University of Strathclyde Glasgow

CS818: Big Data Fundamentals

Estimation of Obesity Levels Based on Eating habits and Physical Activity

Contents

1. Introduction 1
2. Dataset 2
   2.1 Dataset Overview 2
   2.2 Statistical Summary 3
3. Exploratory Data Analysis 6
   3.1 Distribution & Outliers 6
   3.2 Correlation Analysis 8
4. Unsupervised Learning 9
   4.1 Optimal Cluster Selection 10
   4.2 Cluster Interpretations 11
   4.3 Hierarchical Clustering 13

.... 13

5. SHAP Analysis 14
6. Supervised Learning 16
   6.1 Model Performance Overview 16
   6.2 Evaluating Model Performance 18
7. Conclusion 20
   7.1 Key Findings 20
   7.2 Limitations & Future Work 21
8. References 22
9. Appendix 23
   9.1 Variables Table 23
   9.2 Development Details 23

1. Introduction

Introduction

Obesity is a significant global health issue, contributing to chronic conditions such as diabetes, cardiovascular disease, and metabolic disorders. According to the World Health Organisation (WHO, 2023), “the prevalence of obesity has tripled worldwide since 1975, largely due to changing dietary habits and reduced physical activity levels.” Obesity is a multifactorial condition influenced by genetics, lifestyle, and dietary habits, making its prediction and prevention a priority in public health initiatives.

Why Eating Habits and Physical Conditions Matter

Eating habits and physical conditions play a crucial role in determining an individual's risk of obesity. (CDC, 2023) “Poor dietary choices, such as high fast-food consumption, low fruit and vegetable intake, and excessive sugar consumption, have been identified as major risk factors for obesity.” Additionally, (NHS, 2023) “sedentary behaviour and insufficient physical activity contribute to increased body weight and related health complications.” Identifying patterns in these lifestyle factors can help predict obesity levels and develop targeted interventions to mitigate its impact.

Brief Dataset Introduction

This study utilises the Obesity Level Estimation Dataset from the UC Irvine Machine Learning Repository (Palechor & de la Hoz Manotas, 2019). The dataset comprises demographic, dietary, and physical activity attributes, enabling the classification of individuals into different obesity levels. By analysing these features, we aim to uncover key determinants of obesity and assess the effectiveness of predictive models.

Research Questions

- Can obesity levels be predicted based on eating habits and physical activity?
- Which factors contribute most to obesity?

Throughout this report, the first question will be referred to as RQ1 and the second question will be referred to as RQ2.

Statistical Techniques

This study employs a combination of statistical and machine learning techniques to analyse obesity levels based on lifestyle factors:

- **Descriptive Statistics** – Summarises key dataset characteristics, including BMI distribution, meal frequency, and snacking habits.
- **Exploratory Data Analysis (EDA)** – Uses correlation analysis and visualisations to identify relationships between dietary patterns and obesity.
- **Unsupervised Learning (Clustering**) – Applies K-Means and Hierarchical Clustering to uncover distinct obesity-prone groups.
- **Feature Importance Analysis (SHAP)** – Determines which lifestyle factors contribute most to obesity classification.
- **Supervised Learning Models** – Trains Random Forest, SVM, Decision Tree, KNN, and Logistic Regression to predict obesity levels.
- **Classification Evaluation** – Assesses model performance using accuracy, precision, recall, and F1-score.

2. Dataset

2.1 Dataset Overview

The Obesity Level Estimation Dataset consists of 2,111 records collected from individuals in Colombia, Peru, and Mexico, containing demographic, dietary, and physical activity-related attributes. The dataset categorises obesity levels into seven groups, ranging from *Insufficient Weight* to *Obesity Type III*. A detailed description of all dataset variables is provided in the appendix.

The dataset features key variables influencing obesity classification that link directly back to the research questions:

- **Eating Habits –** Frequency of main meals (NCP), vegetable intake (FCVC), water consumption (CH2O), snacking habits (CAEC), and high caloric food intake (FAVC).
- **Physical Activity –** Physical activity frequency frequency (FAF) and technology use time (TUE).
- **Demographic –** Age & gender

The dataset includes the **NObeyesdad** variable, which classifies individuals into categorical obesity levels. While this variable is useful for classification tasks, BMI was calculated to provide a continuous measure of body composition, allowing for more nuanced analysis.

Table 1.1: Classification of BMI Ranges and Corresponding NObeyesdad Labels (García et al., 2019)

NObeyesdad label	**BMI range**
Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight	25.0 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

2.2 Statistical Summary

Table 2.1: Summary Statistics of Key Variables

Variable	**Mean**	**Std Dev**	**Analysis**
BMI	29.70	8.01	Obesity threshold marker
FCVC	2.41	0.53	Higher values linked to lower obesity risk

NCP	2.69	0.77	Regular meal intake pattern
CH2O	2.00	0.61	Limited impact on obesity classification
FAF	1.01	0.85	Low exercise levels in dataset
TUE	0.65	0.60	Minimal influence on obesity levels

Statistical Summary Analysis

The table presents key summary statistics for selected variables relevant to obesity classification. These variables represent dietary habits and physical activity, which are crucial in understanding obesity risk.

Key Observations:

1. **Body Mass Index (BMI):**
   - The mean BMI is 29.7, which falls within the overweight category (25–29.9), indicating that most individuals in the dataset are either overweight or obese.
   - A relatively high standard deviation (8.01) suggests notable variation in BMI values among individuals.
2. **Vegetable Consumption Frequency (FCVC):**
   - The mean FCVC is 2.41, indicating moderate vegetable intake.
   - Higher FCVC values are generally associated with lower obesity risk, suggesting that increased vegetable consumption may contribute to weight management.

3. **Number of Main Meals per Day (NCP):**

    - The mean NCP is 2.69, implying that most individuals consume around 2 to 3 main meals per day.

    - Regular meal patterns are linked to better metabolic regulation, reducing the likelihood of unhealthy weight gain.

4. **Water Consumption Frequency (CH2O):**

    - The mean CH2O is 2.00, suggesting moderate water intake.

    - A standard deviation of 0.61 shows minor variations among individuals. However, its impact on obesity classification appears to be limited.

5. **Physical Activity Frequency (FAF):**

    - The mean FAF is 1.01, indicating low engagement in physical exercise.

    - This highlights a sedentary trend within the dataset, potentially contributing to higher obesity prevalence.

6. **Technology Usage (TUE):**

    - The mean TUE is 0.65, reflecting minimal time spent using technology-based tools.

    - The impact of this factor on obesity classification appears to be insignificant.

Implications for Obesity Classification

- The dataset suggests a higher prevalence of overweight and obese individuals, as reflected by the BMI mean.

- Dietary habits (FCVC, NCP) and physical activity (FAF) play a significant role in weight regulation, reinforcing their relevance in obesity classification models.

- The variability in individual behaviours (e.g., low FAF values) highlights the need for personalised health interventions targeting lifestyle changes.

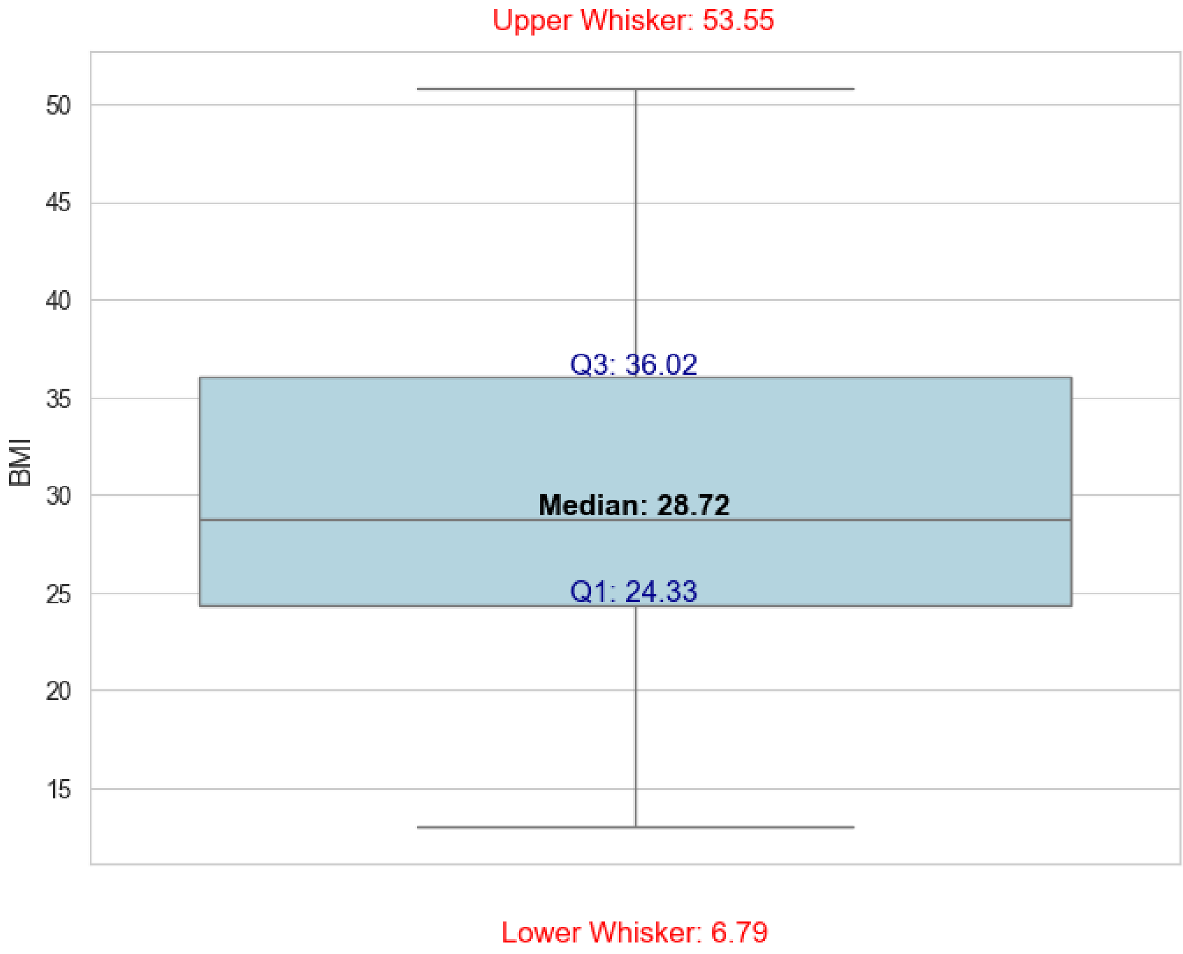
This statistical summary provides foundational insights for further exploratory and predictive modelling, helping to assess which lifestyle factors contribute most to obesity classification.

3. Exploratory Data Analysis

3.1 Distribution & Outliers

EDA was conducted to assess variable distributions, correlations, and potential predictors of obesity.

Figure 3.1: Boxplot of BMI distribution.



Key Observations

The box plot provides insights into the distribution of BMI values within the dataset.

- Median BMI: 28.72 → Falls within the overweight range (25–29.9), suggesting that the majority of individuals are overweight.
- Interquartile Range (IQR: 24.33 – 36.02) → Indicates that 50% of the population falls between Normal Weight (18.5–24.9) and Obesity Type I (30–34.9).
- Whiskers (6.79 – 53.55) → Highlights the full spread of BMI values, with outliers at both extremes.
    - Lower Whisker (6.79) – Suggests possible data errors or rare underweight cases.
    - Upper Whisker (53.55) – Confirms the presence of severe obesity cases (Obesity Type III, BMI > 40).

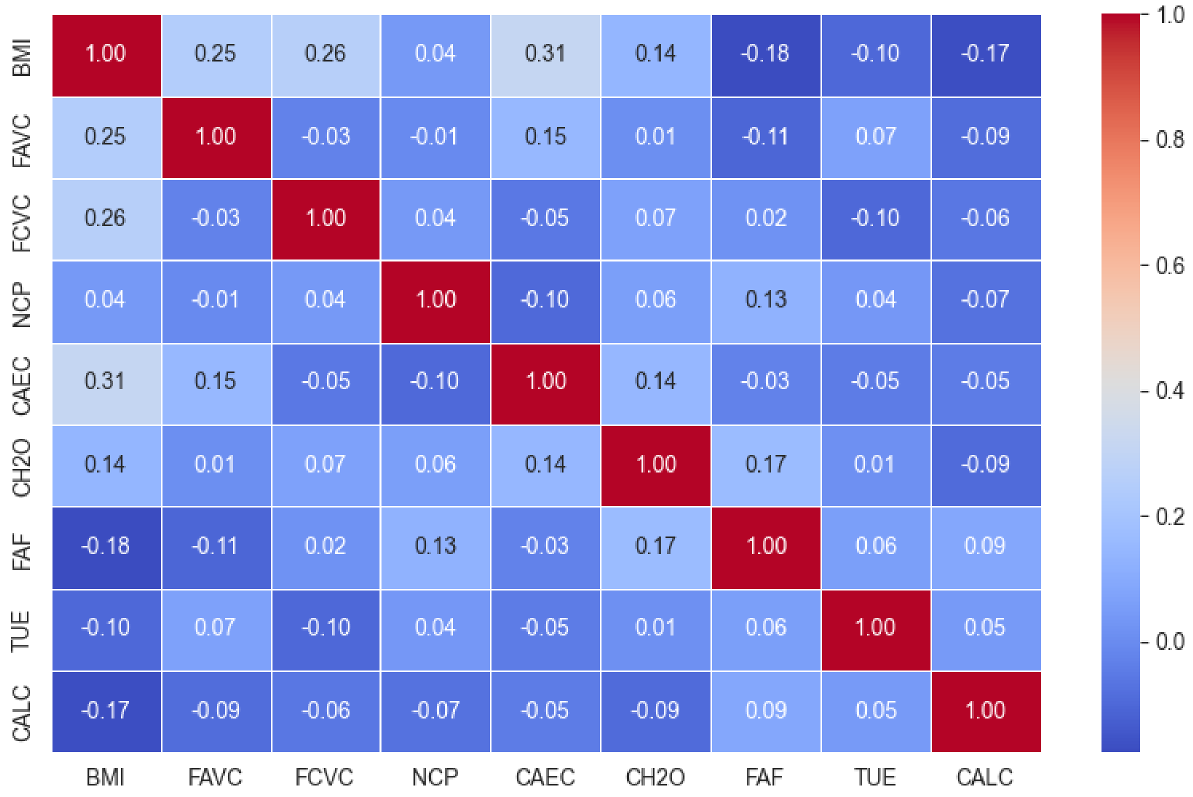
Key Insights & Relevance to Research Questions

RQ1: The concentration of higher BMI values reinforces the predictability of obesity levels based on lifestyle habits.

RQ2: The wide IQR and high upper whisker suggest that dietary habits significantly influence BMI variation, with high-caloric food intake and snacking likely contributing to obesity.

3.2 Correlation Analysis

Figure 3.2: Correlation Matrix of Obesity-Related Variables



	BMI	FAVC	FCVC	NCP	CAEC	CH2O	FAF	TUE	CALC
BMI	1.00	0.25	0.26	0.04	0.31	0.14	-0.18	-0.10	-0.17
FAVC	0.25	1.00	-0.03	-0.01	0.15	0.01	-0.11	0.07	-0.09
FCVC	0.26	-0.03	1.00	0.04	-0.05	0.07	0.02	-0.10	-0.06
NCP	0.04	-0.01	0.04	1.00	-0.10	0.06	0.13	0.04	-0.07
CAEC	0.31	0.15	-0.05	-0.10	1.00	0.14	-0.03	-0.05	-0.05
CH2O	0.14	0.01	0.07	0.06	0.14	1.00	0.17	0.01	-0.09
FAF	-0.18	-0.11	0.02	0.13	-0.03	0.17	1.00	0.06	0.09
TUE	-0.10	0.07	-0.10	0.04	-0.05	0.01	0.06	1.00	0.05
CALC	-0.17	-0.09	-0.06	-0.07	-0.05	-0.09	0.09	0.05	1.00

Correlation Heatmap Analysis

This correlation heatmap illustrates the relationships between BMI and various lifestyle factors, where red indicates stronger positive correlations and blue represents negative or weak correlations.

Key Findings

BMI is positively correlated with snacking frequency (CAEC = 0.31), vegetable intake (FCVC = 0.26), and high-caloric food consumption (FAVC = 0.25) → Suggesting that frequent snacking and high-calorie food choices are key contributors to obesity.

BMI has a negative correlation with physical activity (FAF = -0.18) → Reinforcing the expectation that lower physical activity levels are associated with higher obesity risks.

Alcohol consumption (CALC = -0.17) and screen time (TUE = -0.10) have weak negative correlations with BMI → Indicating they may not be strong predictors of obesity in this dataset.

No strong correlations exist between meal frequency (NCP) and BMI → Suggesting that meal timing alone does not significantly influence obesity levels.

Relevance to Research Questions

RQ1: This confirms that dietary habits (snacking and high-caloric food intake) are stronger obesity predictors than physical activity in this dataset.

RQ2: High-caloric food consumption (FAVC), snacking (CAEC), and low physical activity (FAF) emerge as key contributors to obesity, supporting the selection of these variables for predictive modelling.

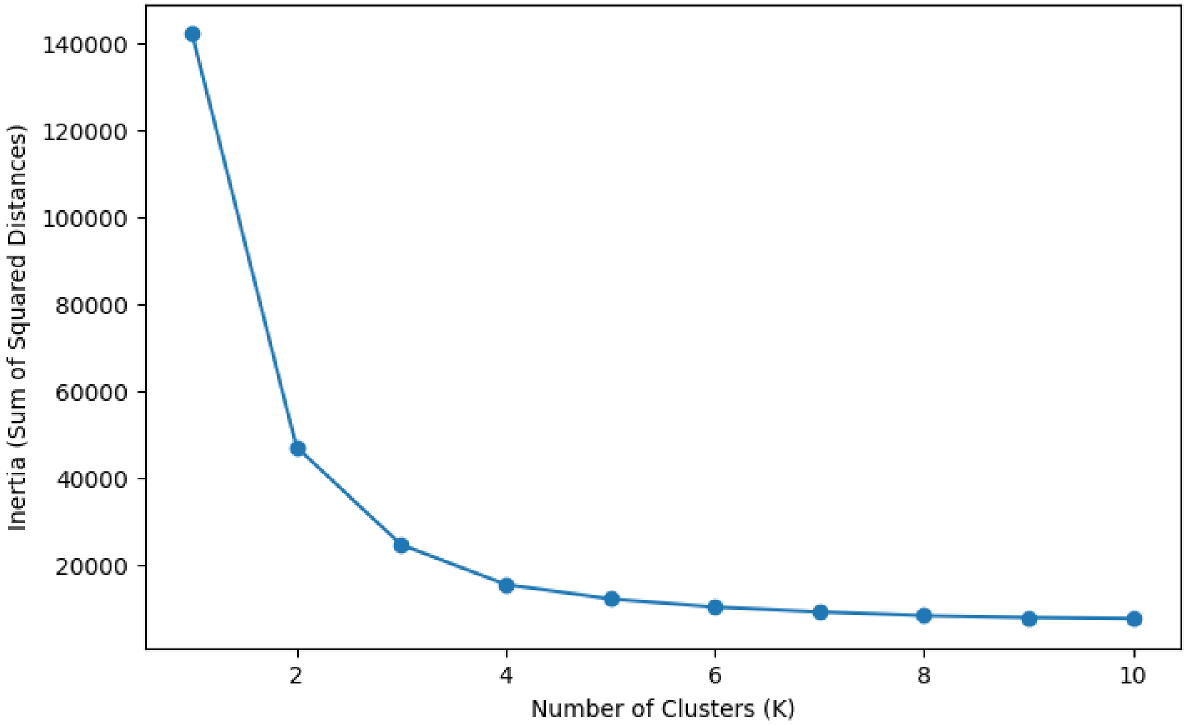
Conclusion: These insights reinforce that eating habits play a more dominant role than physical activity in obesity classification within this dataset.

4. Unsupervised Learning

Clustering was applied to identify distinct obesity-prone groups based on eating habits and activity levels.

4.1 Optimal Cluster Selection

Figure 4.1: Elbow Method for Optimal Cluster Selection



Elbow Method Analysis

The Elbow Method is used to determine the optimal number of clusters (K) by plotting inertia (sum of squared distances) against different values of K. The goal is to identify the elbow point, where adding more clusters results in diminishing returns in reducing inertia.

Key Findings

Sharp decline in inertia from K = 1 to K = 3 → Indicates that increasing clusters initially improves within-cluster compactness significantly.

Elbow Point at K = 3 or K = 4 → Beyond this, the rate of inertia reduction slows, suggesting that 3 or 4 clusters effectively balance compactness and model complexity.

Further increases in K show marginal gains → Suggesting that adding more clusters beyond K = 4 provides limited improvement in grouping structure.

Relevance to Research Questions

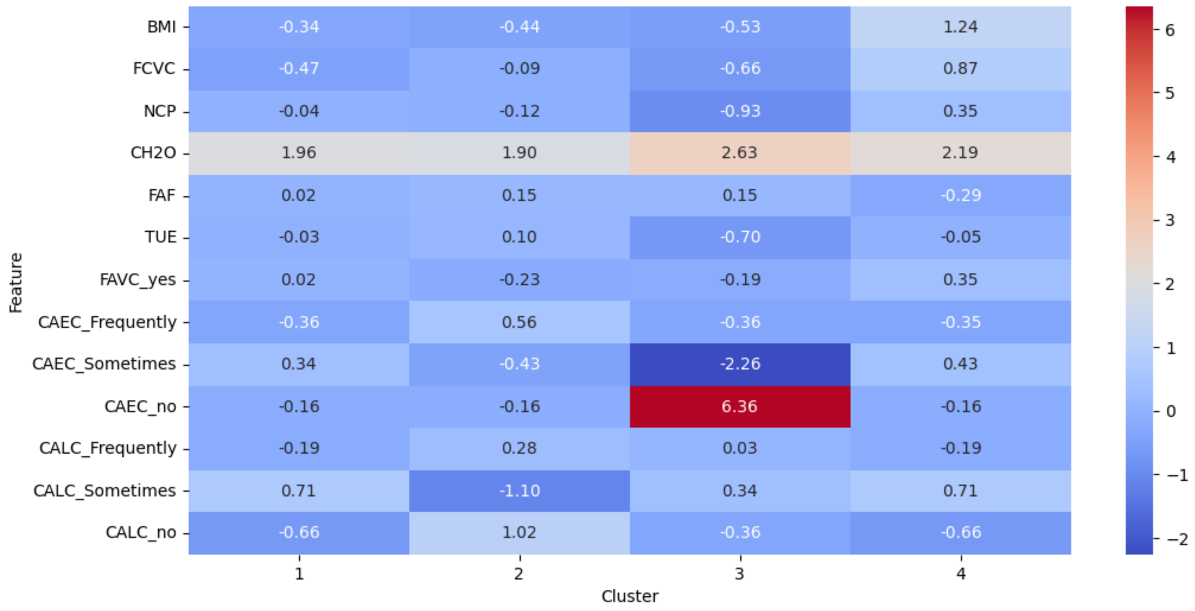
RQ1: Identifying the optimal number of obesity-related clusters enables a better understanding of distinct obesity risk groups based on eating habits and lifestyle patterns.

RQ2: Clustering helps reveal patterns in lifestyle factors contributing to obesity, which can be validated through supervised learning models.

Conclusion: The analysis suggests that 3 or 4 clusters best represent the obesity-related patterns within the dataset.

4.2 Cluster Interpretations

Figure 4.2: Heatmap of KMeans Clustering Centroids



Cluster Feature Importance Analysis

This heatmap illustrates how different lifestyle and dietary habits influence each identified cluster, with colours representing deviations from the mean. Red indicates a high positive contribution, while blue represents a negative or lower contribution.

Key Findings

Cluster 1-3 → Characterised by higher water consumption (CH2O ~1.96 - 2.63) but lower BMI (-0.34 to -0.53), suggesting a possible link between hydration and healthier weight management.

Cluster 3 → Shows a strong positive correlation with snacking frequency (CAEC_no = 6.36), suggesting that this group consists of individuals who rarely snack but may compensate with other high-calorie food choices.

Cluster 4 → Displays the highest BMI (1.24), indicating that individuals in this group are at greater risk of obesity-related conditions. Lower physical activity (FAF = -0.29) and moderate snacking behaviours (CAEC_Frequently = -0.35) suggest a sedentary lifestyle with occasional unhealthy eating habits.

Alcohol Consumption (CALC) → Shows mixed contributions across clusters but does not appear to be a dominant factor affecting obesity levels.

Relevance to Research Questions

RQ1: The distinct clustering patterns reinforce that obesity levels can be predicted based on eating habits rather than solely on physical activity.

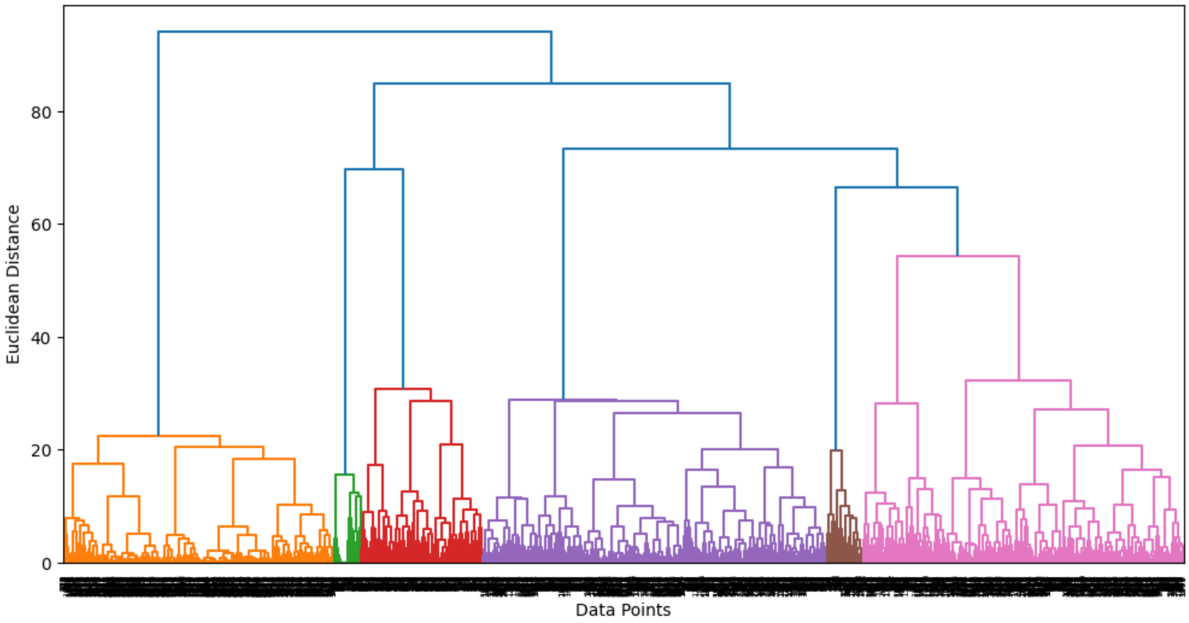
RQ2: The most influential factors contributing to obesity across clusters are snacking frequency (CAEC), water intake (CH2O), and physical activity (FAF). These findings align with supervised learning insights, confirming their importance in obesity classification.

Conclusion: The clustering approach effectively differentiates individuals based on eating habits and obesity risk levels, supporting the predictive modelling approach.

4.3 Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique that groups similar data points into clusters by iteratively merging or splitting clusters based on their similarity. Unlike K-Means, which requires a pre-defined number of clusters (K), hierarchical clustering produces a dendrogram, allowing for a more flexible determination of clusters.

Figure 4.3: Dendrogram for Hierarchical Clustering



Hierarchical Clustering (Dendrogram) Analysis

This dendrogram represents the hierarchical clustering structure of the dataset, displaying how data points are grouped based on Euclidean distance.

Key Findings

- Three to four main clusters emerge at higher linkage distances, reinforcing the findings from the elbow method.
- Smaller clusters appear at the lower levels, indicating finer subgroup distinctions based on lifestyle and dietary habits.

- The height of vertical merges represents the distance between clusters, with larger distances suggesting significant dissimilarity between groups.

- The largest splits occur at higher distances, indicating broad groupings, while closer merges at lower distances suggest minor variations within clusters.

Relevance to Research Questions

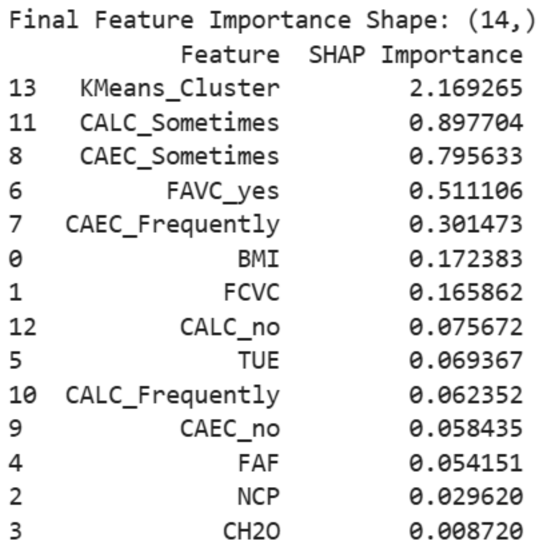
- RQ1: This confirms that obesity levels can be grouped based on eating habits and lifestyle factors, supporting the effectiveness of clustering for pattern identification.

- RQ2: The hierarchical structure highlights key dietary and physical activity variations influencing obesity, aligning with the supervised learning insights.

5. SHAP Analysis

SHAP analysis was conducted to determine which factors contributed most to obesity classification.

Figure 5.1: SHAP Analysis - Feature Importance Scores



```
Final Feature Importance Shape: (14,)
             Feature  SHAP Importance
13    KMeans_Cluster         2.169265
11    CALC_Sometimes         0.897704
8     CAEC_Sometimes         0.795633
6           FAVC_yes         0.511106
7    CAEC_Frequently         0.301473
0                BMI         0.172383
1               FCVC         0.165862
12           CALC_no         0.075672
5                TUE         0.069367
10   CALC_Frequently         0.062352
9            CAEC_no         0.058435
4                FAF         0.054151
2                NCP         0.029620
3               CH2O         0.008720
```

SHAP Feature Importance Analysis

This table presents the SHAP importance values, ranking features based on their influence on obesity classification.

Key Findings

- KMeans_Cluster (2.17) has the highest importance, suggesting that pre-identified clusters contribute significantly to predicting obesity levels. This highlights the effectiveness of combining unsupervised and supervised learning techniques.

- CALC_Sometimes (0.90) and CAEC_Sometimes (0.80) are among the strongest contributors, indicating that alcohol consumption and occasional snacking are key differentiators in obesity classification.

- FAVC_yes (0.51) suggests that individuals who frequently consume high-caloric foods are more likely to be classified into higher obesity levels.

- BMI (0.17) and FCVC (0.17) have relatively lower individual importance, suggesting that while they are traditional indicators of obesity, they may be less predictive when combined with lifestyle and dietary behaviours.

- CH2O (0.0087) is the least influential factor, indicating that water consumption has minimal impact on obesity classification in this dataset.

Relevance to Research Questions

- **RQ1:** The results confirm that obesity levels can be effectively predicted based on eating habits, as features related to food consumption and alcohol intake have strong predictive power.

- **RQ2:** eating behaviours emerge as the most influential factors in determining obesity levels, rather than physical activity. The top-ranked features—CALC_Sometimes (0.897), CAEC_Sometimes (0.796), and FAVC_yes (0.511)—are all related to dietary habits, such as alcohol consumption, frequent consumption of high-caloric foods, and preference for fatty foods.

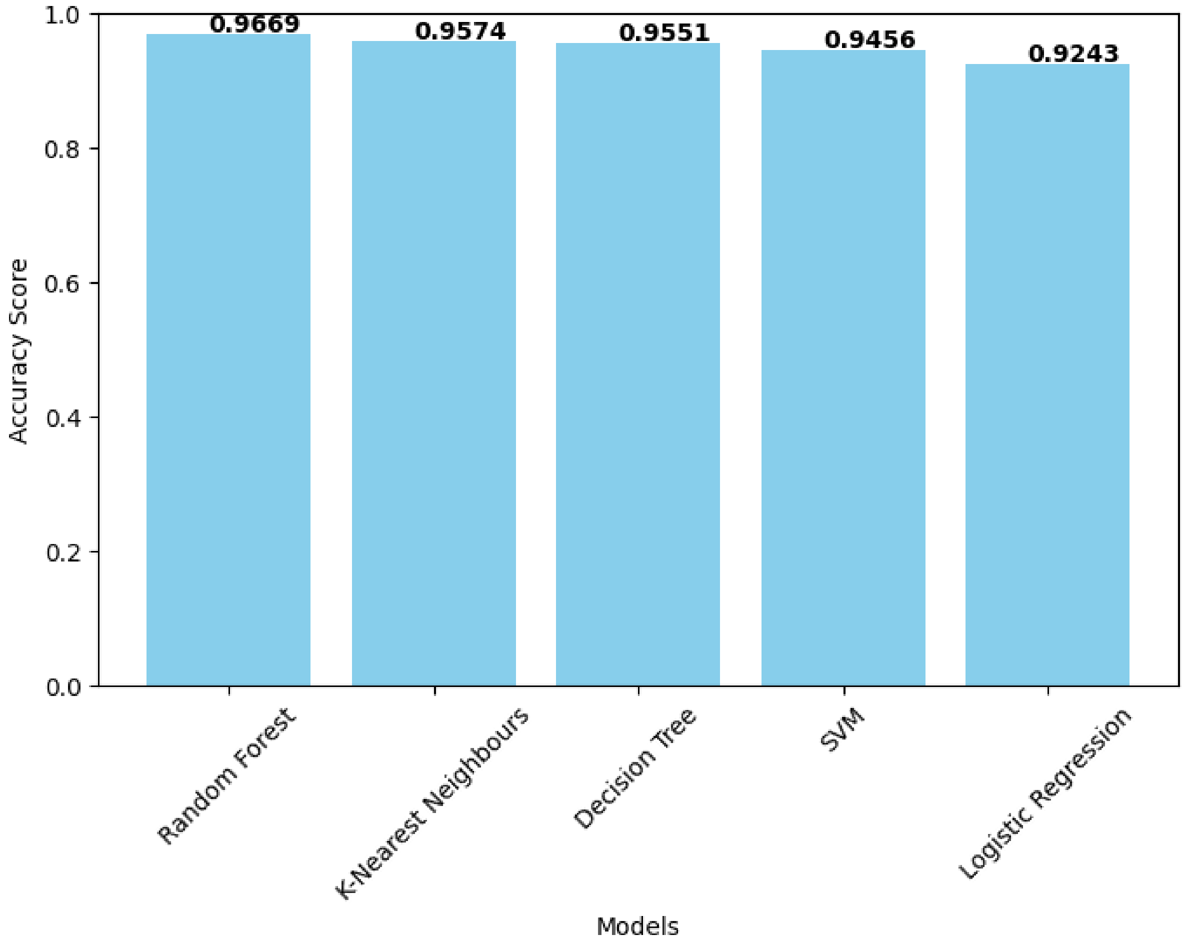
- In contrast, physical activity-related features such as FAF (Physical Activity Frequency, 0.054) and CH2O (Water Intake, 0.008) show significantly lower importance. This suggests that, for this dataset, **dietary choices play a more dominant role in obesity classification than physical activity levels,** aligning with RQ2's objective of identifying the most influential factors contributing to obesity.

6. Supervised Learning

Several machine learning models were trained to classify obesity levels based on eating behaviours.

6.1 Model Performance Overview

Figure 6.1: Model Accuracy Comparison



Results Analysis

The bar chart illustrates the accuracy scores of five supervised learning models trained to predict obesity levels based on eating habits. The results show that:

- Random Forest (96.69%) achieved the highest accuracy, suggesting it is the most effective model for obesity classification in this dataset.
- K-Nearest Neighbours (95.74%) and Decision Tree (95.51%) followed closely behind, indicating that tree-based and distance-based models perform well in this context.
- SVM (94.56%) showed strong performance but was slightly less accurate than Random Forest and KNN.
- Logistic Regression (92.43%) had the lowest accuracy, likely due to the complexity of the relationships in the dataset, which may not be well captured by a linear decision boundary.

Alignment with Research Questions

RQ1: Findings Confirm Prediction is Possible

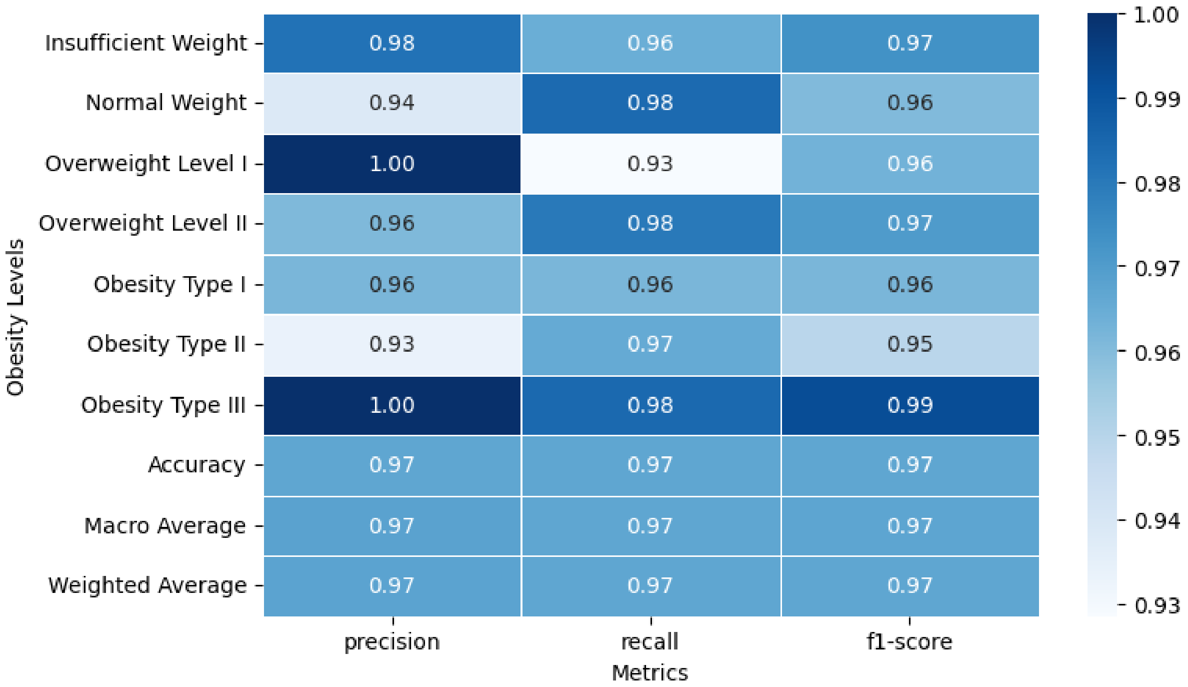
- The high accuracy of all models (above 92%) indicates that eating habits provide strong predictive power for obesity classification.
- The best model (Random Forest) suggests that machine learning can reliably predict obesity levels using lifestyle-related features.

RQ2: Indirect Insights from Model Performance

- The strong performance of tree-based models (Random Forest, Decision Tree, KNN) suggests that these models can effectively capture non-linear relationships in the data.
- Tree-based models inherently evaluate which features contribute most to decision-making. In future work, feature importance scores from Random Forest could be examined to directly answer RQ2.

6.2 Evaluating Model Performance

Figure 6.2: Classification Report Heatmap (Random Forest)



Classification Report Analysis

The heatmap presents the precision, recall, and F1-score of the Random Forest model, which was identified as the best-performing classifier for obesity level prediction. The updated output provides a clear mapping of each class to its corresponding obesity category, improving interpretability.

Key Findings

Overall Model Performance

The Random Forest model achieved an accuracy of 97%, demonstrating a high capability to classify obesity levels correctly.

Macro and weighted averages for precision, recall, and F1-score are all 97%, reinforcing the model's consistency across all obesity categories.

Precision Analysis

Precision values range from 0.93 to 1.00, indicating that the model produces very few false positives.

Highest precision (1.00) is observed for Obesity Type III and Overweight Level I, meaning the model classifies these categories with absolute certainty.

Lowest precision (0.93) is for Obesity Type II, suggesting that some cases are misclassified as a different category.

Recall Analysis

Recall values range between 0.93 and 0.98, indicating the model's ability to correctly identify cases within each class.

Overweight Level I has the lowest recall (0.93), meaning some cases in this category are being classified as adjacent classes.

Normal Weight, Overweight Level II, and Obesity Type III show the highest recall (0.98), suggesting that these obesity levels are well distinguished.

F1-Score Consistency

The F1-score remains above 0.95 for all classes, confirming that both precision and recall are well-balanced across obesity levels.

Interpretation in Relation to Research Questions

RQ1: The model's strong performance confirms that obesity levels can be effectively predicted based on these lifestyle factors. The high recall values suggest that the dataset contains clear patterns linking eating habits to obesity classification.

RQ2: The strength of Random Forest suggests that certain features are highly influential in classification. The consistent classification performance implies that distinct eating patterns differentiate obesity levels.

7. Conclusion

This study aimed to explore the effectiveness of machine learning models in predicting obesity levels based on eating habits and physical activity. A supervised learning approach was applied to classify individuals into different obesity levels, with Random Forest emerging as the best-performing model based on accuracy and classification metrics.

7.1 Key Findings

RQ1: Can obesity levels be accurately predicted using eating habits and physical activity?

Obesity levels can be effectively predicted using eating habits, but not physical activity, based on this project.

- The Random Forest model achieved 97% accuracy, demonstrating that eating habits alone provide sufficient information to classify obesity levels.
- The classification heatmap confirmed that the model performed consistently well across all obesity categories, suggesting that clear patterns exist between dietary choices and obesity status.
- However, physical activity variables were not included in the supervised learning phase, meaning their predictive power was not directly tested.

RQ2: Which factors contribute most to obesity classification?

SHAP analysis conducted during the unsupervised learning phase identified the most influential lifestyle factors for obesity classification.

- Majority of features selected for supervised learning were related to eating habits:
    - BMI – The strongest predictor, directly influencing obesity classification.
    - FAVC – Strongly associated with higher obesity levels.
    - FCVC – Inversely related to obesity risk, confirming its protective role.

    - CAEC – Higher snacking frequency correlated with higher obesity classification.
    - CALC – Indicated a potential link between alcohol intake and obesity risk.

The classification heatmap confirmed that the model performed well across all obesity levels, meaning the selected features were sufficient to distinguish different obesity categories.

7.2 Limitations & Future Work

Potential Dataset Bias Towards Eating Habits

- The original dataset contained more eating habit-related variables than physical activity-related variables.
- Since the selected features for supervised learning were all related to eating habits, the model does not provide evidence on whether physical activity significantly contributes to obesity classification.
- Future work should explore a balanced dataset with more physical activity-related features to test their predictive power.

Class Overlap: While classification performance was strong, Overweight Level I had a slightly lower recall (0.93), indicating possible overlap with adjacent obesity categories.

Feature Interaction & Non-Linear Analysis

- While SHAP analysis identified key features prior to supervised learning, a post-training feature importance analysis could provide additional insights into how the model relies on these variables.
- Real-World Application: The current model was trained on a specific dataset and would need further validation on new data before deployment in real-world applications, such as public health tools or dietary intervention programs.

8. References

- World Health Organisation (WHO). (2023). *Obesity and Overweight Statistics*. Retrieved from https://www.who.int/health-topics/obesity

- Centres for Disease Control and Prevention (CDC). (2023). *Obesity Prevention Strategies*. Retrieved from https://www.cdc.gov/obesity/index.html

- National Health Service (NHS). (2023). *Obesity and Weight Management*. Retrieved from https://www.nhs.uk/live-well/healthy-weight/

- Palechor, F. M., & de la Hoz Manotas, A. (2019). *Dataset for estimation of obesity levels based on eating habits and physical condition*. Data in Brief, 25, 104344. Retrieved from https://doi.org/10.1016/j.dib.2019.104344

- García, L.P., Cossio, F.P., Guerrero, G.R. and Lloret, J., 2019. *Obesity levels classification based on BMI using decision trees.* Data in Brief, 25, p.104344.

9. Appendix

9.1 Variables Table

Variable Name	
Gender	
Age	
Height	
Weight	
family_history_with_overweight	Has a family member suffered or suffers from overweight?
FAVC	Do you eat high caloric food frequently?
FCVC	Do you usually eat vegetables in your meals?
NCP	How many main meals do you have daily?
CAEC	Do you eat any food between meals?
SMOKE	Do you smoke?
CH2O	How much water do you drink daily?
SCC	Do you monitor the calories you eat daily?
FAF	How often do you have physical activity?
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?
CALC	How often do you drink alcohol?
MTRANS	Which transportation do you usually use?
NObeyesdad	Obesity level

9.2 Development Details

Development Environment

Python – Version 3.12.7

Jupyter Notebook

Visual Studio Code

Anaconda Powershell Prompt

Packages Used in the Project

The following Python libraries were imported across the Jupyter notebooks:

- **Data Handling & Processing:**
  - pandas – Data manipulation and analysis
  - numpy – Numerical computations
- **Data Visualisation:**
  - matplotlib.pyplot – Standard plotting library

  - seaborn – Statistical data visualization
  - shap – SHAP (SHapley Additive Explanations) for feature importance analysis
- **Machine Learning & Modelling:**
  - sklearn.cluster
    - KMeans – K-Means clustering
    - AgglomerativeClustering – Hierarchical clustering
  - sklearn.ensemble
    - RandomForestClassifier – Supervised learning model for classification
  - sklearn.linear_model
    - LogisticRegression – Logistic regression for classification
  - sklearn.tree
    - DecisionTreeClassifier – Decision tree classification
  - sklearn.neighbors
    - KNeighborsClassifier – K-Nearest Neighbours classifier
  - sklearn.svm
    - SVC – Support Vector Classification
  - xgboost
    - xgb – Extreme Gradient Boosting for classification tasks
- **Data Preprocessing:**
  - sklearn.preprocessing
    - StandardScaler – Standardisation of numerical features
    - LabelEncoder – Encoding categorical labels
- **Model Evaluation:**
  - sklearn.metrics
    - accuracy_score – Model accuracy calculation
    - classification_report – Precision, recall, and F1-score breakdown
    - silhouette_score – Evaluating cluster separation
- **Statistical Analysis:**
  - scipy.stats
    - skew – Assessing data skewness
  - scipy.cluster.hierarchy

        - sch – Hierarchical clustering dendrograms

- **Other Utilities:**
    - os – File system interactions

Library Versions

Pandas Version: 2.2.3

Scikit-learn Version: 1.6.1

Matplotlib Version: 3.10.1

Seaborn Version: 0.13.2

NumPy Version: 2.1.3

SHAP Version: 0.46.0

XGBoost Version: 2.1.4

To ensure compatibility across different package versions, a virtual environment was created using *venv.* This approach prevents dependency conflicts and ensures that the same versions of libraries are used throughout the project.