author12hash=789a9375127043a3d28c4daf73828381family=Alcazar, familyi=A., given=Julian, giveni=J.hash=3d16071133096dd58a1849bcc7cf6db2family=Navarrete-Villanueva, familyi=N., given=David, giveni=D.hash=1b1d2c4b202608a3b2e150346462c350family=Mañas, familyi=M., given=Asier, giveni=A.hash=71cbe8765d66b80a08af735ccdf8fdb3family=Gómez-Cabello, familyi=G., given=Alba, giveni=A.hash=8d0a8f34968579292a81ba85011c027dfamily=Pedrer Chamizo, familyi=P., given=Raquel, giveni=R.hash=dbb4f86d589cfe910ad6df581173a2effamily=Alegr familyi=A., given=Luis M., giveni=L. M.hash=65855fae92440fed151644e0cd9d96befamily=Villa, familyi=V., given=Gerardo, giveni=G.hash=77503dc081925407ad51ae46f8058559family=Gusi, familyi=G., given=Narcís, giveni=N.hash=4346936ae9d5812912ea48128c59fb69family=González-Gross, familyi=G., given=Marcela, giveni=M.hash=c63829a1ea60880d39731e9640c307b2family=Casa familyi=C., given=Jose Antonio, giveni=J. A.hash=f26801b94c1a907db802bce9cf2e6e36family=Vicent Rodriguez, familyi=V., given=German, giveni=G.hash=f33d45088e248b0393cba2d5367baafcfamily=A familyi=A., given=Ignacio, giveni=I. author2hash=677774606dcde5657361415e2b2a28 familyi=B., given=Sanghamitra, giveni=S.hash=9f4e9584bc598469e23ea48a96dd8ab2family=Saha, familyi=S., given=Sriparna, giveni=S. author6hash=faa0cb812dece9c32989e35eec3277dffa familyi=B., given=Oscar, giveni=O.hash=2b6d8a0291cd8e4c453c6f6c73c11ac6family=Veen, familyi=V., given=Jort, giveni=J.hash=60ea5da347803d3f26d1eae5ec3e02eefamily=Montiel-Rojas, familyi=M., given=Diego, giveni=D.hash=09fb12aca6cd84a46d778a4996d7cc58family=Edholm familyi=E., given=Peter, giveni=P.hash=0500e35dc48cc34db5006b5c4e0d4fe2family=Kadi, familyi=K., given=Fawzi, giveni=F.hash=bed297f62f74dbf0ee97a25e60ba1063family=Nilsson, familyi=N., given=Andreas, giveni=A. author1hash=bba68e77287d1c29690bbb7fdaafdc familyi=B., given=Matthias, giveni=M. author6hash=e61efa035511cb8e1804a8360d1249 familyi=D., given=Pauline, giveni=P.hash=c4b2c3bc62758eb4660d3481280d9d3dfamily=Méjean, familyi=M., given=Caroline, giveni=C.hash=566178444ff6c5af69ea20a9e832b9cafamily=Bellisle, familyi=B., given=France, giveni=F.hash=2448113206b73428db7120d11615f94bfamily=Allès, familyi=A., given=Benjamin, giveni=B.hash=2621eedc6da98c62818c5aac27a00f53family=Hercberg, familyi=H., given=Serge, giveni=S.hash=ca2c4b6a5971ab722eb5e9218f454c8afamily=Péneau, familyi=P., given=Sandrine, giveni=S. publisher1UCI Machine Learning Repository author1hash=fb48b65407d03cee0c164d17c1ea7eacfamily=Geron, familyi=G., given=Aurelien, giveni=A. author2hash=bbae5355dcc684680961ec621eb053c1fam

familyi=G., given=Mauro, giveni=M.hash=81dd4c4d2b8a3e0c9299333789d20eecfamily=Shung, familyi=S., given=Dennis L., giveni=D. L. author5hash=babd1cb6f580135b9dd8bd4313fd4b73family=Hebert, familyi=H., given=James R, giveni=J. R.hash=bff3685eb5dae204eee9f872d232b1e3family=Clemow, familyi=C., given=Lynn, giveni=L.hash=c9423883e5b9886877e5af6700b29d8bfamily=PBERT, familyi=P., given=LORI, giveni=$.hash = e95a20220e2f29705cfc95c2ddca55cafamily = OCKENE, fan$ familyi=H., given=Rink, giveni=R.hash=5867adc526123a9efec91959bb94f6e5family=Kiers, familyi=K., given=Henk A. L., giveni=H. A. L.hash=a3a1bd4dbfe02a1e551acfc08210c66dfamily=John familyi=J., given=Addie, giveni=A. author3hash=ca063c15e63c4f373b3866da0893d15bfa familyi=K., given=D, giveni=D.hash=1b752f72b335bf2263d40de70dd6ea45family=Cody, familyi=C., given=D, giveni=D.hash=6b648acdcb8e9087d3c45f4770c00864family=O'Shea, familyi=O., given=D, giveni=D. author4hash=62bf0fe2ed3b9e02268632b433336372famil familyi=N., given=Monica, giveni=M.hash=0606b2ccaf802e9451d1aeeba0dffd3efamily=Lutze, familyi=L., given=Sarah Alice, giveni=S. A.hash=3e783d63bfba54dd9bba5a5946cf1d73family=Grech, familyi=G., given=Amanda, giveni=A.hash=b53a0c9b43f2d7b0f21c99bc1a746ccbfamily=Allman-Farinelli, familyi=A., given=Margaret, giveni=M. author2hash=6dd120842c469b3b0085301a322e2526family=Palechor, familyi=P., given=Fabio Mendoza, giveni=F. M.hash=edcf2af9026185b3f156709cd30a08d3family=Manotas, familyi=M., given=Alexis De La Hoz, giveni=A. D. L. H. author1hash=65eabc6887f06518e1098fff1609e90afamily=Piaggi, familyi=P., given=Paolo, giveni=P. author5hash=cce05cab2c437e9c128f853684ee3137family=Wang, familyi=W., given=Liang, giveni=L.hash=90e24063be3dab0369505809e437b908family=Wang, familyi=W., given=Huijun, giveni=H.hash=44d73e639c535123777db68e270b754efamily=Zhang, familyi=Z., given=Bing, giveni=B.hash=cf3a94e4104e1a9d2704cf1e9e96875dfamily=Popkin, familyi=P., given=Barry M., giveni=B. M.hash=dc3a1206c1fa821aa2e6e818f26960b0family=Du, familyi=D., given=Shufa, giveni=S.

University of Strathclyde Glasgow

CS818 Assessment

Factors impacting obesity rates

Word count (from TexCount): 3241

March 28, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Obesity is a critical public health challenge, with links to elevated risk of type 2 diabetes, cardiovascular disease, and certain cancers [**Kinlen2018**]. Epidemiological research has long established that lifestyle factors - most notably diet and physical activity — are closely associated with body mass index (BMI) and general metabolic health [**Bergens2020**]. Public health officials base much of their obesity advice on these factors. However, the strength and nature of these associations can be heterogeneous across different populations. For example physically active and metabolically healthy individuals may nevertheless be obese [**Alcazar2021**], and this combination is more prevalent amongst females and younger age groups [**Bluher2020**]. Similarly, whilst a diet rich in saturated fats has been linked to increased obesity risk [**Wang2020**], countries like France maintain comparatively low levels of obesity with comparatively high saturated fat intake [**Ducrot2018**].

This heterogeneity challenges conventional statistical methods, which often assume homogenous relationships between variables [**Hoekstra2012**]. Researchers need more advanced analytical techniques to disentangle these complex relationships between lifestyle factors and obesity.

We explore the multifactorial nature of obesity with both supervised and unsupervised learning techniques. Our findings highlight key lifestyle factors associated with obesity and their interrelationships, which may enable public health officials to develop more targeted interventions.

Chapter 2

Data

2.1 Overview of dataset

The analysis in this study is based on a dataset from the UC Irvine Machine Learning Repository [**Obesity**], comprising 2111 observations across 17 variables collected from participants in Mexico, Peru and Colombia. Approximately 77% of the data is synthetic, which was used to address an unbalanced distribution in obesity levels [**Palechor2019a**]. The target variable is obesity level, labelled 'NObeyesdad' in the dataset. This categorises individuals based on their BMI into the groups Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. The calculation for BMI is

$$\text{BMI} = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$

and obesity is defined as a BMI equal or greater to 30.00 in line with the World Health Organisation's definition [**Palechor2019a**]. The coded variables that appear in the dataset are detailed in table **??**.

Variable Name	Description
family_history_with_overweight	Family history of overweight
FAVC	Frequent consumption of high-caloric food
FCVC	Frequency of vegetable consumption in meals
NCP	Number of main meals per day
CAEC	Frequency of food consumption between meals
SMOKE	Smoking status
CH2O	Daily water consumption
SCC	Monitoring of daily calorie intake
FAF	Frequency of physical activity
TUE	Daily usage time of technological devices
CALC	Frequency of alcohol consumption
MTRANS	Usual mode of transportation
NObeyesdad	Obesity level

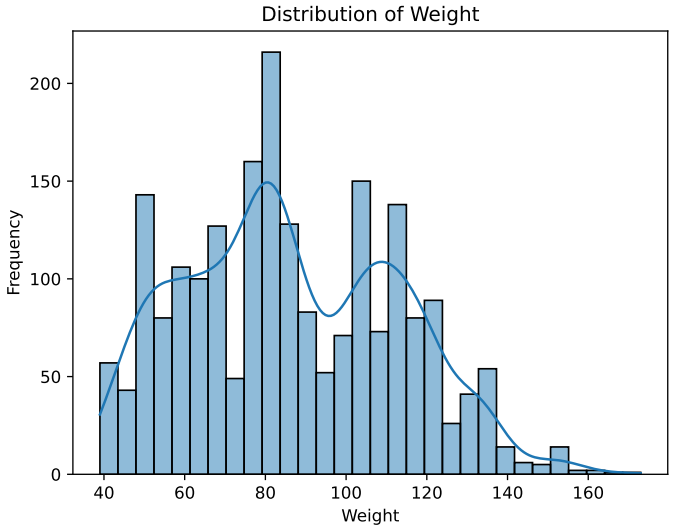
Table 2.1: Variables and Descriptions

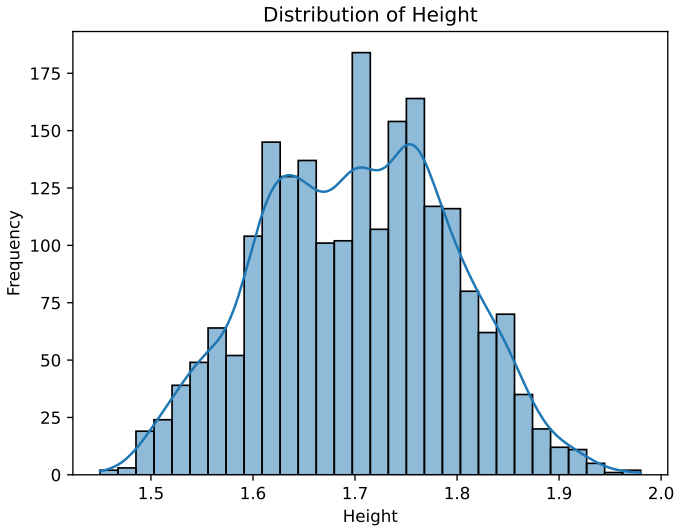
2.2 Exploratory analysis: results

Figure **??** captures the distribution of the demographic variables, which shows a right-skew for age with a mean of 24.3 years and few entries from participants aged 40 and above. This could limit how well any findings generalise to older populations. Weight's distribution is unimodal and slightly right-skewed, with a mean of 86kg, whilst height's distribution is more symmetrical, with a mean of 1.7m.

Expanding out the exploratory analysis, we can consider the correlation between numeric variables. To ensure obesity rate is captured, the category 'BMI' will be added using the formula already given. And since BMI is calculated from height and weight, those categories will be removed to avoid duplication and to focus in on variables beyond body-measurements.

As per figure **??** and table **??**, the strongest correlation is the negative relationship between age and time on electronic devices, though this relationship is not directly applicable to obesity. Of greater relevance are positive relationships between BMI and both vegetable consumption (FCVC) and age. There is also a negative relationship between BMI and physical activity (FAF), though in all cases the strength of the relationship is fairly weak.





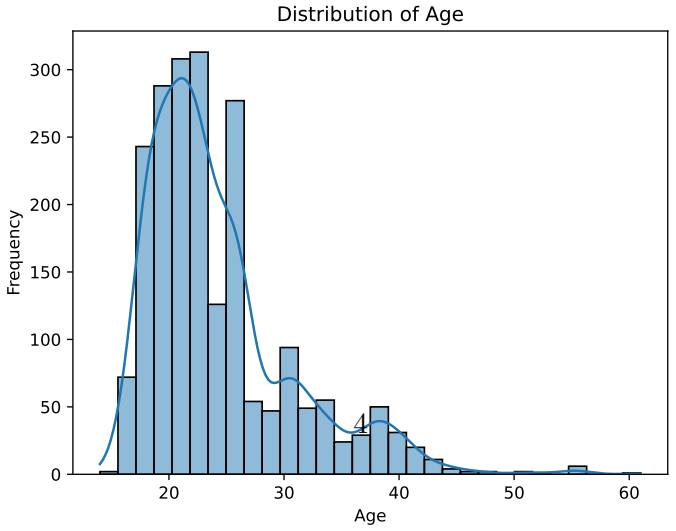


Figure 2.1: Distributions of weight, height, and age.

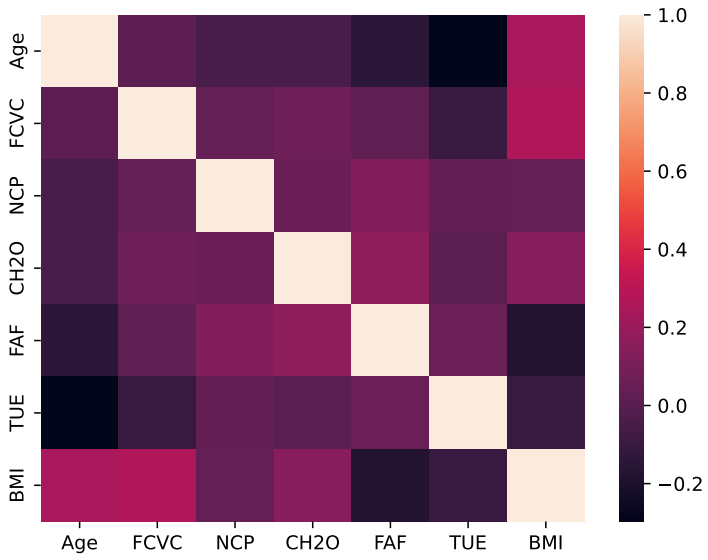


Figure 2.2: Correlation Heatmap

Table 2.2: Strongest Correlation Pairs

Variable 1	Variable 2	Correlation
Age	TUE	-0.297
BMI	FCVC	0.264
Age	BMI	0.244
BMI	FAF	-0.178
CH2O	FAF	0.167
Age	FAF	-0.145
BMI	CH2O	0.144
FAF	NCP	0.130
FCVC	TUE	-0.101
BMI	TUE	-0.100

The dataset also contains the categorical variables CALC (how often a respondent drinks alcohol) and CAEC (how frequently a respondent eats between meals). In both cases, the data is heavily concentrated in the 'Sometimes' category, though there were a reasonable number of respondents who noted they never drink alcohol, as per **??**. Similarly, the MTRANS (mode of transport) variable in figure **??** shows the vast majority of respondents travelling by public transport, with a smaller minority travelling by automobile, and very few travelling by either bike, motorbike, or walking. In their current skewed form, these variables could disproportionately influence model estimates or clustering outcomes.

For the final section of this exploratory analysis, we can consider how the binary variables relate to obesity levels. Here, 'Obese' follows the WHO definition as falling in category Obesity Type I, II and III, whilst 'Not obese' is all remaining categories, representing a BMI of below 30. As shown in figure **??**, whilst an individual's gender or smoking status has little relation to obesity levels, individuals reporting a family history of obesity are significantly more likely to be obese than those with no such family history. Similarly, high consumption of calorific foods (FAVC) is perhaps unsurprisingly associated with higher obesity, and monitoring calories with lower obesity.

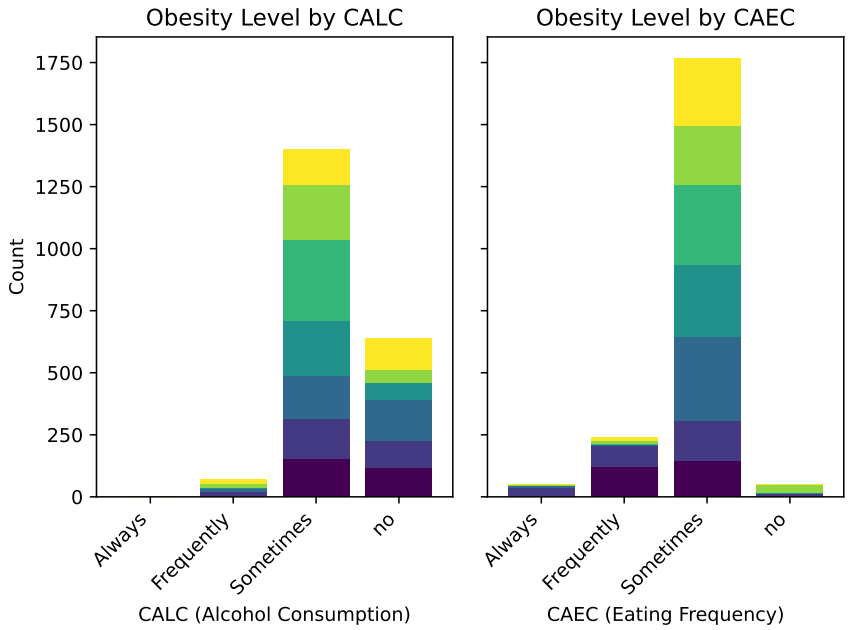


Figure 2.3: Distribution of Obesity Level by CALC and CAEC

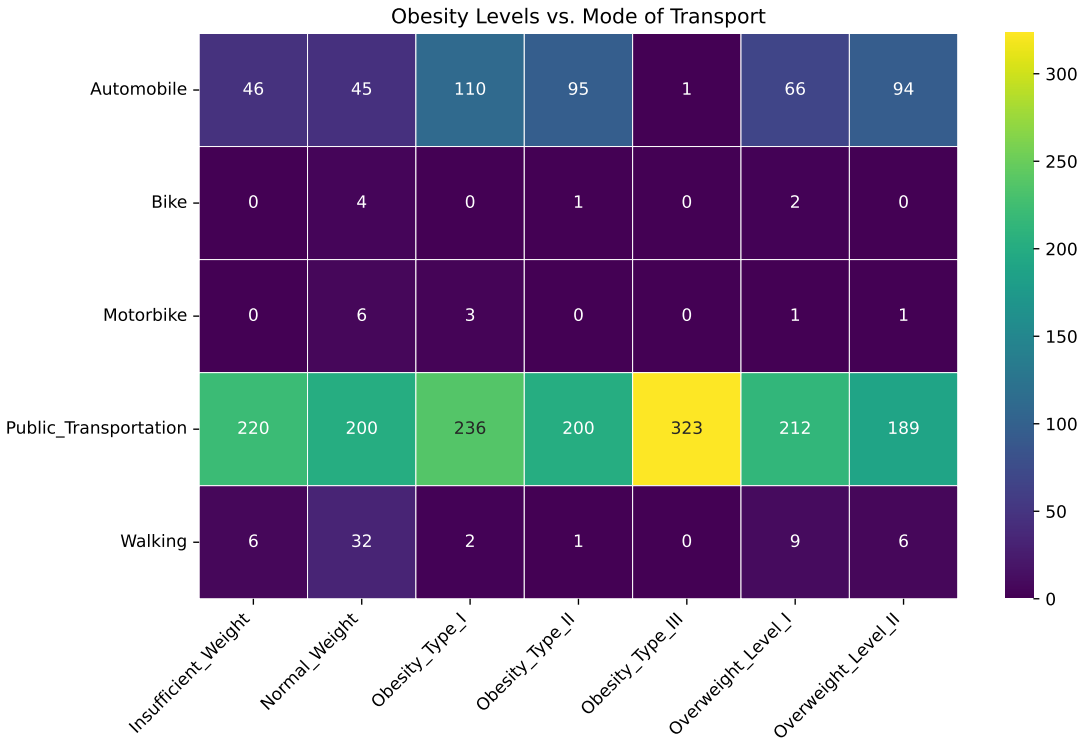


Figure 2.4: Obesity Levels vs. Mode of Transport

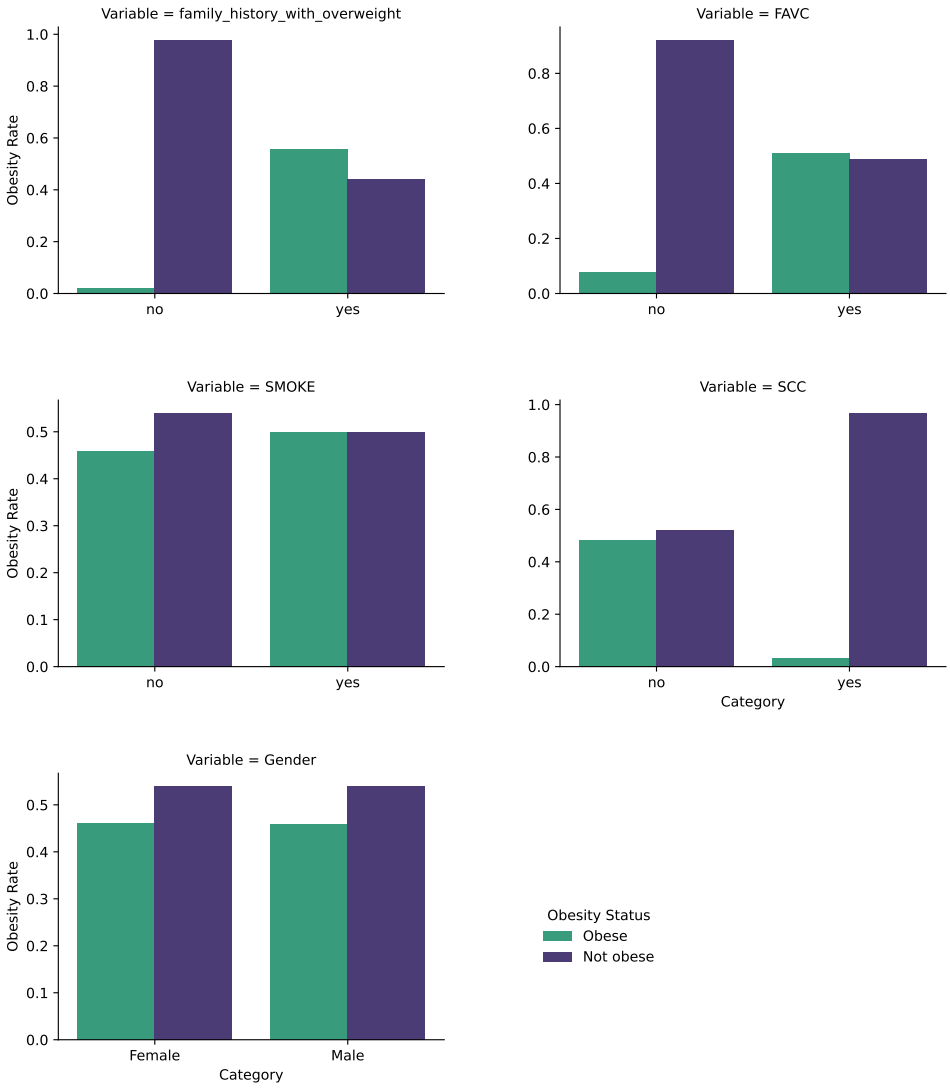


Figure 2.5: Obese vs. Not Obese for Binary Variables

2.3 Discussion

This initial exploratory analysis has provided a clear understanding of the dataset. Bivariate analysis was used to identify correlations. The link between low physical activity and higher BMI is consistent with research showing the positive impact exercise has on obesity risk [**Bergens2020**]. Similarly, individuals may become more sedentary as they grow older, which could contribute to the link between age and BMI. However, the fact that higher vegetable consumption is correlated to higher BMI directly contradicts findings and public health advice which encourages greater vegetable consumption as a means of reducing obesity risk [**Nour2018**]. This will be considered when conducting further analysis in the report's remaining sections. Similarly, analysis highlighted the concentration of responses to certain categorical variables such as alcohol consumption, which could indicate a predisposition amongst respondents towards particular behavioural patterns which will be borne in mind. Finally, the dataset is primarily made up of synthetic data. Whilst the SMOTE technique used by the creators is methodologically robust, there is always a risk with synthetic data that it fails to capture real-world variability [**Giuffre2023**]. These potential limitations will be considered when evaluating the generalisability of any findings.

Whilst this approach has allowed us to uncover trends, univariate and bivariate analysis does not adequately capture the interactions between variables. For example, the correlation matrix shows a strong correlation between age and time on electronic devices, and separately between age and BMI, suggesting but leaving unexplored an interplay between these factors. To better understand potential interrelationships, more advanced multivariate techniques must be employed, which is the subject of the following section.

Chapter 3

Unsupervised Analysis

Unsupervised learning is a machine learning technique which seeks to identify patterns and insights in unlabelled data. One common approach to unsupervised learning is cluster analysis, whereby an algorithm sorts unlabelled data into clusters which share some common associations. K-Means clustering employs an algorithm that partitions data into a predetermined number of clusters (defined as K) by iteratively assigning data to a randomly chosen centre known as a centroid. The centroid is then recalculated based on the mean of the data points assigned to it, and the process is repeated until the centroids have stabilised and subsequent recalculations produce similar results [**Geron2022**].

K-Means can evaluate the behaviour of multiple variables simultaneously, which may permit identification of latent structures or groupings within datasets. As such, it can be used here to uncover distinct profiles of obesity risk that incorporate multiple variables, thereby moving us beyond the uni- and bivariate analyses conducted so far.

3.1 Data preprocessing

Because K-Means analysis works by calculating the distance between data points, it requires numeric variables. Non-numeric variables must therefore be appropriately encoded. The binary variables 'family_history_with_overweight', 'FAVC', 'SMOKE', and 'SCC', together with 'Gender' (a categorical variable with only two unique answers

in the dataset) can be encoded with 1 for 'yes' and 'Male' and 0 for 'no' and 'Female'.

The dataset also contains two ordinal categories, 'CALC' and 'CAEC', with categories denoting frequency in descending order. These can be encoded such that 'no' = 0, 'Sometimes' = 1, 'Frequently' = 2, and 'Always' = 3. The final non-numerical categories are 'MTRANS' and 'NObeyesdad'. As the options for these variables are not ordinal, the most appropriate encoding method is one-hot, which converts each unique category into a separate binary variable [**Geron2022**]. This preserves the nominal nature of the variables, and avoids imposing any artificial order on the category options.

Another important consideration is scaling. With significant variability in the spread across categories, larger scales can end up dominating the clustering process. In this dataset, the scale for weight and to a lesser extent age, have the potential to dominate as captured in **??**. To avoid distortions, a z-score normalisation has been used to centre those variables around zero with a standard deviation of one. This helps ensure each variable contributes equally to the distance calculations.

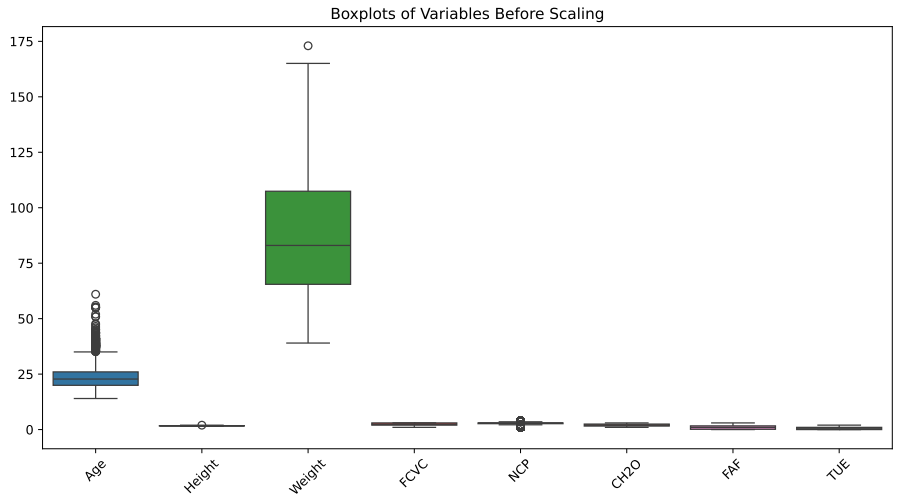


Figure 3.1: Boxplots of the dataset's variables before scaling

Finally, Principal Component Analysis has been used to reduce dimensionality. PCA transforms data with potentially correlated variables into a set of uncorrelated variables called principal components, ordered by the amount of variance they capture

from the original data. This reduces noise and can make it easier for methods like K-Means to identify patterns. It is particularly effective where datasets have a large number of variables and where variables are likely to be highly correlated as with obesity, height and weight in this dataset [**Bandyopadhyay2013**]. A condition has been set for the model to use the number of components necessary to capture 95% of the variance, which ensures the model minimises the loss of potential data patterns.

3.2 K-Means: results

Before the analysis is run, K must be determined. An Elbow plot is used to identify the optimal number of clusters, found at the 'elbow point' where additional clusters no longer significantly reduce inertia. This is used here in conjunction with a Silhouette Score which evaluates cluster cohesion and separation, with higher scores indicating clearer boundaries between clusters [**Geron2022**]. As can be seen in **??**, the Elbow plot suggests a change in inertia around k=4, which is also the highest silhouette score, making 4 the best option.

The K-Means analysis was therefore run with k=4. The output cross-tabulated with the obesity categories is shown in **??**, which shows Cluster 0 is most strongly associated with Overweight Level 1, Cluster 1 with Normal Weight, Cluster 2 with Obesity Type III and Cluster 3 with Overweight Level II. The association between Cluster 2 and the highest level of obesity is particularly strong, indicating that analysis has identified a grouping with significant obesity risk. To better understand what characterises each group, **??** shows the five highest absolute mean values for variables in each cluster, excluding height and weight as the determinants of obesity rates in order to focus in on demographic and lifestyle factors.

Cluster 0 is characterised by few meals but frequent snacking, use of public transport and some prevalence of family obesity. For Cluster 1, frequency of snacking is an even more significant driver, and is coupled with high calorie food consumption. In Cluster 2 the most significant variables are family history and consumption of high calorie food, suggesting both genetic risk and unhealthy eating habits. Frequent snacking also

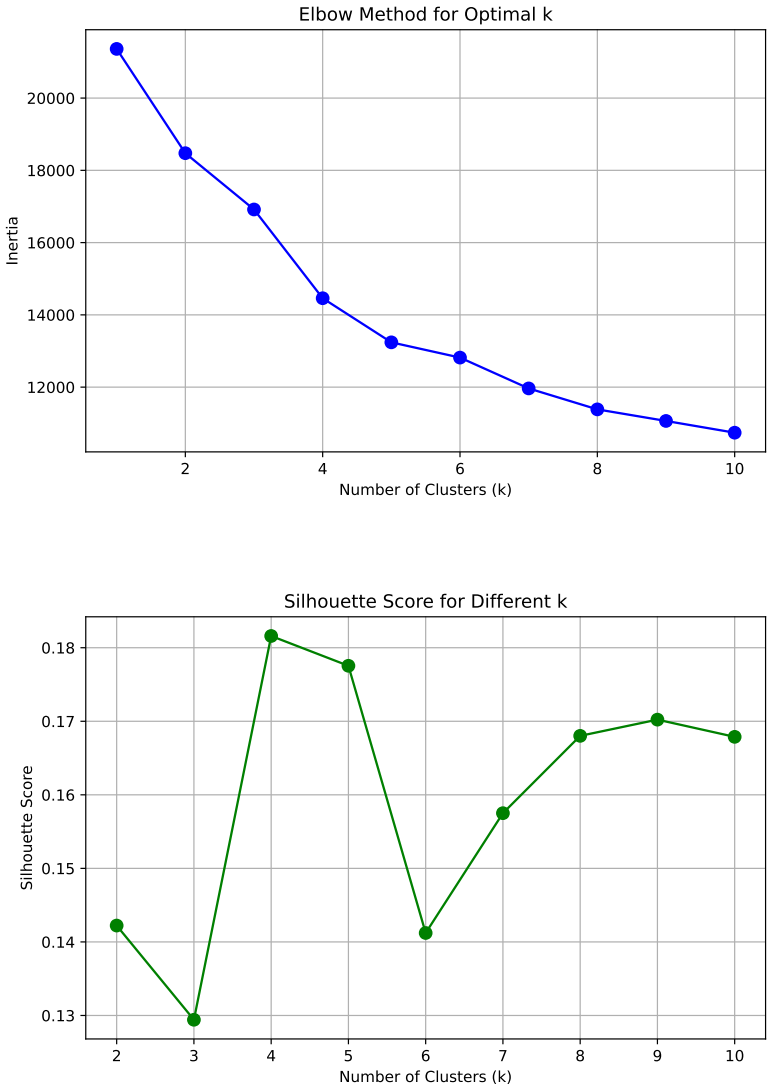


Figure 3.2: Comparison of Elbow Method and Silhouette Score

appears and vegetable consumption is also in the top five most significant variables. Finally, in Cluster 3 age is the dominant factor, though family history remains influential, as do high calorie food consumption and frequent snacking. Assessing clusters as a whole, public transport appears frequently, though as addressed earlier in this study, that may reflect the fact that few respondents use alternative methods of transport.

In the next section, supervised analysis is used to generate further insights from the

data, before the findings from both analyses are considered in the Discussion section.

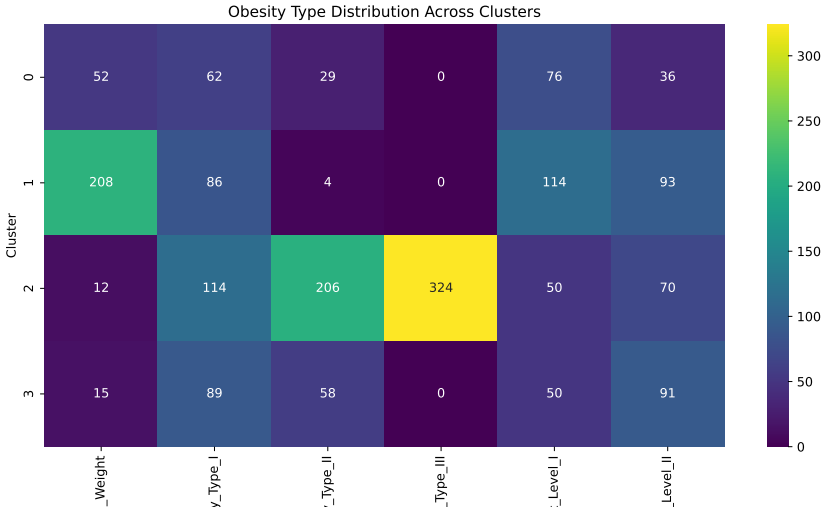


Figure 3.3: Obesity Category Distribution Across Clusters

Table 3.1: Top 5 Strongest Relationships per Cluster (Excluding Height, Weight, and Obesity Categories) with Strength Values

Cluster 0

Rank	Feature	Strength
1	NCP	-1.937
2	CAEC	1.110
3	FAVC	0.851
4	Public Transport	0.929
5	Family History	0.640

Cluster 1

Rank	Feature	Strength
1	CAEC	1.291
2	Public Transport	0.812
3	FAVC	0.793
4	NCP	0.562
5	Family History	0.671

Cluster 2

Rank	Feature	Strength
1	Family History	0.980
2	FAVC	0.969
3	CAEC	1.043
4	Public Transport	0.852
5	FCVC	0.338

Cluster 3

Rank	Feature	Strength
1	Age	1.906
2	Family History	0.925
3	FAVC	0.912
4	CAEC	1.069
5	Public Transport	0.154

Chapter 4

Supervised Analysis

Supervised learning is a machine learning technique that uses labelled data to build predictive models. Decision tree analysis is one such supervised method, whereby an algorithm learns a sequence of decision rules to classify or predict outcomes based on input features. Decision trees work by iteratively partitioning a dataset, using a predictor variable and a criterion such as variance reduction as the basis for determining where to split the data. This process continues until a stopping criterion such as a maximum number of splits is reached [**Geron2022**].

Similar to cluster analysis, decision tree analysis can simultaneously evaluate multiple variables. However, whilst K-Means analysis has been used to uncover latent groupings in the data, a decision tree will instead permit identification of key predictors. This can further refine our understanding of obesity risk's key drivers and their interactions.

4.1 Decision Tree Analysis: results

The analysis was run three times. First, it was completed for the full dataset. As can be seen in **??**, this achieved a strong accuracy score of 0.93, however the most important feature by a significant margin was 'Weight', followed by 'Height'. Given that weight is the primary determinant of BMI, it is dominating the model and overshadowing the potential predictive power of other factors. For the second run, 'Height' and 'Weight'

were excluded to reduce the impact of body-measurement data on the results. Similarly, 'Age' and 'Gender' were excluded to focus specifically on lifestyle factors. The interplay between demographic features and obesity can be instructive for identifying high-risk groups. However, in focusing in on lifestyle factors, the analysis may identify behaviours that are within individuals' power to change. This model achieved an accuracy of 0.75, and the clusters were identified as the most important feature. The third and final run removed clusters to focus solely on lifestyle factors, and achieved an accuracy of 0.71. Such a small decline in accuracy indicates that whilst the clusters have helped identify groupings in the data, they do not have significant predictive power.

Table 4.1: Comparison of Decision Tree Runs: Accuracy and Top 5 Features

(a) Run 1: 0.93

Feature	Imp
Weight	0.555
Height	0.169
FCVC	0.146
Age	0.030
FAVC	0.024

(b) Run 2: 0.75

Feature	Imp
Cluster	0.196
FCVC	0.149
TUE	0.097
FAF	0.097
NCP	0.096

(c) Run 3: 0.71

Feature	Imp
TUE	0.158
FCVC	0.154
CH20	0.128
NCP	0.111
FAF	0.100

The full list of variables and their importance in this final run is captured in **??**, and shows the most significant features were TUE (time on devices) and FCVC (vegetable consumption). The effectiveness of the model itself is captured in the classification report in table **??**. A decision was made not to set a max depth for the decision tree to allow the model to capture more complex patterns. However, this can also increase the risk of overfitting. Looking at the classification report, while some classes, notably Obesity Type I, show lower recall and F1-scores, the overall balanced performance - with an accuracy of 0.71 and similar macro and weighted averages — suggests the model is not severely overfitting. This gives confidence that the model is capturing meaningful patterns across most classes rather than just memorizing the training data.

Class	Precision	Recall	F1-score	Support
Normal_Weight	0.75	0.77	0.76	179
Obesity_Type_I	0.64	0.58	0.61	102
Obesity_Type_II	0.77	0.83	0.80	88
Obesity_Type_III	0.94	0.98	0.96	98
Overweight_Level_I	0.52	0.48	0.50	88
Overweight_Level_II	0.53	0.54	0.54	79
Accuracy			0.71	634
Macro Avg	0.69	0.70	0.69	634
Weighted Avg	0.71	0.71	0.71	634

Table 4.2: Classification Report

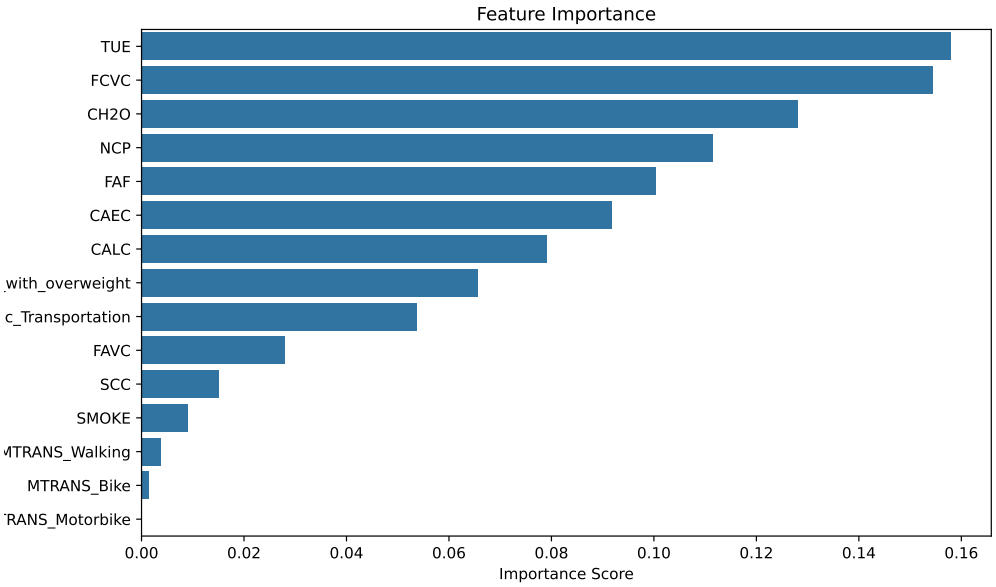


Figure 4.1: Feature Importance Plot for Final Decision Tree

Chapter 5

Discussion

The cluster analysis showed that the cluster associated with the highest obesity group was characterised by a family history of obesity, use of public transport, as well as frequent snacking and consumption of high calorie food. However, the cluster most strongly linked to normal weight also had strong associations with snacking and high calorie food, underscoring the high variability of diet's impact on individuals. It also suggests that whilst dietary habits play a significant role in obesity risk, their influence could be moderated in normal weight groups by other factors such as physical activity levels or a hereditary predisposition towards a certain weight. It could also be indicative of factors not well captured in the data such as variance in individuals' metabolic response to foods, which may allow some individuals to maintain a healthy weight despite a high caloric intake [**Piaggi2019**]. Any cause-effect relationships therefore appear to be obscured by layers of behavioural, demographic and genetic or hereditary interactions. These relationships were further interrogated with the decision-tree.

While the complete decision-tree contains too many branches for straightforward interpretation, examination of the initial branches in Figure **??**, constrained to a maximum depth of 3, illuminates the model's most significant predictive features.

This identifies some surprising relationships. For example, lower vegetable consumption appears to be generally associated with higher weight classes, however higher vegetable consumption together with more time spent on devices is linked to Obesity Type III, the group with the highest BMI. The relationship between FCVC and BMI

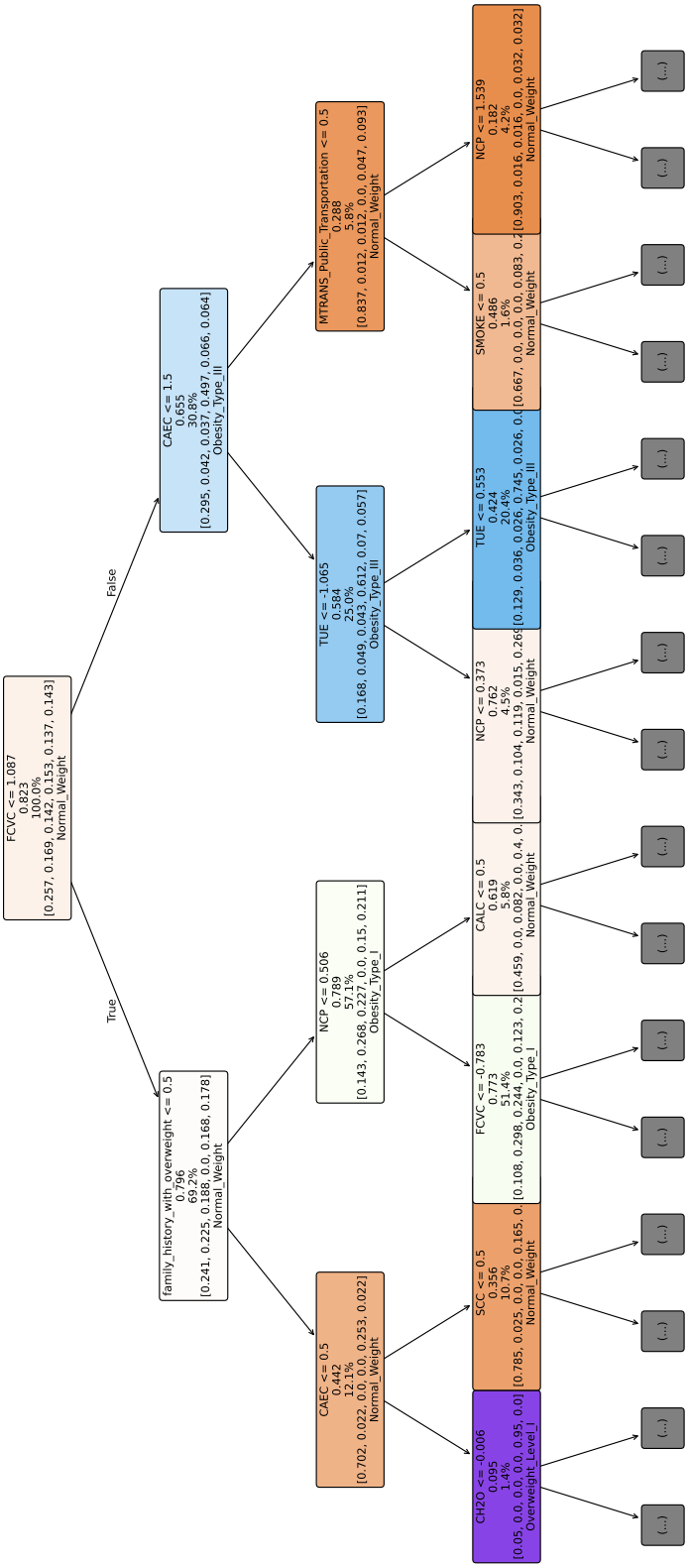


Figure 5.1: Decision tree with a max depth of 3

was picked up in the initial exploratory analysis, but the decision-tree permits a deeper dive into the association, in particular by highlighting the influence of TUE. Whilst vegetable consumption may be predictive of obesity rates, the relationship here is not linear or omnidirectional. Instead, it is highly context dependent. The fact that high vegetable consumption appears to be generally linked to normal weight suggests it may have a protective effect in line with existing literature [**Nour2018**], however this impact may then be cancelled out by low physical activity. There may also be other factors not captured by the model that influence the high consumption amongst the highest BMI individuals. For example, people who are conscious of their weight may increase vegetable intake given public health messaging around the benefits, or it could reflect a 'social desirability bias'. This is where individuals are inclined to self-report socially desirable behaviours, which would lead them to over-report healthy behaviours and under-report unhealthy ones [**Hebert1995**]. Seeking a better understanding of this relationship would be an interesting avenue for future research, given the extent to which it contradicts consensus in the literature.

This identifies some surprising relationships and then there was the other thing and then this thing and then that thing. And then there was another thing and then that other thing.

Chapter 6

Conclusion

This analysis reconfirms that obesity risk is not monolithic, but rather varies greatly across different subpopulations defined by both lifestyle and demographic factors. Whilst traditional metrics remain critical, using more sophisticated statistical tools achieves a richer understanding of obesity and its multifactorial nature. The findings of the study also suggest some potential areas of focus for public health interventions, notably physical activity. Use of public transport and high screentime were two of the study's most significant variables relating to obesity risk, with these factors even potentially nullifying some of the benefits of other healthy behaviours. One potential intervention could therefore be promoting active transport. The initial exploratory analysis showed respondents overwhelmingly rely on public transport, and to a lesser degree automobiles. The low prevalence of cycling and walking at present suggests this could be an impactful way to reduce overall sedentary time in these countries.

Despite these strengths, there are also a number of weaknesses and limitations in the research. The analysis draws from a rich dataset which nevertheless consists primarily of synthetic data, and the utility and generalisability of the data is dependent on the strength of the data generation methods used. Finally, this study is cross-sectional, examining data at a single point in time. The variables studied here and their relationships may vary over time, which could impact the conclusions drawn. Whilst we have already highlighted particular relationships, such as that between high vegetable consumption and the highest BMI, as potentially useful avenues for future

research, further studies should also include longitudinal studies that observe how these factors interact over extended time-frames. Similarly, rerunning the analysis on other datasets could further validate the findings, improve the generalisability of results, and address some of the data quality concerns outlined above. In this way, we can continue to build on the analysis set out here and shed further light on a complex and multifaceted public health issue.

Chapter 7

Appendix

7.1 Development environment

All analyses were implemented in Python 3.12, using Jupyter Notebooks and Neovim to write and test the code. The paper was written in Neovim using the VimTex plugin.

7.2 Supplementary figures

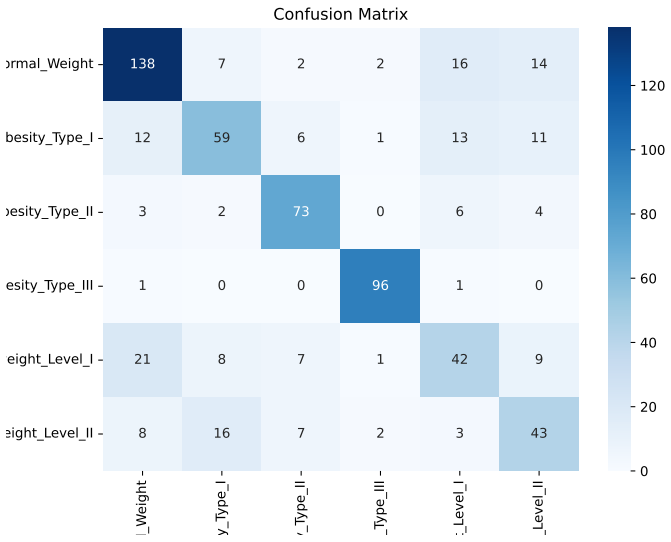


Figure 7.1: Confusion Matrix for Final Decision Tree

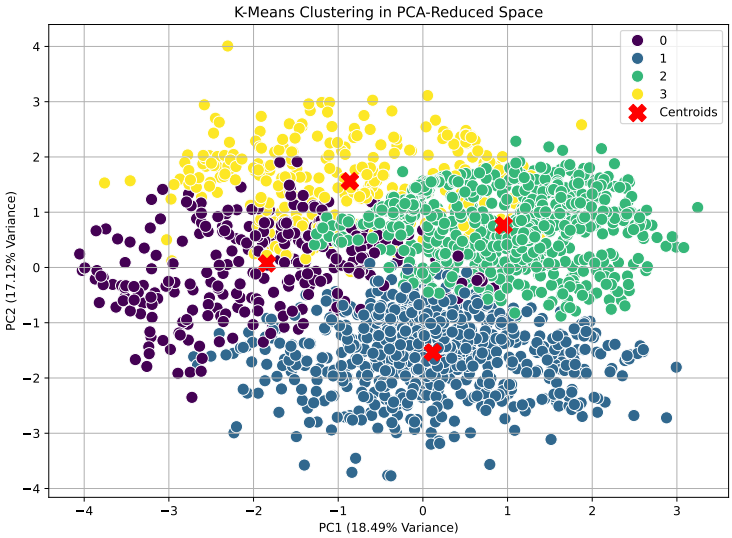


Figure 7.2: K-Means Clustering in PCA-Reduced Space