Estimation of obesity levels based on eating habits and physical condition

University of Strathclyde

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

Obesity has become a critical public health challenge that transcends socioeconomic boundaries. Since 1975, the global incidence of obesity has nearly tripled, transcending its previous association with affluent nations to emerge as a pressing concern in low- and middle-income countries (World Health Organisation, 2021).

But what underpins this surge? Whilst factors such as calorie-dense diets and sedentary lifestyles are evident contributors, the situation is multifaceted. Notably, Latin America presents an alarming trend, with the prevalence of childhood overweight surpassing the global average (UNICEF, 2023).

This report aims to dissect and understand the complex relationship between dietary habits, physical condition, and obesity levels through a thorough analysis of the 2019 dataset provided by Palechor and de la Hoz Manotas. The dataset includes variables from individuals in Colombia, Peru, and Mexico, offering a unique opportunity to explore the intricate dynamics at play within these populations.

This report employs both unsupervised and supervised learning techniques to process and analyse the data. It is structured to first identify and describe the key challenges inherent in the dataset, followed by the application of appropriate statistical methods to uncover underlying patterns and insights.

The overarching goal is to assess the application of classical statistical techniques in contemporary data analysis. This encompasses evaluating the capabilities and limitations of these tools in addressing the identified issues.

In fulfilling these aims, the report will deliver a comprehensive review of the dataset, derive summary statistics, and draw on unsupervised and supervised analysis methods to address the research questions posed. A reflection on the methods used and the complexities encountered will provide a closing evaluation of the analytical processes undertaken.

Chapter 2 - Eating Habits Dataset

The dataset comprises of 2,111 entries, each representing an individual's record spanning across 17 distinct categories. The data has been gathered from individuals from the countries of Mexico, Peru and Colombia. (UC Irvine Machine Learning Repository, 2019)

These entries are complete with no missing values, indicating a well-maintained and comprehensive dataset for analysis. The attributes encapsulate a blend of categorical and numerical data types which include demographic variables such as 'gender' and quantifiable measures including 'Age', 'Height' and 'Weight'. The dataset incorporates factors like 'family_history_with_overweight', 'FAVC' (Frequent consumption of high caloric food), 'FCVC' (Frequency of consumption of vegetables), and 'NCP' (Number of main meals), amongst others, providing a multifaceted view of lifestyle patterns that may correlate with the dependent variable 'NObeyesdad'—an indicator of obesity level. Table 1 shows the categories within 'Obesity_Level', arrayed from the least to the most severe on a scale from 1 to 7. This granular classification enables a nuanced analysis of obesity trends and informs the subsequent application of both unsupervised and supervised learning techniques.

Table 1 - Categories Obesity Levels

1	Insufficient Weight
2	Normal Weight
3	Overweight Level I
4	Overweight Level II
5	Obesity Type I
6	Obesity Type II
7	Obesity Type III

77% of the data was generated synthetically using the Weka tool and the SMOTE Filter while 23% of the data was collected directly from users through a web platform. (Palechor & Manotas, 2019)

The Synthetic Minority Over-Sampling Technique (SMOTE) is an algorithm used to address class imbalance by generating synthetic examples rather than by over-sampling with replacement. It operates by joining the points of the minority class with line segments and producing new points along these lines, thereby synthesizing new minority class instances. (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)

Weka, on the other hand, is a collection of machine learning algorithms for data mining tasks, which includes tools for data pre-processing, classification, regression, clustering, association rules, and visualisation (Witten, 2016).

Chapter 3 – Summary Statistics

Summary statistics are essential in data analysis aw they provide a foundational understanding of the key characteristics of a dataset. They of a description of the sample and measures of the data's variability and overall distribution. (Black, 2019)

Renaming of Variables

To facilitate a more intuitive understanding, certain variables have been renamed for clarity with there changes shown in Table 2 below:

Table 2 - Variable Name Changes

Old Variable Name	New Variable Name
NObeyesdad	Obesity_Level
FAVC	High_Caloric_Food_Frequency
FCVC	Vegetable_Consumption_Frequency
NCP	Main_Meals_Per_Day
CAEC	Snacks_Between_Meals
SMOKE	Smoker
CH2O	Daily_Water_Intake
SCC	Calorie_Intake_Monitoring
FAF	Physical_Activity_Frequency
TUE	Technology_Use_Time
CALC	Alcohol_Consumption_Frequency
MTRANS	Main_Transportation_Mode

Overview

Figure 1 shows an overview of the dataset. The age histogram shows a unimodal distribution skewed towards younger individuals. In terms of weight, males generally outweigh females on average, as shown in the gender comparison bar chart. The weight versus height scatter plot indicates a wide variation in body measurements, notably with many females being shorter yet heavier than males. The obesity level bar chart reveals a uniform distribution across categories, without any significant overrepresentation.

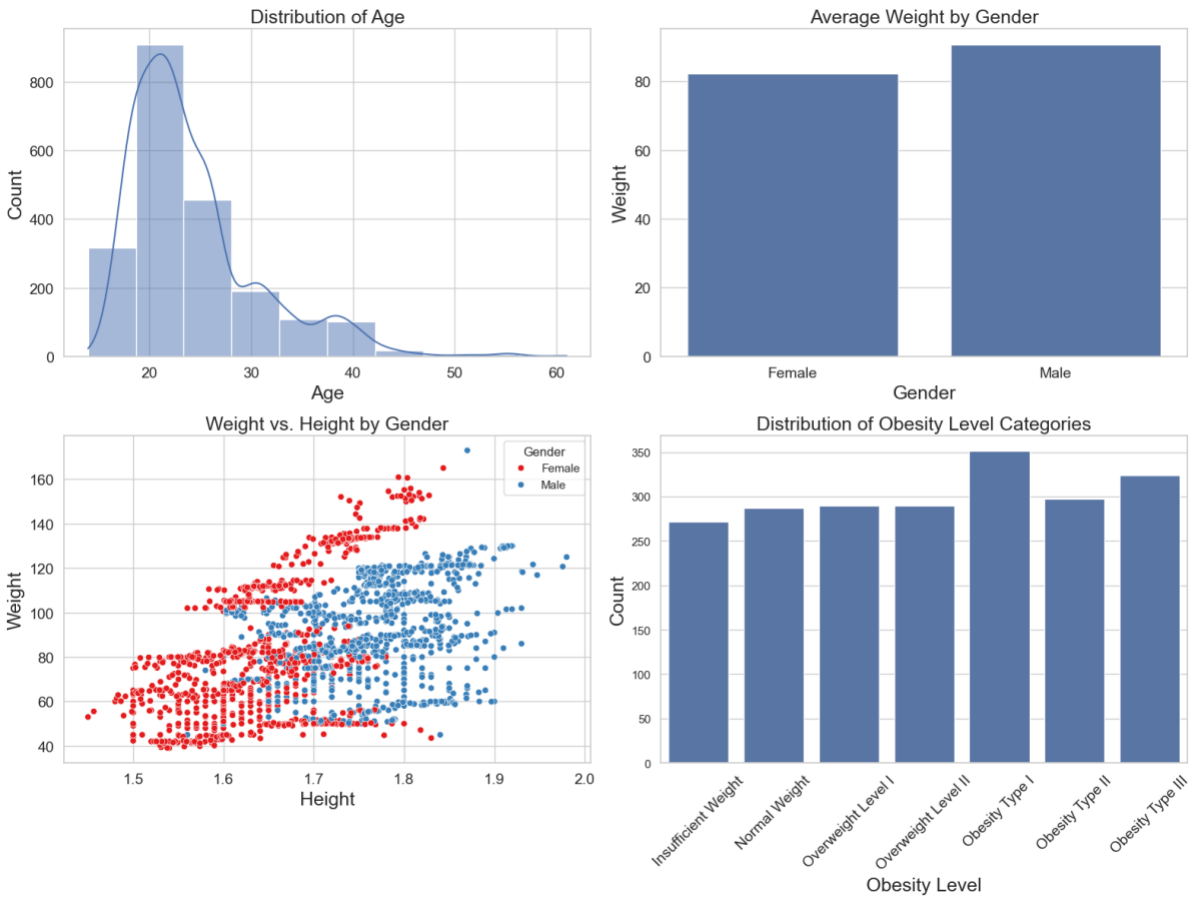

Figure 1- General Summary of Dataset

Figure 2 illustrates the variable distributions within the dataset. The weight distribution is approximately normal, with a majority of data points clustering near the median, and a tail extending towards higher weights hinting at the prevalence of overweight or obesity. The height distribution is multimodal, possibly reflecting diverse subgroups within the population. Daily water intake is notably skewed, with most individuals consuming amounts within a specific range, concentrated heavily around 2 litres, suggesting a possible reporting bias. Lastly, the bimodal distribution of physical activity frequency suggests distinct lifestyle clusters within the dataset.

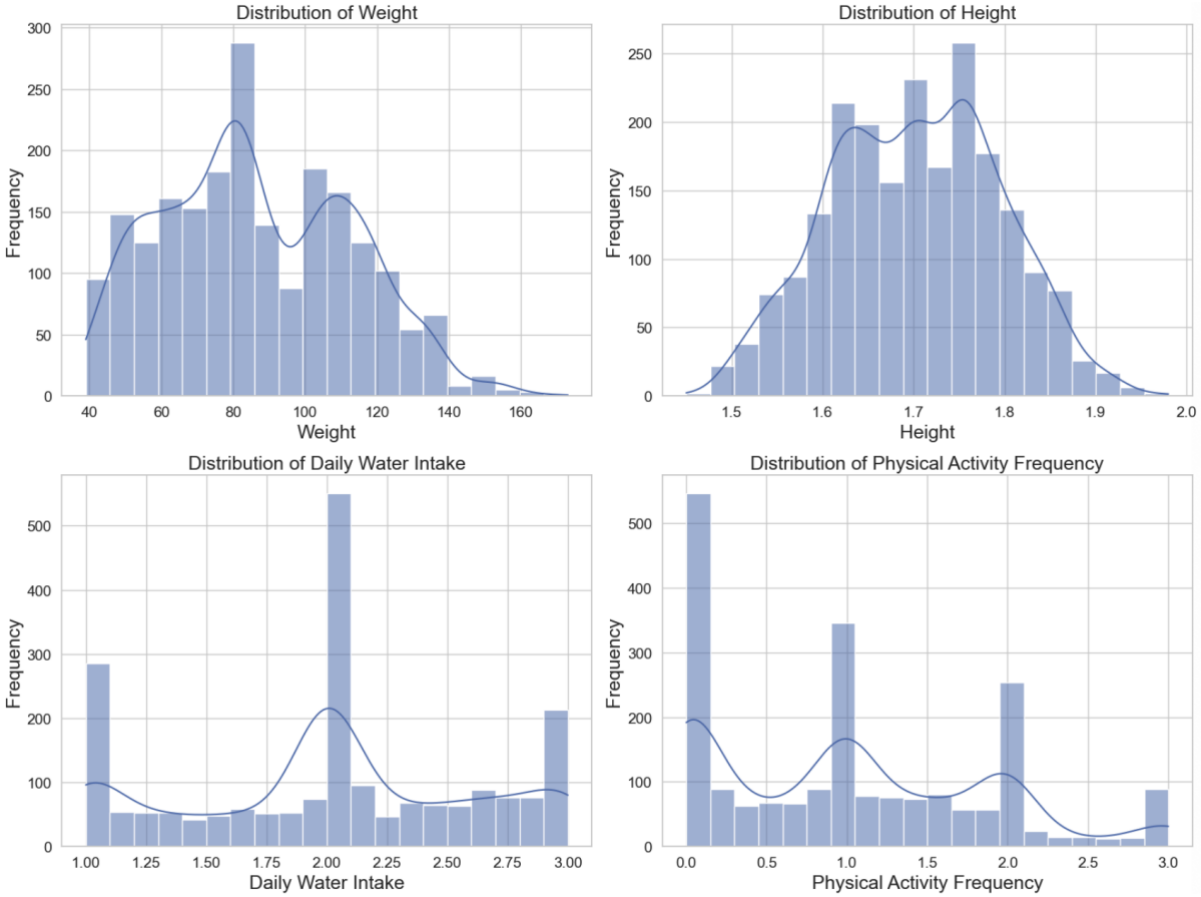

Figure 2 - Histograms for weight, height, water intake, physical activity

BMI

In this report, our focal point is the categorical variable 'obesity level' and its influencing factors. Thus, we will temporarily introduce the Body Mass Index (BMI), a crucial quantitative measure associated with obesity. BMI is calculated as follows:

$$BMI = \frac{weight\ (kg)}{\sqrt{height\ (m)}}$$

Figure 3 illistrates the relationship between BMI and the categorised obesity levels. The incremental increase in median bmi values demonstrates a clear positive correlation between BMI and the severity of obesity. The use of a numerical value will aide in the visusaliation and analysis of the data however, the BMI variable will not be used in the models.

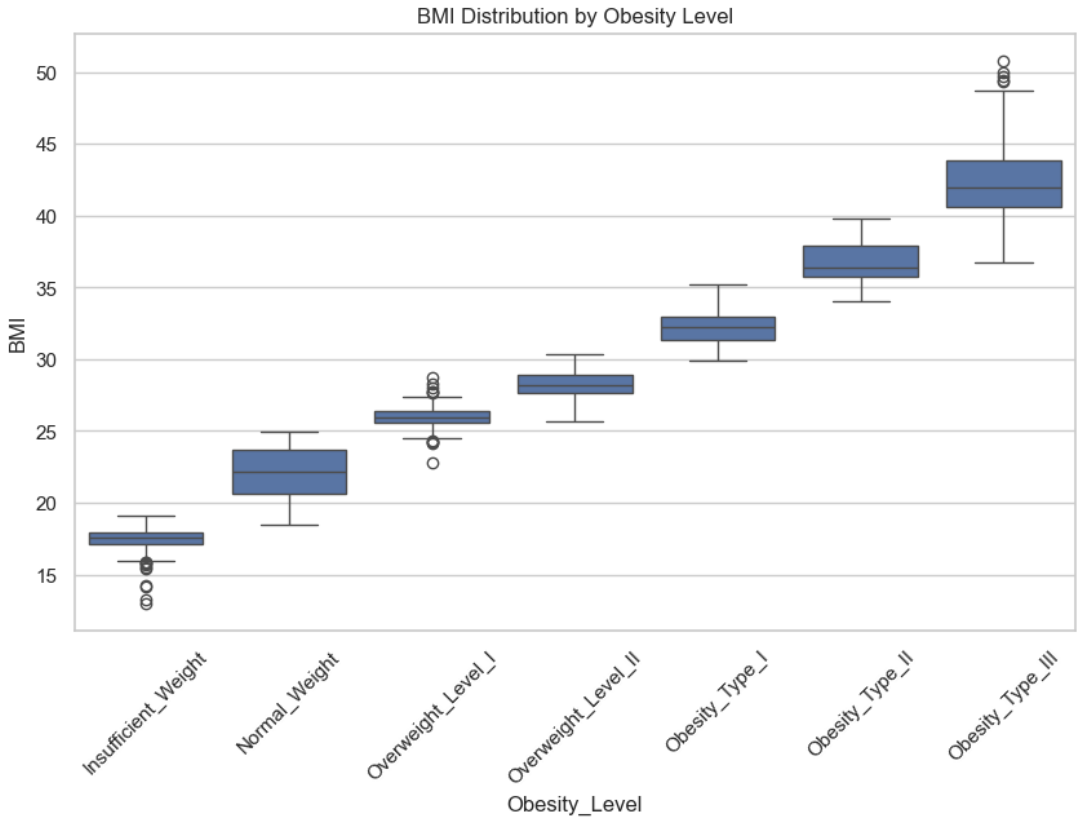

Figure 3 - BMI and Obesity Levels

Analysis

The boxplots in Figure 4 offer a comparative look at BMI distribution across various lifestyle and dietary factors. Gender differences in BMI are observable, with the box plot revealing that females have a slightly wider range of BMI values, while males tend to have a more concentrated distribution around the median BMI. High caloric food consumption shows a marked impact on BMI, with frequent consumers displaying greater variability. Snacking between meals also appears to be associated with higher variation. Notably, smokers tend to have a tighter distribution compared to smokers. Calorie intake monitoring is linked with a lower median BMI, suggesting a potential role in weight management. Alcohol consumption frequency does not show a clear trend, indicating other factors might play a

more significant role. Finally, distribution by main transportation mode reveals that active modes like walking and biking are associated with lower median weights.

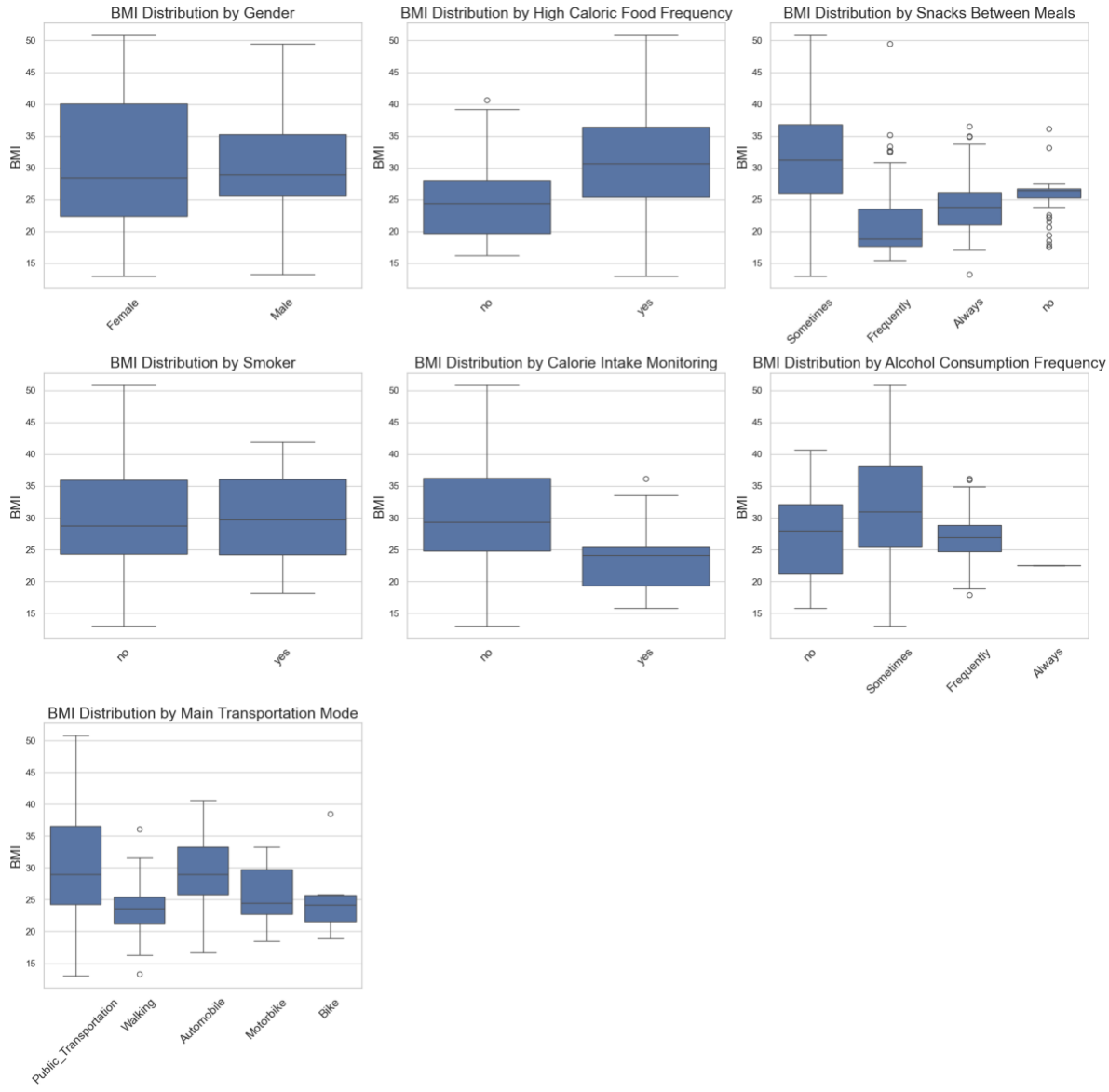

Figure 4 - Boxplots showing weight distribution

Figure 5 reveals robust correlations between BMI and variables such as age, vegetable consumption, and daily water intake. However, the correlation between BMI and physical activity frequency, as well as the number of main meals per day, appears to be more subtle. These observations hint at the intricate connections between lifestyle choices and body mass index.

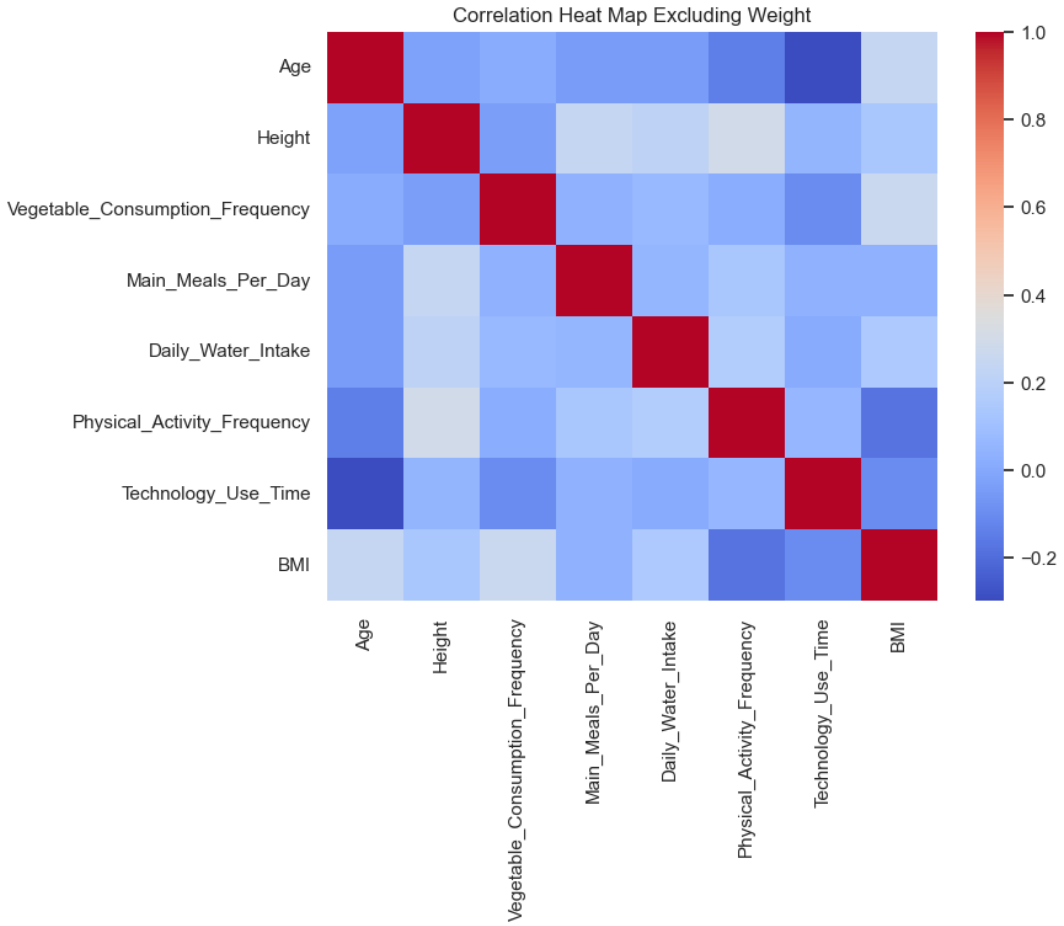

Figure 5 - Correlation Heat Map with BMI

Pre-Processing

Data pre-processing plays a crucial role in preparing the dataset for analysis. In this study, categorical variables such as 'Gender' and 'Main_Transportation_Mode' were encoded using one-hot encoding to convert them into a machine-readable form. Numerical variables like 'Age', 'Height', and 'Weight' underwent scaling to standardise their range, ensuring that the logistic regression model could interpret these features on a comparable basis.

Chapter 4 - Unsupervised Analysis – Clustering

Hierarchal Clustering

Hierarchal clustering constructs a tree over data, the leaves are individual data items while the root is a single cluster that contains all data. There are two main approaches to forming such a tree. This report will use Hierarchical agglomerative clustering (HAC) which starts at the bottom, with every datum in its own singleton cluster, and mergers the groups together. The other method is called divisive clustering, it starts with all the data in one group and then splits it up until every datum is in its own singleton group (Adams, 2016).

Linkage

The linkage criterion determine set within Scikit-Learn determines which distance to use between sets of observation, the algorithm will merge the pairs of clusters that minimise this criterion (Scikit-Learn Agglomerative Clustering, 2024). The below linkage criterion was used:

- Ward – Minimises the variance of the clusters being merged.
- Average – Uses the average of the distances of each observation of the two sets.
- Complete – Uses the maximum distances between all observations of the two sets.
- Single – Uses the minimum of the distances between all observations of the two sets.

Figure 6 below shows the dendrogram for each linkage method, one solution to calculate the number of clusters would be to use the “lastp” truncation method, which involves drawing a horizontal line and observing its intercepts. However, in this report we will use a different method to find the optimal number of clusters.

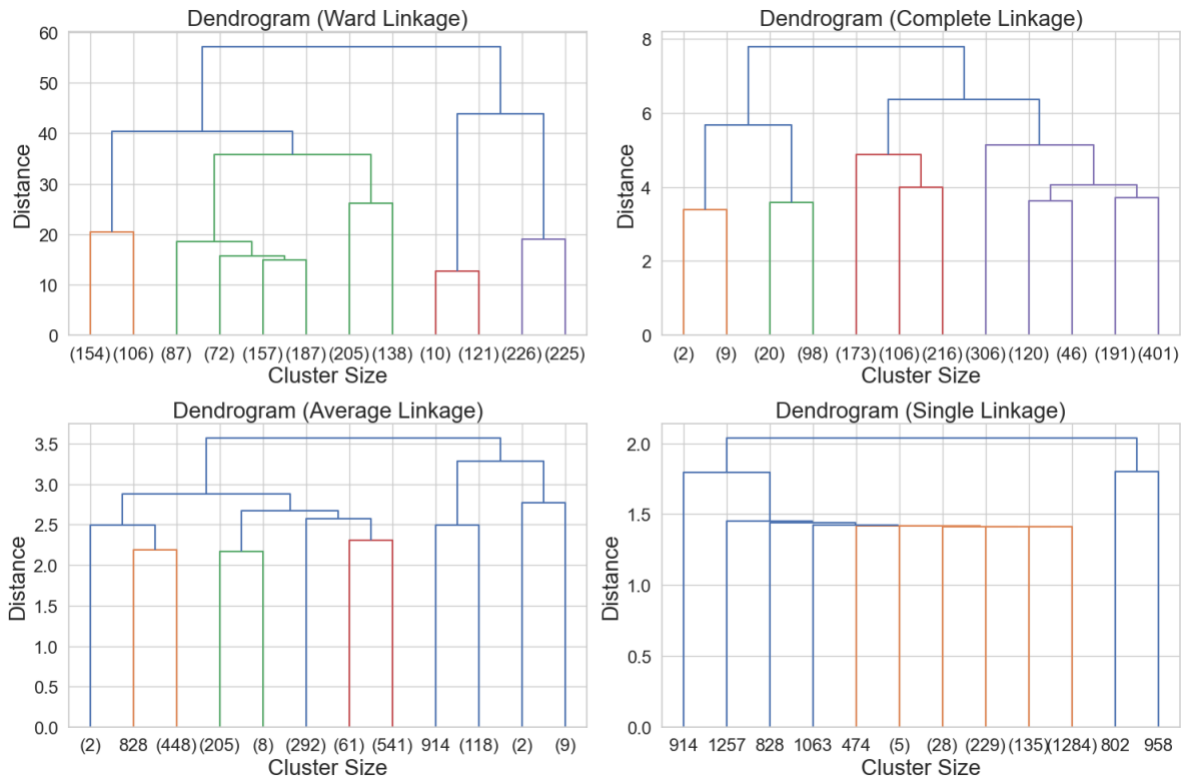

Figure 6 - Dendrograms for different Linkages

Number of Clusters

Determining the optimal number of clusters is a critical step in cluster analysis, ensuring that the resultant clusters are meaningful and actionable. While there are multiple approaches to ascertain this number, each method offers distinct insights based on varying principles. This report emphasises the utilisation of the Silhouette method due to its comprehensive evaluation of cluster quality. However, for a holistic understanding, it's beneficial to briefly discuss alternative methods and their comparative advantages.

- Elbow Method – Consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. (Ketchen & Shook, 1996)
- Davies-Bouldin Index – Internal evaluation method where the optimal number of clusters is determined by the cluster arrangement that yields the lowest Davies-Bouldin index. This index signifies the ‘similarity’ between clusters.

- Silhouette Analysis – Preferred in this study, measures how similar an object is to its own cluster compared to other clusters. A high silhouette value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

In the context of this report Silhouette Analysis was chosen for its balance between interpretability and robustness. By calculating the silhouette scores across a range of cluster numbers for Ward, Complete, Average and Single linkage methods, we could identify the optimal cluster count that maximises the score.

Predictions

The Single linkage method was anticipated to underperform due to its propensity for generating a "chaining" effect, resulting in elongated clusters rather than compact and distinct ones. In contrast, Ward linkage is recognised for its efficacy in creating well-defined clusters of equal variances by optimising the minimisation of within-cluster variance, making it a potentially suitable choice for this dataset. The performance of Complete linkage was considered uncertain, given the initial assessment that the dataset might not feature well-demarcated clusters. However, average linkage emerged as a promising alternative, anticipated to yield favourable results in environments characterised by uniform density and moderate separation among data points.

Results

Table 3 shows the optimal number of clusters used for each option based upon the Silhouette score.

Table 3- Number of Custers

Linkage Option	Number of Clusters
Ward	10
Complete	2
Average Linkage	7
Single Linkage	2

In addition to the Silhouette score, the clustering analysis also employs Completeness and Homogeneity scores. Completeness measures if all members of a given class are within the same cluster, aiming for a score close to 1 for optimal clustering. Homogeneity checks if each cluster contains only members of a single class, with a score of 1 indicating perfect homogeneity. Together, these metrics offer a comprehensive view of clustering performance, ensuring that clusters are both internally consistent and exclusively populated by members of the same class. (Rosenberg & Hirschberg, 2007)

Figure 7 illustrates the clustering performance metrics for various linkage methods. As anticipated, the Ward method outperforms the others in terms of homogeneity, which implies that its clusters are more consistent and contain elements that are more like each other. In stark contrast, the single linkage method has a notably lower homogeneity score, indicating a tendency to form clusters that are less pure with a wider variance within clusters. This underscores the effectiveness of the Ward method in creating distinct and homogeneous groupings within the dataset, which is critical for the reliability of any subsequent analysis.

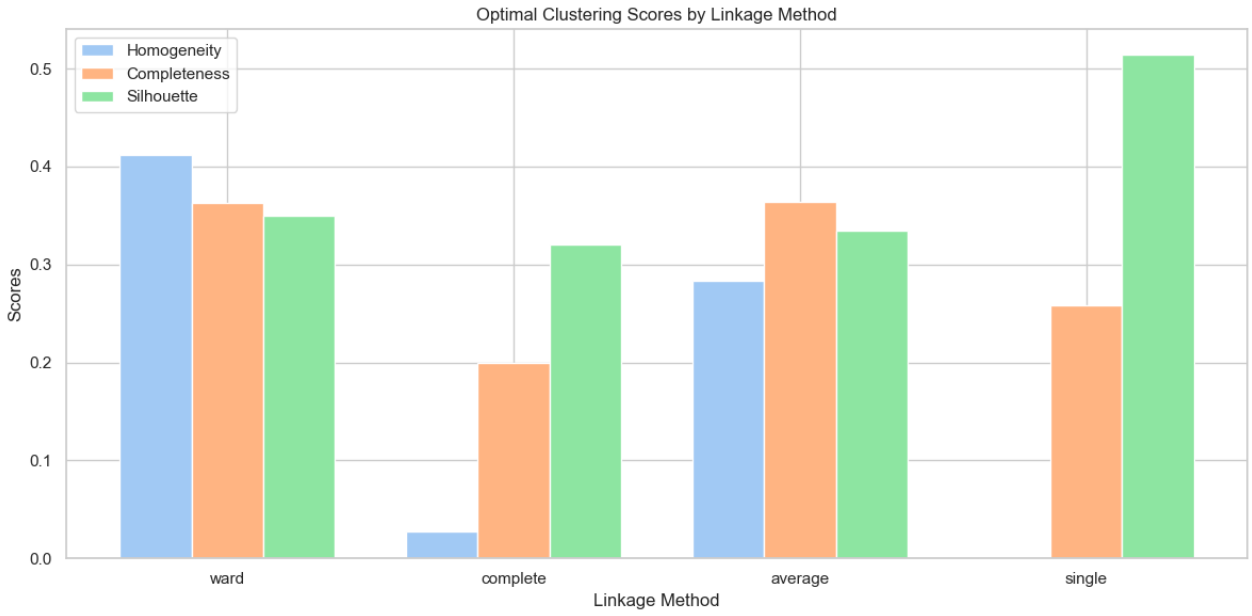

Figure 7 - Cluster Results

K-Means

The K-Means algorithm is another clustering method, its a simple algorithm capable of clustering datasets quickly and efficiently in just a few iterations. Originally proposed by Stuart Lloyds at the Bell Labs in 1957 as a technique for pulse-code modulation (Géron, 2019), some cases today include customer segmentation and cybercrime identification.

The graph associated with K-Means clustering demonstrates the evolution of the Silhouette, Completeness, and Homogeneity scores as the number of clusters increases. The Silhouette score, which gauges the consistency within clusters, remains relatively stable across a range of cluster numbers, suggesting a uniform compactness regardless of the cluster count. On the other hand, the Completeness score shows a steady increase, indicating an improvement in the grouping effectiveness with more clusters, possibly capturing more nuanced groupings within the data. The Homogeneity score, assessing the extent to which clusters contain only data points which are members of a single class, also improves with more clusters, but eventually plateaus, suggesting a limit to the benefits of adding more clusters.

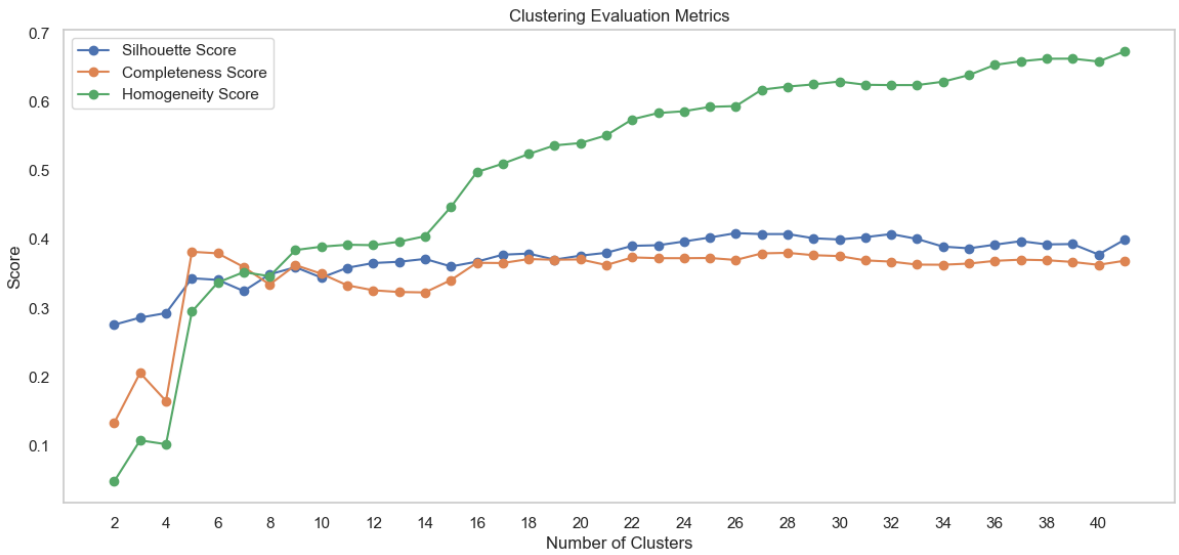

Figure 8 - K-means clustering for different numbers of clusters

From Figure 8 the homogeneity, completeness and Silhouette score rapidly increase after 5 clusters which suggest this is an appropriate number of clusters that balances between having too many small and too few large clusters. Figure 9 visualises the dataset with the clusters colour coded and demonstrates the separation between clusters.

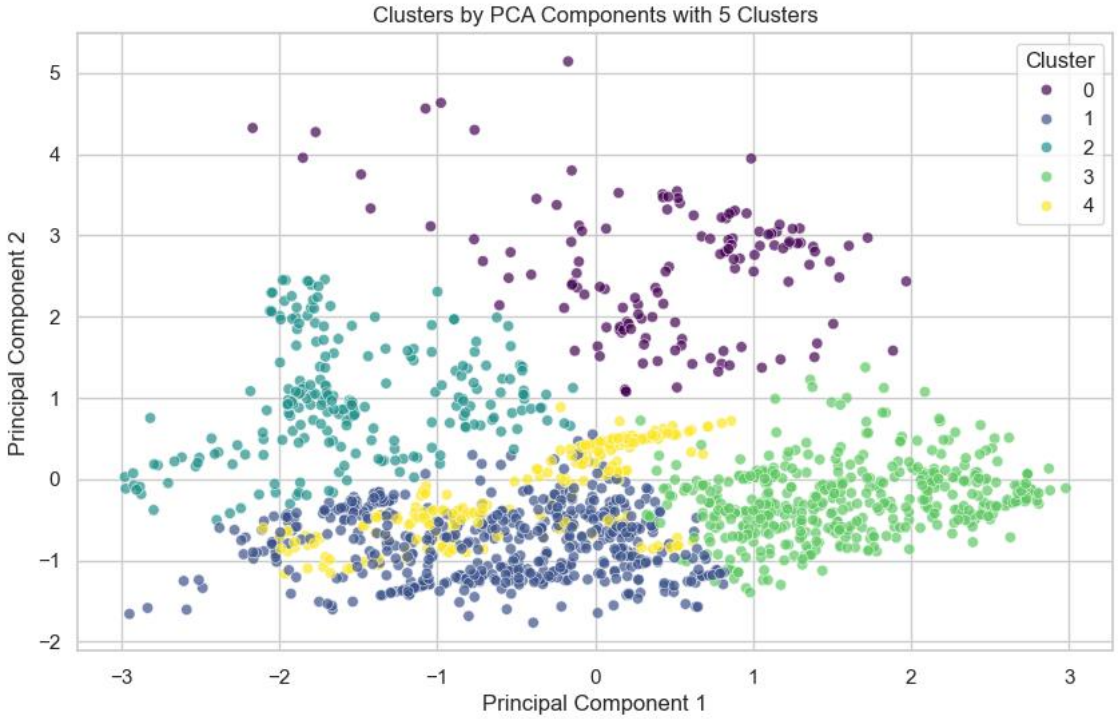

Figure 9 - Principal Component Analysis Scatter Graph

Chapter 5 - Supervised Analysis – Logistic Regression

Supervised learning methods involve training a model on a labelled dataset, where the input attributes (features) and the output (target variable) are known. The model learns to associate patterns in the input data with the corresponding outcomes, enabling it to make predictions on new, unseen data. In the context of this study, supervised learning will be applied to discern which parameters significantly influence the obesity level, a categorical outcome representing different degrees of obesity. The goal is to build a predictive model that can accurately classify individuals into the correct obesity category based on their lifestyle choices and physical characteristics. (James, Witten, Hastie, Tibshirani, & Taylor, 2013)

Logistic Regression

Logistic regression is a statistical technique to evaluate the relationship between various predictor variables (either categorical or continuous) and an outcome which is binary (Ranganathan, Aggarwal, & Pramesh, 2017), in this example the obesity level. Unlike linear regression which outputs continuous number, logistic regression transforms its output using the logistic sigmoid function to return a probability value that can be mapped to two or more discrete classes.

The Scikit-Learn package inside python implements regularised logistic regression with regularisation applied by default (Scikit-Learn Agglomerative Clustering, 2024). Inside the class the ‘max_iter’, which refers to the maximum number of iterations the solver performs to reach convergence has been set to 1000.

Results

To analyse the results several metrics have been used to provide a different perspective on the model’s performance. Together

- Accuracy – Represents the % of model predictions that match the actual obesity levels.

- Precision – Measures the accuracy of positive predictions, it is crucial factor when the cost of a false positive is high.
- Recall – Indicates the model's ability to find all the relevant cases within a dataset.
- F1-Score – The harmonic mean of precision and recall, providing a balance between the two and is useful when the class distribution is uneven.
- Support – Refers to the actual number of references within the class.

As shown in Table 4 the results indicate a strong performance of the logistic regression model in classifying the various levels of obesity. The overall accuracy of 91% is quite high, which suggests that the model is generally effective at making the correct predictions.

Table 4 - Classification Report

Obesity Level	Precision	Recall	F1-Score	Support
Insufficient Weight	0.93	0.98	0.95	54
Normal Weight	0.90	0.79	0.84	58
Obesity Type I	0.90	0.99	0.94	70
Obesity Type II	0.98	1.00	0.99	60
Obesity Type III	1.00	0.98	0.99	65
Overweight Level I	0.75	0.88	0.81	58
Overweight Level II	0.91	0.71	0.80	58
Average	0.91	0.91	0.91	0.91

The precision values are mostly over 0.9 which indicates across these classes the model is usually correct. However, for Obesity Level I it seems to have some difficulty recording a value of 0.75. This lower precision suggests there may be some confusion with this class and other classes. The high recall of 0.71 to 1 suggests that the model can identify most of the true positive for each class. However, the recall for Normal Weight and Overweight Level II is slightly lower which means it could have difficulty distinguishing between these classes. The F1-score, which is a balance between precision and recall is also high for most classes, indicating a balanced performance. The support reflects a balanced distribution of classes, which is good for model evaluation.

While these results are promising, they should be interpreted in the context of the dataset size and complexity. A dataset of 423 instances for the test set is moderate but might not cover all possible variations in a real-world scenario.

Cross Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models. For the logistic model used in this report, 5-fold cross-validation was performed meaning the data was split into 5 parts and trained/ evaluated 5 times, each time using a different part as the test set and the remaining parts as the training set.

Table 5 - Cross-Validation Results

Test Number	Cross-Validated Scores
1	0.88
2	0.89
3	0.88
4	0.89
5	0.85
Average	0.88

The cross-validation resulted in an average score of 0.87 across the 5 iterations of cross-validation. The scores are consistent, with only slight variations between each fold which indicates that the model is stable and provides reliable predictions.

Chapter 6 – Reflections

Reflecting on the methods used for analysis in this report, both un-supervised hierarchal clustering and supervised logistic regression have played a pivotal role in unravelling the complexity of obesity levels based upon a range of variables provided in the dataset.

Hierarchical agglomerative clustering offered an exploration of the data structure, enabling the identification of natural groupings within the dataset. The utilisation of various linkage criteria – Ward, Average, Complete and Single – allowed for a comprehensive examination of how different approaches to measuring distances between clusters can impact the resulting hierarchical tree. Silhouette analysis, employed to ascertain the optimal number of clusters, offered a robust framework for discerning the most suitable cluster count, ensuring clusters are both meaningful and distinct.

It's evident that the choice of linkage criteria significantly influenced the analysis outcome. The Ward linkage proved to be the all-round most effective for this dataset, with its relatively balanced scores across the silhouette, homogeneity, and completeness metrics. However, the limitations of the complete and single linkage methods, particularly in achieving high homogeneity, underscore the necessity of carefully selecting linkage criterion based on the dataset's characteristics.

Logistic regression was a pertinent choice for predicting categorical outcomes such as obesity levels. The method's ability to handle binary outcomes and provide probabilistic predictions made it suitable for classifying individuals into distinct obesity categories based on a set of predictors. However, the lower precision observed for certain classes, such as Overweight Level I, highlights potential areas for model refinement. This suggests that the model may benefit from additional features or alternative algorithms that could provide clearer boundaries between similar classes. Moreover, while the cross-validation scores

indicate a stable and reliable model, further investigation into the model's performance with a larger and more varied dataset would be beneficial to ensure its robustness and applicability in broader scenarios.

Through rigorous analysis using hierarchical clustering and logistic regression, we've uncovered patterns and relationships that underscore the complexity of this global health issue. The methodologies applied here—reflective of the robustness of statistical learning techniques—have revealed insights that extend beyond the data, hinting at the broader societal and behavioural factors at play.

References

Adams, R. (2016). *Hierarchical Clustering.* Princeton: Princeton University.

Black, K. (2019). *Business for Contemporary Decision Making Statistics.* Houstan: John Wiley & Sons.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.* San Francisco: O'Reilly.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2013). *An Introduction to Statistical Learning.* Toronto: Springer.

Ketchen, D., & Shook, C. (1996). *The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique* . Baton Rouge: Strategic Management Journal.

Ranganathan, P., Aggarwal, R., & Pramesh, C. (2017). *Common pitfalls in statistical analysis: Logistic regression.* Mumbai: Department of Anaesthesiology, Tata Memorial Centre.

Rosenberg, A., & Hirschberg, J. (2007). *V-Measure: A conditional entropy-based external cluster evaluation measure.* New York: Columbia University.

Scikit-Learn Agglomerative Clustering. (2024, 02 21). Retrieved from Scikit-Learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#:~:text=The%20linkage%20criterion%20determines%20which,observation%20of%20the%20two%20sets.

UC Irvine Machine Learning Repository. (2019, August 26). *Estimation of obesity levels based on eating habits and physical condition*. Retrieved from UC Irvine Machine Learning Repository:

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

UNICEF. (2023, August 31). *More than 4 million children under 5 have overweight in Latin America and the Caribbean*. Retrieved from UNICEF: https://www.unicef.org/lac/en/press-release/more-4-million-children-under-5-overweight-latin-america-caribbean

World Health Organisation. (2021, June 9). *WHO Inline*. Retrieved 02 20, 2024, from Obseity and Overweight: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Appendicies

- Language: Python 3.12.2
- IDE: Visual Studio Code 1.86.0
- Packages Used:
    1. Numpy
    2. Pandas
    3. Matplotlib
    4. Seaborn
    5. Sklearn
    6. Scipy
- Code stored as a separate “ipynb” file and attached separately to this report.