (2025, July 23). Bayesian information criterion (BIC). https://www.geeksforgeeks.org/machine-learning/bayesian-information-criterion-bic/*

Appendix A: Figures

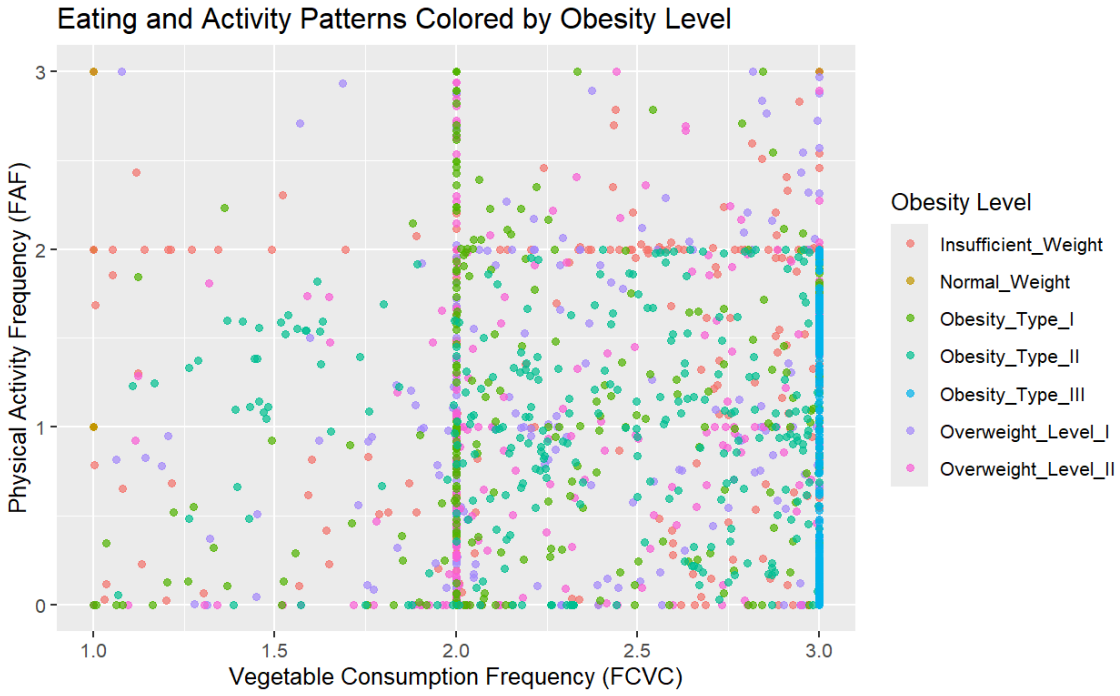

Figure 1: Eating and Activity Patterns Colored by Obesity Level

Figure 2: Alcohol and Smoking Behavior Across Obesity Levels

Figure 3: Dataset projected onto PC1 and PC2. Each point represents an individual and the colors indicate their obesity levels.

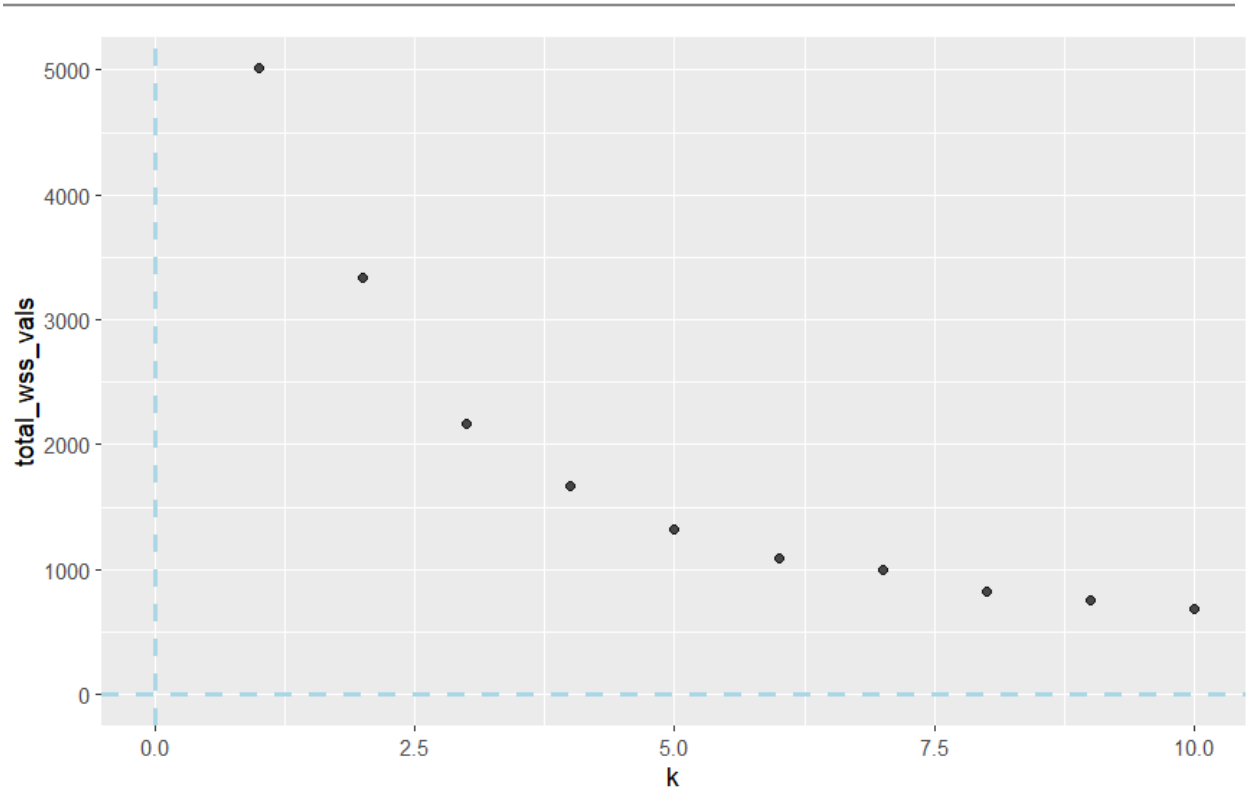

Figure 4: Elbow Plot used to find Optimal value of k

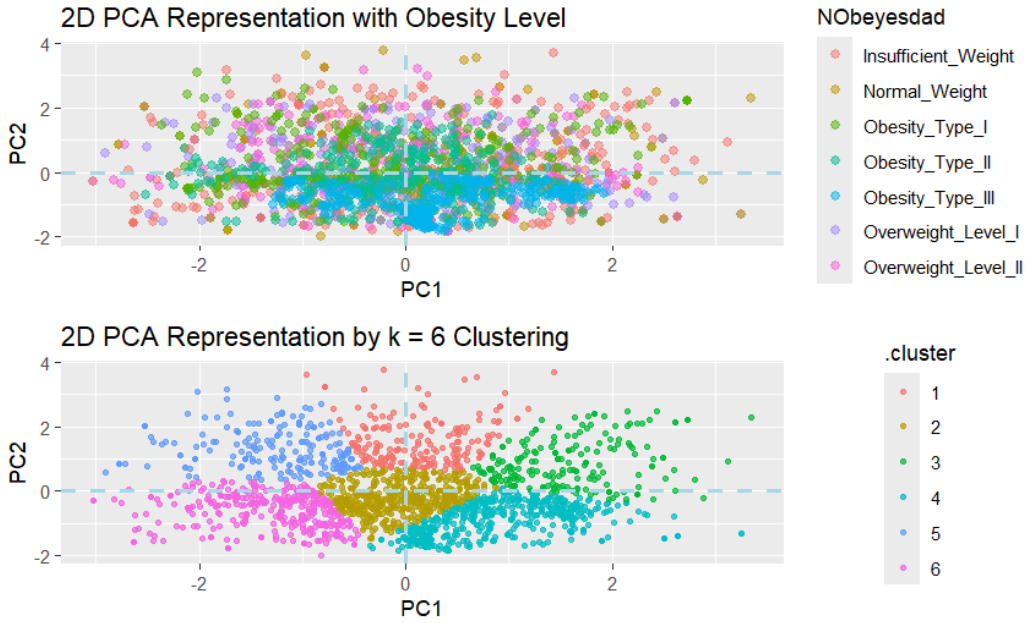

Figure 5: PCA Visualization of the obesity dataset. The top panel is obesity level mapped onto the PCA. The bottom panel shows the same PCA space grouped by k-means cluster with k= 6.

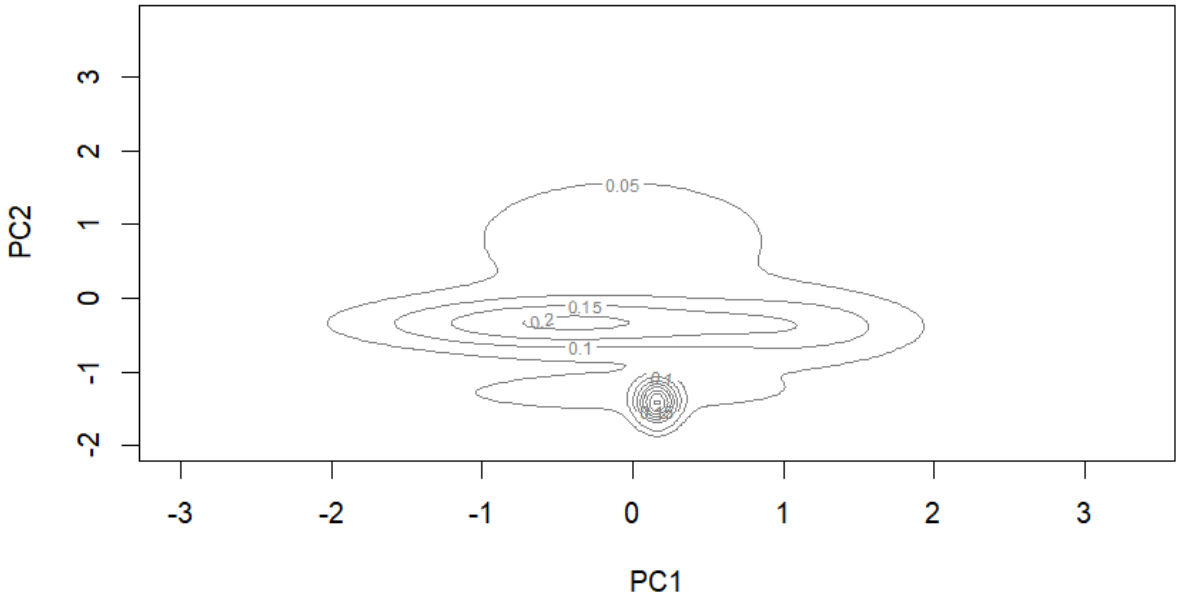

Figure 6: Density Plot of PC1 and PC2 showing areas of high concentration marked by contour lines.

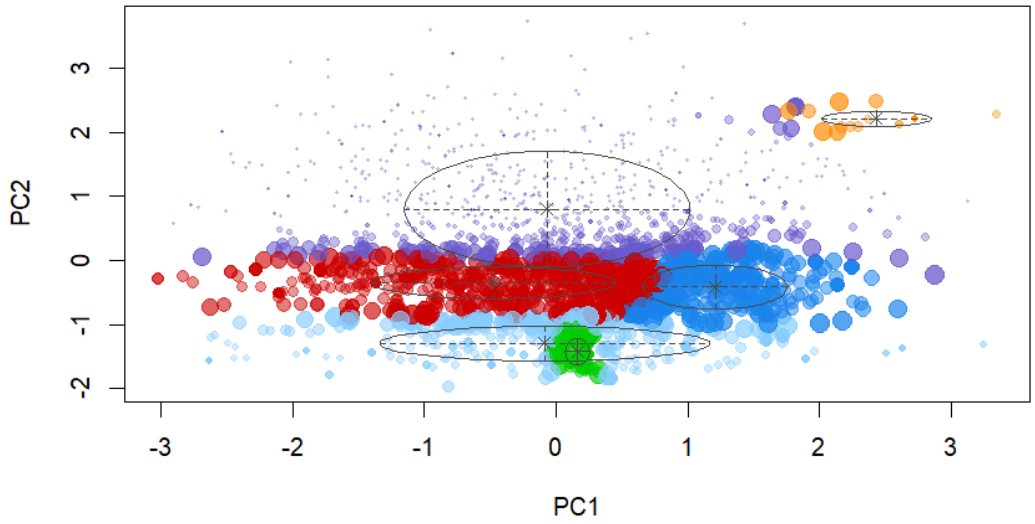

Figure 7: Uncertainty Plot of PC1 and PC2. Larger, darker dots indicate observations assigned to their Gaussian component with less uncertainty whereas smaller, lighter dots indicate higher uncertainty

---

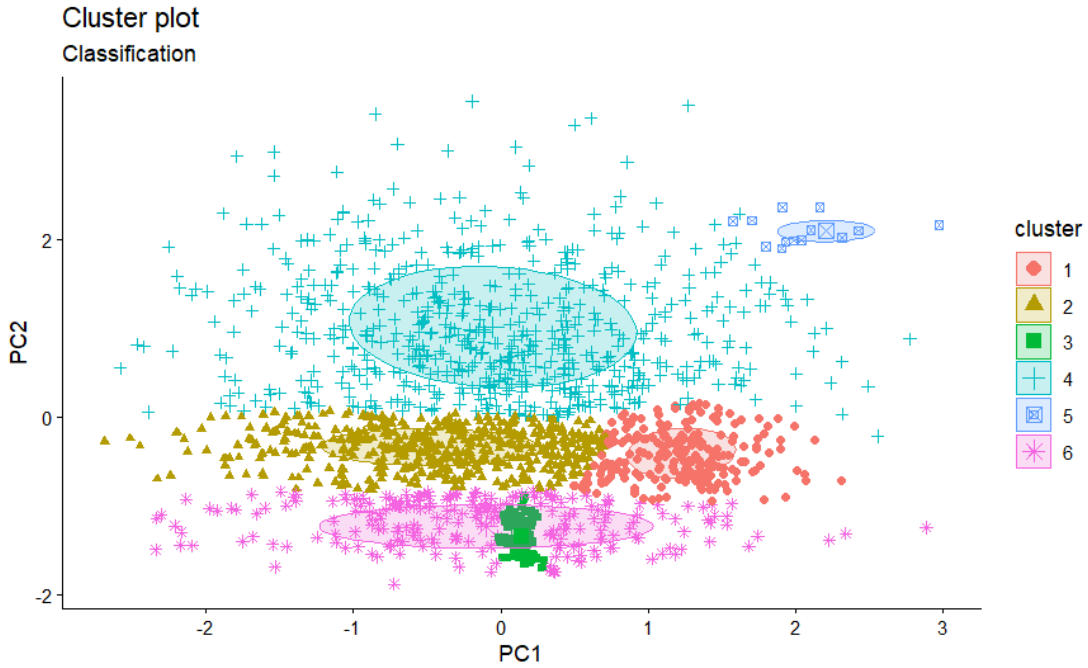

Figure 8: Classification Plot for PC1 and PC2. More compact Gaussians indicate a tighter clustering of similar characteristics. Elongated Gaussians suggest a looser correlation between the two PCs.

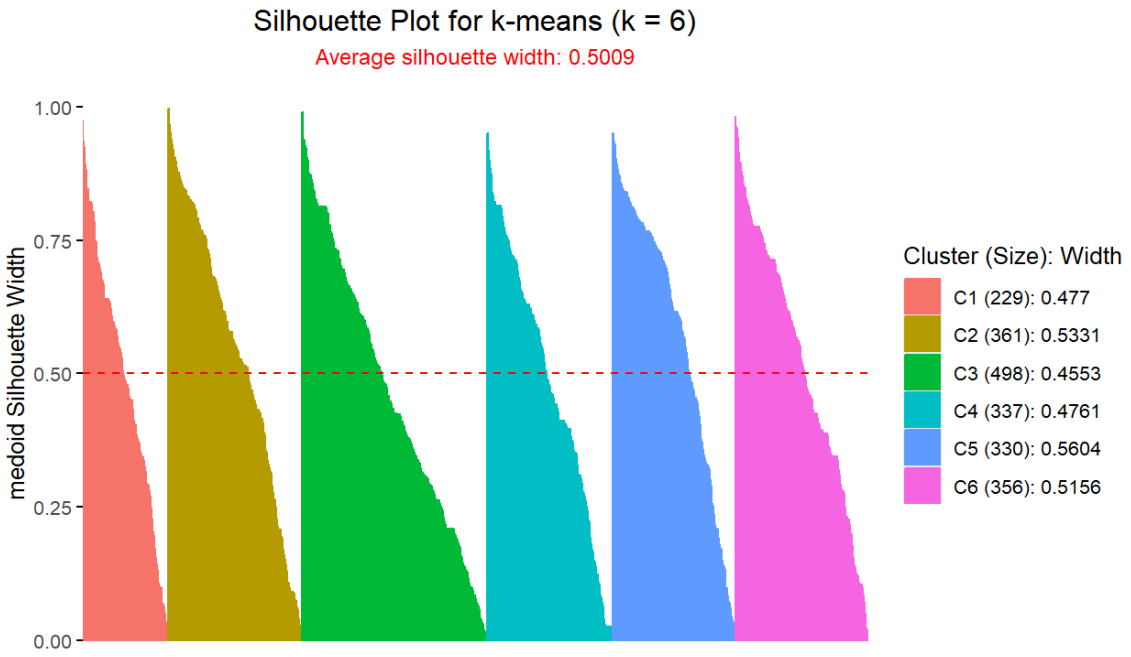

Figure 9: Silhouette Plot for k-means for k = 6

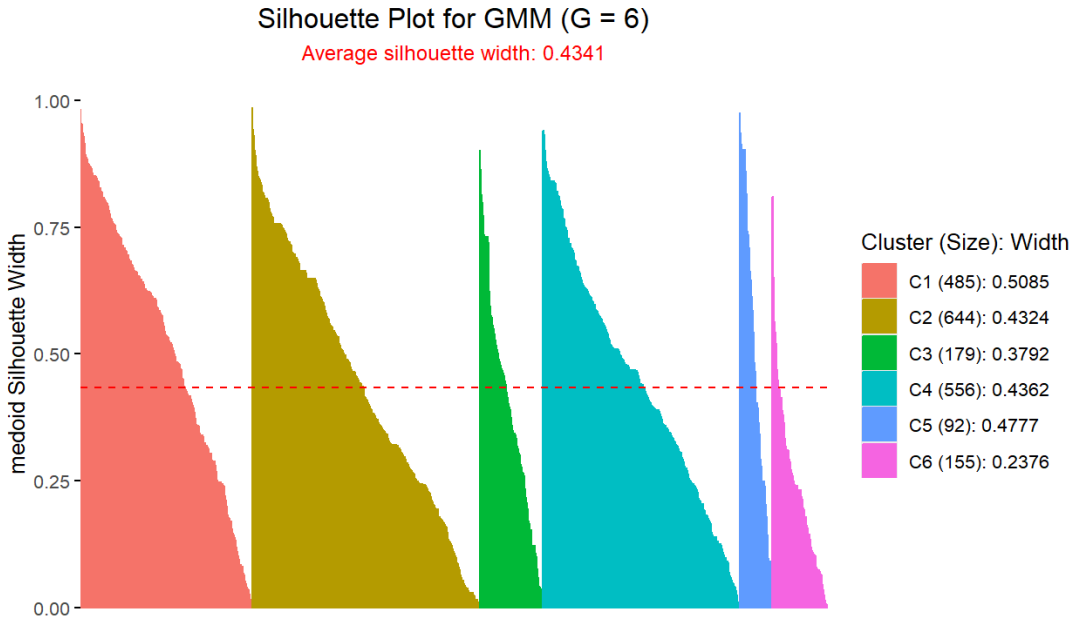

Figure 10: Silhouette Plot for GMM for G = 6

Appendix B: Tables

	PC1	PC2
Frequency of Vegetable Consumption	0.2090413	-0.68388403
Number of Meals	0.4777915	0.05658398
Water Intake	0.5568172	-0.12514939
Physical Activity Frequency	0.6291376	0.12822453
Technology Use Time	0.1488540	0.70498013

Table 1: PCA Loadings for the Five Variables Analyzed