**Student ID Number:** _____

# Assignment 1

# Data Exploration and Classification

**Semester 1 2024**

**Student Name: Thuan Nguyen**
**Student ID: 23194073**
**PAPER NAME:** Foundations of Data Science

**PAPER CODE:** COMP615

**Due Date:** Sunday 14 April 2024 (midnight)

**TOTAL MARKS:** 100

**INSTRUCTIONS:**

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section**.

## Table of Contents



# Table of figures

# Task 1: INTRODUCTION

The global obesity epidemic continues to escalate, posing significant health risks and economic burdens across the population. There are some key facts from WHO ("Obesity and overweight," 2024):

,

- In 2022, 1in 8 people in the world were living with obesity.
- Worldwide adult obesity has more than doubled since 1990, and adolescent obesity has quadrupled.
- In 2022, 2.5 billion adult (18 years old and older) were overweight.

This project uses 'Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico' which includes demographic details (age, height, weight), dietary habits, physical activity levels, sedentary behaviour, and obesity level to investigate the intricate relationship between various lifestyles factors and obesity level.

Because the aetiology of obesity is complicated and involves behavioural, genetic, and environmental variables, necessitates a multifaceted approach to understanding. However, this project primarily aims to investigate how different lifestyles choice contributes to varying obesity level among individuals.
Based on the aim of the project, there are one question that need to be answer:

- What dietary habits and daily living habits are significantly affected obesity levels?

Assumption for the research:

- Data completeness: The dataset is assumed to be complete with no missing value in any of the critical fields used for analysis.
- Normality: For certain statistical analyses, it may be assumed that the continuous variables (age, weight, and height) follow a normal distribution.
- Accurate reporting: it is assumed that all participants accurately reported their dietary intake and daily living habits.
- Independence of observation: observation should be independent within the dataset, which means each entry value assumed to represent and independent individual.

# Task 2: DATA EXPLORATION

Let explore the dataset to understand completely about each attribute. Because this research is about dietary and lifestyles habits, so features related to that purposed will be focused on.

### Part 1: Initial Information about dataset

Below are questions of the survey used for initial recollection of information: (Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico). This table is necessary to understand deeply about each feature, how they are collected, what question it answer, which can be used for further analysis or decision making.

## Variables Table

| Variable Name | Role | Type | Demographic | Description | Units | Missing Values |
|---|---|---|---|---|---|---|
| Gender | Feature | Categorical | Gender | | | no |
| Age | Feature | Continuous | Age | | | no |
| Height | Feature | Continuous | | | | no |
| Weight | Feature | Continuous | | | | no |
| family_history_with_overweight | Feature | Binary | | Has a family member suffered or suffers from overweight? | | no |
| FAVC | Feature | Binary | | Do you eat high caloric food frequently? | | no |
| FCVC | Feature | Integer | | Do you usually eat vegetables in your meals? | | no |
| NCP | Feature | Continuous | | How many main meals do you have daily? | | no |

*Figure 1: Dataset information.*

| | | | | | |
|---|---|---|---|---|---|
| CAEC | Feature | Categorical | | Do you eat any food between meals? | no |
| SMOKE | Feature | Binary | | Do you smoke? | no |
| CH2O | Feature | Continuous | | How much water do you drink daily? | no |
| SCC | Feature | Binary | | Do you monitor the calories you eat daily? | no |
| FAF | Feature | Continuous | | How often do you have physical activity? | no |
| TUE | Feature | Integer | | How much time do you use technological devices such as cell phone, videogames, television, computer and others? | no |
| CALC | Feature | Categorical | | How often do you drink alcohol? | no |
| MTRANS | Feature | Categorical | | Which transportation do you usually use? | no |
| NObeyesdad | Target | Categorical | | Obesity level | no |

*Figure 2: Dataset information.*

Below are questions that used for collecting data and its possible answer.

| Questions | Possible Answers |
|---|---|
| ¿What is your gender? | • Female<br>• Male |
| ¿what is your age? | Numeric value |
| ¿what is your height? | Numeric value in meters |
| ¿what is your weight? | Numeric value in kilograms |
| ¿Has a family member suffered or suffers from overweight? | • Yes<br>• No |
| ¿Do you eat high caloric food frequently? | • Yes<br>• No |
| ¿Do you usually eat vegetables in your meals? | • Never<br>• Sometimes<br>• Always |

*Figure 3: Dataset information.*

| | |
|---|---|
| ¿How many main meals do you have daily? | • Between 1 y 2<br>• Three<br>• More than three |
| ¿Do you eat any food between meals? | • No<br>• Sometimes<br>• Frequently<br>• Always |
| ¿Do you smoke? | • Yes<br>• No |
| ¿How much water do you drink daily? | • Less than a liter<br>• Between 1 and 2L<br>• More than 2L |

*Figure 4: Dataset information.*

| | |
|---|---|
| ¿Do you monitor the calories you eat daily? | • Yes<br>• No |
| ¿How often do you have physical activity? | • I do not have<br>• 1 or 2 days<br>• 2 or 4 days<br>• 4 or 5 days |
| ¿How much time do you use technological devices such as cell phone, videogames, television, computer and others? | • 0–2 hours<br>• 3–5 hours<br>• More than 5 hours |
| ¿how often do you drink alcohol? | • I do not drink<br>• Sometimes<br>• Frequently<br>• Always |

*Figure 5: Dataset information.*

| | |
|---|---|
| ¿Which transportation do you usually use? | • Automobile<br>• Motorbike<br>• Bike<br>• Public Transportation<br>• Walking |

*Figure 6: Dataset information.*

From the figure 3, 4,5 and 6, we can see that the author of the dataset convert answer to numerical discrete data. For example: the answer to question 'Do you usually eat vegetables in your meals?', the answer converts to 0(Never), 1(Sometimes), and 2(Always). This also applies to other questions but not all.

Part 2: Original features, instance datatypes, summary statistics

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
None
```

*Figure 7: Dataset Information*

From figure 1, the dataset contains 2111 entries and 17 features (columns). All the entries are non-null, which is indicating no missing values across the dataset.
There are two datatypes for all features, 9 features are categorical (object types) and 8 are numerical (float64).

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 21.0 | 1.62 | 64.0 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 0.0 | 1.0 | no | Public_Transportation | Normal_Weight |
| 1 | Female | 21.0 | 1.52 | 56.0 | yes | no | 3.0 | 3.0 | Sometimes | yes | 3.0 | yes | 3.0 | 0.0 | Sometimes | Public_Transportation | Normal_Weight |
| 2 | Male | 23.0 | 1.80 | 77.0 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 2.0 | 1.0 | Frequently | Public_Transportation | Normal_Weight |
| 3 | Male | 27.0 | 1.80 | 87.0 | no | no | 3.0 | 3.0 | Sometimes | no | 2.0 | no | 2.0 | 0.0 | Frequently | Walking | Overweight_Level_I |
| 4 | Male | 22.0 | 1.78 | 89.8 | no | no | 2.0 | 1.0 | Sometimes | no | 2.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_II |

*Figure 8: Dataset first few rows.*

As we can see above (figure 8) , there are 16 attributes, demographic details (Age, Height, Weight, family_history_with_overweight), dietary habits( FAVC, FCVC, NCP,…) and lifestyle habits ( FAF, TUE,…) which can be considered as factors affected the obesity levels can be considered as predictor attributes, and 'Nobeyesdad' is considered as target that we want to research.

```
Summary Statistics for numberical features:
                Age       Height       Weight         FCVC          NCP  \
count  2111.000000  2111.000000  2111.000000  2111.000000  2111.000000
mean     24.312600     1.701677    86.586058     2.419043     2.685628
std       6.345968     0.093305    26.191172     0.533927     0.778039
min      14.000000     1.450000    39.000000     1.000000     1.000000
25%      19.947192     1.630000    65.473343     2.000000     2.658738
50%      22.777890     1.700499    83.000000     2.385502     3.000000
75%      26.000000     1.768464   107.430682     3.000000     3.000000
max      61.000000     1.980000   173.000000     3.000000     4.000000

              CH2O          FAF          TUE
count  2111.000000  2111.000000  2111.000000
mean      2.008011     1.010298     0.657866
std       0.612953     0.850592     0.608927
min       1.000000     0.000000     0.000000
25%       1.584812     0.124505     0.000000
50%       2.000000     1.000000     0.625350
75%       2.477420     1.666678     1.000000
max       3.000000     3.000000     2.000000

--------------
```

*Figure 9: Summary Statistics for numerical features.*

From figure 9:

- The average FCVC score of 2.42 indicates that people eat vegetables on a reasonably frequent basis in general. This component may operate as a barrier against obesity as it is frequently associated with diets that are healthier.
- The average of almost three meals a day suggests that most individuals follow a regular eating schedule.
- Participants are generally well-hydrated, consuming 2.01 litres of water on average per day. This can impact appetite and metabolism, two critical aspects of body weight management.

Part 3: Checking Dataset cleanliness.

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | \ |
|---|---|---|---|---|---|---|---|---|
| 97 | Female | 21.0 | 1.52 | 42.0 | no | no | 3.0 | |
| 98 | Female | 21.0 | 1.52 | 42.0 | no | no | 3.0 | |
| 105 | Female | 25.0 | 1.57 | 55.0 | no | yes | 2.0 | |
| 106 | Female | 25.0 | 1.57 | 55.0 | no | yes | 2.0 | |
| 145 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 174 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 179 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 184 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 208 | Female | 22.0 | 1.69 | 65.0 | yes | yes | 2.0 | |
| 209 | Female | 22.0 | 1.69 | 65.0 | yes | yes | 2.0 | |
| 282 | Female | 18.0 | 1.62 | 55.0 | yes | yes | 2.0 | |
| 295 | Female | 16.0 | 1.66 | 58.0 | no | no | 2.0 | |
| 309 | Female | 16.0 | 1.66 | 58.0 | no | no | 2.0 | |
| 443 | Male | 18.0 | 1.72 | 53.0 | yes | yes | 2.0 | |
| 460 | Female | 18.0 | 1.62 | 55.0 | yes | yes | 2.0 | |
| 466 | Male | 22.0 | 1.74 | 75.0 | yes | yes | 3.0 | |
| 467 | Male | 22.0 | 1.74 | 75.0 | yes | yes | 3.0 | |
| 496 | Male | 18.0 | 1.72 | 53.0 | yes | yes | 2.0 | |
| 523 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | |
| 527 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | |
| 659 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | |
| 663 | Female | 21.0 | 1.52 | 42.0 | no | yes | 3.0 | |
| 763 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 764 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 824 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 830 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 831 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 832 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 833 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 834 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 921 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 922 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |
| 923 | Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | |

*Figure 10: duplicated rows.*

| | MTRANS | NObeyesdad |
|---|---|---|
| 97 | Public_Transportation | Insufficient_Weight |
| 98 | Public_Transportation | Insufficient_Weight |
| 105 | Public_Transportation | Normal_Weight |
| 106 | Public_Transportation | Normal_Weight |
| 145 | Public_Transportation | Overweight_Level_I |
| 174 | Public_Transportation | Overweight_Level_I |
| 179 | Public_Transportation | Overweight_Level_I |
| 184 | Public_Transportation | Overweight_Level_I |
| 208 | Public_Transportation | Normal_Weight |
| 209 | Public_Transportation | Normal_Weight |
| 282 | Public_Transportation | Normal_Weight |
| 295 | Walking | Normal_Weight |
| 309 | Walking | Normal_Weight |
| 443 | Public_Transportation | Insufficient_Weight |
| 460 | Public_Transportation | Normal_Weight |
| 466 | Automobile | Normal_Weight |
| 467 | Automobile | Normal_Weight |
| 496 | Public_Transportation | Insufficient_Weight |
| 523 | Public_Transportation | Insufficient_Weight |
| 527 | Public_Transportation | Insufficient_Weight |
| 659 | Public_Transportation | Insufficient_Weight |
| 663 | Public_Transportation | Insufficient_Weight |
| 763 | Public_Transportation | Overweight_Level_I |
| 764 | Public_Transportation | Overweight_Level_I |
| 824 | Public_Transportation | Overweight_Level_I |
| 830 | Public_Transportation | Overweight_Level_I |
| 831 | Public_Transportation | Overweight_Level_I |
| 832 | Public_Transportation | Overweight_Level_I |
| 833 | Public_Transportation | Overweight_Level_I |
| 834 | Public_Transportation | Overweight_Level_I |
| 921 | Public_Transportation | Overweight_Level_I |
| 922 | Public_Transportation | Overweight_Level_I |
| 923 | Public_Transportation | Overweight_Level_I |

*Figure 11: duplicated rows.*

| | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | \ |
|---|---|---|---|---|---|---|---|---|---|
| 97 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 98 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 105 | 1.0 | Sometimes | no | 2.0 | no | 2.0 | 0.0 | Sometimes | |
| 106 | 1.0 | Sometimes | no | 2.0 | no | 2.0 | 0.0 | Sometimes | |
| 145 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 174 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 179 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 184 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 208 | 3.0 | Sometimes | no | 2.0 | no | 1.0 | 1.0 | Sometimes | |
| 209 | 3.0 | Sometimes | no | 2.0 | no | 1.0 | 1.0 | Sometimes | |
| 282 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 1.0 | no | |
| 295 | 1.0 | Sometimes | no | 1.0 | no | 0.0 | 1.0 | no | |
| 309 | 1.0 | Sometimes | no | 1.0 | no | 0.0 | 1.0 | no | |
| 443 | 3.0 | Sometimes | no | 2.0 | no | 0.0 | 2.0 | Sometimes | |
| 460 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 1.0 | no | |
| 466 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 0.0 | no | |
| 467 | 3.0 | Frequently | no | 1.0 | no | 1.0 | 0.0 | no | |
| 496 | 3.0 | Sometimes | no | 2.0 | no | 0.0 | 2.0 | Sometimes | |
| 523 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 527 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 659 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 663 | 1.0 | Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | |
| 763 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 764 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 824 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 830 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 831 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 832 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 833 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 834 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 921 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 922 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |
| 923 | 1.0 | no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | |

*Figure 12: duplicated rows.*

As can see from figure 10,11, and 12, there are plenty of duplicated rows in this dataset, which can skew analysis result, especially for visualization distribution. In another hand, for this circumstance, the number of repeated rows is not too much, compared to whole dataset. It might be caused by computing mistake or noticed one value many times. Because there is no indication that the duplicates are intentional or represent different observations, the best way to handle is removing them.

visualization method to detect outliers

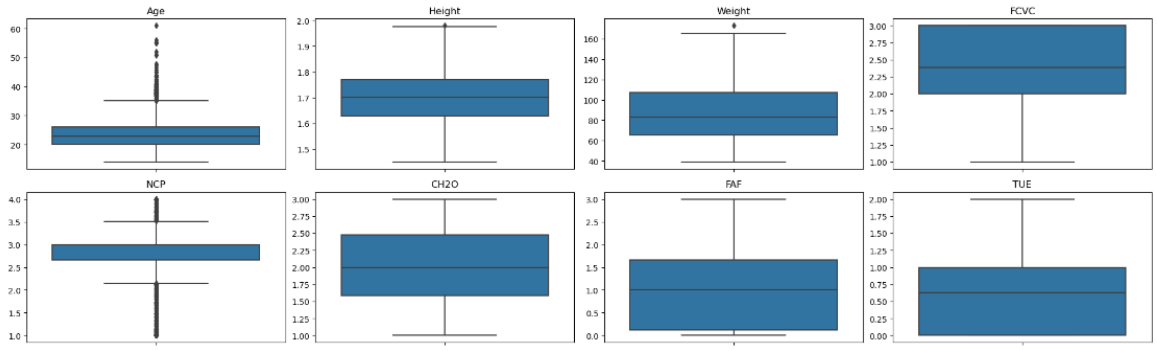



*Figure 13: Visualization method for detecting outliers.*

From figure 13, easily observe that attributes contain potential outliers are Age, Height, Weight, and NCP. Additionally, Age and NCP seems to have a lot more noticeable potential outliers compared to Height and Weight.

Furthermore, it is very hard to approach significant potential outliers and to identify the number of those outliers if only using visualization methods. Thus, using IQR method is more prudent to calculate and reinforce affirmation.

```
Amount of outliers by using IQR method:
 Age       168
Height       1
Weight       1
FCVC         0
NCP        579
CH2O         0
FAF          0
TUE          0
dtype: int64
```

*Figure 14: Statistical method for detecting outliers.*

Figure 14 is the result after using the IQR method to calculate. Similar results arise after compare two method. There are four features that contain outliers, whereas the amount of Age and NCP are significant high.

For feature ‘Age’: from information from figure 9, the highest age occurs is 61 and the lowest is 14, which mean even those value considered as outliers from the result of outliers detecting method, those value still make sense and reflect the real input from people who did the survey.
For features ‘NCP’: as mentioned above, the values of this attributes are converted to numerical value, from 1 to 4. Therefore, the values also represent real input from the survey.
Therefore, keep all the outliers because it represents true value.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| emale | 25.19621 | 1.686306 | 104.5727 | yes | yes | 3 | 3 | Sometimes | no | 1.152736 | no | 0.319156 | 1 |
| emale | 18.50334 | 1.683124 | 126.6738 | yes | yes | 3 | 3 | Sometimes | no | 1.115967 | no | 1.541072 | 1 |
| emale | 26 | 1.622397 | 110.7926 | yes | yes | 3 | 3 | Sometimes | no | 2.704507 | no | 0 | 0.29499 |
| emale | 21.85383 | 1.755643 | 137.7969 | yes | yes | 3 | 3 | Sometimes | no | 2.184707 | no | 1.978631 | 0.838957 |
| emale | 21.90012 | 1.843419 | 165.0573 | yes | yes | 3 | 3 | Sometimes | no | 2.406541 | no | 0.10032 | 0.479221 |
| emale | 18.30662 | 1.7456 | 133.0344 | yes | yes | 3 | 3 | Sometimes | no | 2.984323 | no | 1.586525 | 0.62535 |
| emale | 26 | 1.630927 | 111.4855 | yes | yes | 3 | 3 | Sometimes | no | 2.444125 | no | 0 | 0.26579 |
| emale | 26 | 1.629191 | 104.8268 | yes | yes | 3 | 3 | Sometimes | no | 2.654702 | no | 0 | 0.555468 |
| emale | 21.84971 | 1.770612 | 133.9633 | yes | yes | 3 | 3 | Sometimes | no | 2.825629 | no | 1.399183 | 0.928972 |
| ale | 19.79905 | 1.743702 | 54.92753 | yes | yes | 2 | 3.28926 | Sometimes | no | 2.847264 | no | 1.680844 | 2 |
| ale | 17.18875 | 1.771915 | 55.69504 | yes | yes | 2 | 4 | Sometimes | no | 2.884033 | no | 2 | 1.340107 |
| ale | 22.28502 | 1.75376 | 55.87926 | yes | yes | 2.450218 | 3.995147 | Sometimes | no | 2.147746 | no | 2 | 0.58998 |
| emale | 22 | 1.675446 | 51.1542 | yes | yes | 3 | 3 | Frequently | no | 2.815293 | no | 1.978631 | 1 |
| emale | 21.02497 | 1.666203 | 49.86979 | yes | yes | 3 | 3 | Frequently | no | 2.593459 | no | 2 | 1 |
| emale | 22.03833 | 1.711467 | 51.96552 | yes | yes | 2.880161 | 3 | Frequently | no | 1.031354 | no | 2.206738 | 1.37465 |
| emale | 21.24314 | 1.598019 | 44.84566 | no | no | 3 | 1.72626 | Frequently | no | 2.444125 | no | 1.31817 | 0 |
| emale | 22.14243 | 1.59611 | 42.84803 | no | no | 3 | 2.581015 | Frequently | no | 2.654702 | no | 0.902095 | 0 |
| emale | 21.96243 | 1.57206 | 43.91984 | no | no | 3 | 1.600812 | Frequently | no | 2.651258 | no | 0.600817 | 0 |
| emale | 21.49106 | 1.586952 | 43.08751 | no | no | 2.00876 | 1.73762 | Frequently | no | 1.792022 | no | 0.119643 | 0 |
| emale | 22.71794 | 1.59559 | 44.58116 | no | no | 2.596579 | 1.10548 | Frequently | no | 1.490613 | no | 0.345684 | 0 |
| emale | 23.50125 | 1.6 | 45 | no | no | 2.591439 | 3 | Frequently | no | 2.074048 | no | 1.679935 | 0 |
| emale | 18.53508 | 1.688025 | 45 | no | yes | 3 | 3 | Sometimes | no | 3 | yes | 2.539762 | 1.283673 |
| emale | 19 | 1.556211 | 42.33977 | no | yes | 3 | 2.0846 | Sometimes | no | 2.13755 | yes | 0.196152 | 0.062488 |
| emale | 19 | 1.564199 | 42.09606 | no | yes | 3 | 1.894384 | Sometimes | no | 2.456581 | yes | 1.596576 | 0.9974 |
| emale | 20.25453 | 1.56948 | 41.32456 | no | yes | 2.392665 | 1 | Frequently | no | 1 | no | 0 | 0.738269 |

*Figure 15: Dataset shown in Excel.*

From figure 7 above, it is shown that this dataset has no missing value, indicating that authors of this dataset might handled those missing values using imputing method. However, after comparing with values shown in Excel table, those features such as FCVC, NCP, FAF. CH2O, TUE have some continuous values, which is wrong categorised. Because those values generated by converted from categorical survey answer to discrete numerical values, so the values should be discrete values. Assume that the imputing method are done inaccurately, the handle method that suggested to use is round those continuous values to nearest integer values.

## Part 4: Original Data Visualization

```
------ Distribution of Continous Numerical Data:
```

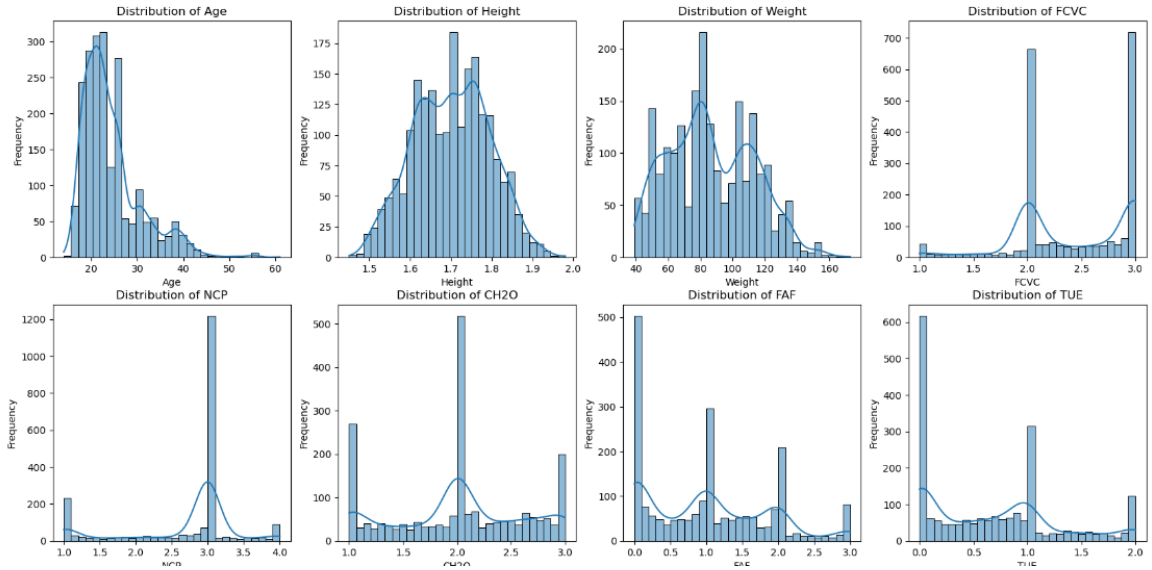



*Figure 16: Distribution for numerical data.*

As mentioned above, noticed that some attributes such as FCVC, NCP, CH2O, FAF, TUE are categorized wrong type. However, those attributes are related to purpose of the research:

- FCVC: This histogram shown that values are around 2 and 3, suggesting that many participants often consume vegetables regularly.
- NCP: significant peak at 3, indicating that most participants stick with 3 meals a day.
- CH2O: From the distribution, the peak is 2, based on possible answer from the survey, most participants consume 1 to 2 L per day.
- FAF: The physical activity frequency is skewed to lower values, with many participants reporting low to no physical activity.
- TUE: The histogram is similar to FAF, skewed to lower values.

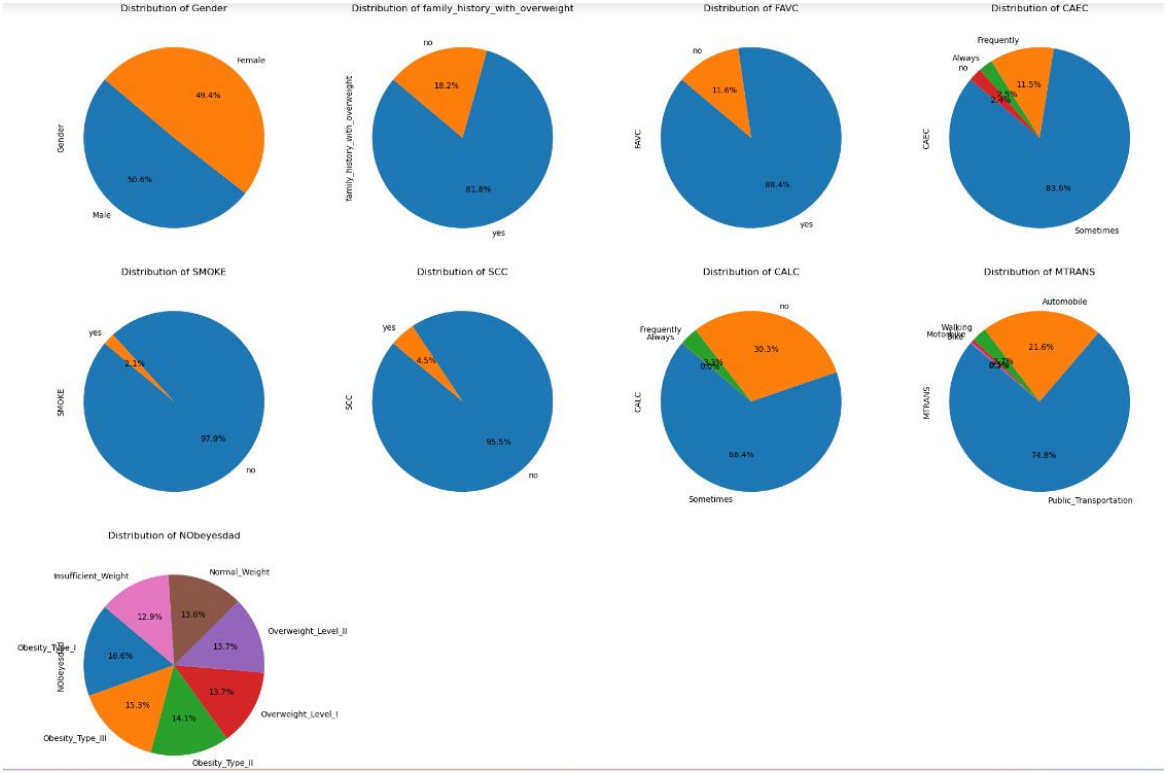



*Figure 17: Visualization for categorical data.*

From figure 17:

- The dataset appears to have a balanced distribution of genders, which could be good for analysis process because it might reduce bias.
- FAVC: Most participants frequently consume high-caloric foods (88.4%).
- CAEC: Most participants sometimes eat between meals (83.6%), with a smaller number frequently (11.5%). Rare cases always (2.5%) or never eat between meals (2.4%).
- SMOKE: Smoking is not prevalent among the participants (90.9%).
- SCC: Most participants do not monitor their calorie intake (90.5%). Lack of monitoring potentially leads to weight gain.
- CALC: Many participants never consume alcohol (66.4%)
- MTRANS: The predominant mode of transportation is public transportation (74.8%), while walking (2.7%), which can be a significant factor in physical activity levels.

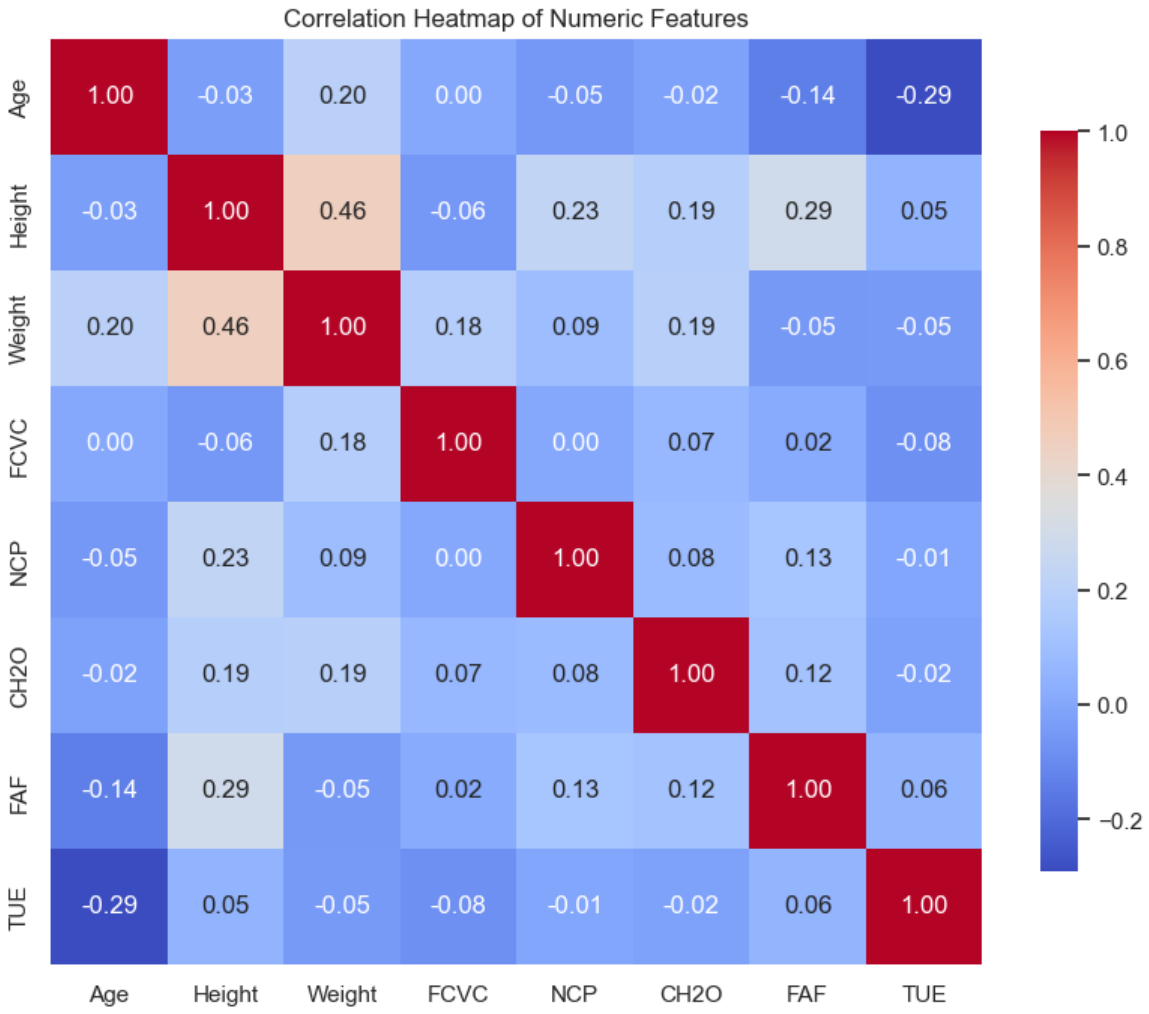



*Figure 18: correlation heatmap for relationship between continuous features.*

From figure 18:

- Weight and Height are moderate positive correlation (0.46), suggesting that taller individuals tend to weight more.
- Weight and age also show weak positive correlation (0.20), indicating that weight tends to increase with age.
- Height and FAF (physical activity frequency) show weak positive correlation (0.29), indicating that individuals who train more could be taller.
- Age and TUE (time using technology) have moderate negative correlation (0.3), indicating that the older spend less time to use technology.

## Part 5: Conclusion

The data and visuals highlight how many different factors contribute to obesity. They stress the significance of lifestyle decisions (food, exercise, mode of transportation) as well as demographic considerations (age, gender) in comprehending and combating obesity. Crucially, the statistics point to specific intervention strategies that might be successful in controlling obesity, like encouraging physical activity and better eating habits. The results further reinforce the necessity for tailored strategies in healthcare and policy development to address obesity, considering the various variables that exist among various populations. These discoveries can direct further investigation, aid in the formulation of public health policies, and educate people about obesity-related health-related issues.

# Task 3: CLASSIFICATION MODELS

PART 1: PREPROCESSING

As showing at Part 3 Task 2, this dataset contains duplicates, outliers, and wrong categorized values. Those noises need to be handled for Classification Models, this is crucial step to ensure that the model is effective, efficient, and reliable.

For duplicates, as discussed above, remove those duplicates is necessary.

```
number of duplicated value after remove: 0
```

*Figure 19: Duplicates after handle.*

For outliers, as discusses above, keep all outliers because those outliers represent valid, real-world information.

For wrong categorized values that might be caused from inaccurately handling technique. Therefore, to fix that, round them to the nearest integer value and convert into integer types is the most feasible option for further analysis and creating model

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2087 non-null   object
 1   Age                             2087 non-null   float64
 2   Height                          2087 non-null   float64
 3   Weight                          2087 non-null   float64
 4   family_history_with_overweight  2087 non-null   object
 5   FAVC                            2087 non-null   object
 6   FCVC                            2087 non-null   int32
 7   NCP                             2087 non-null   int32
 8   CAEC                            2087 non-null   object
 9   SMOKE                           2087 non-null   object
 10  CH2O                            2087 non-null   int32
 11  SCC                             2087 non-null   object
 12  FAF                             2087 non-null   int32
 13  TUE                             2087 non-null   int32
 14  CALC                            2087 non-null   object
 15  MTRANS                          2087 non-null   object
 16  NObeyesdad                      2087 non-null   object
dtypes: float64(3), int32(5), object(9)
memory usage: 252.7+ KB
None
```

*Figure 20: Dataset after handling wrong imputing value.*

Figure 19 shows that those features such as FCVC, NCP, CH2O, FAF, TUE are converted to integer (int32) types, which is believed that more reasonable based on questions survey.

After handling all noises, below are visualizations for clean dataset.

```
------ Distribution of Continous Numerical Data:
```

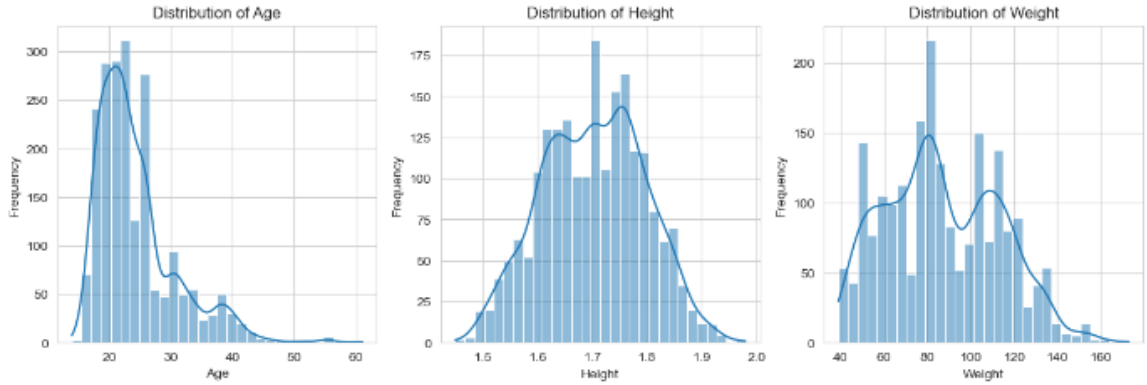



*Figure 21: Distribution for continuous data.*

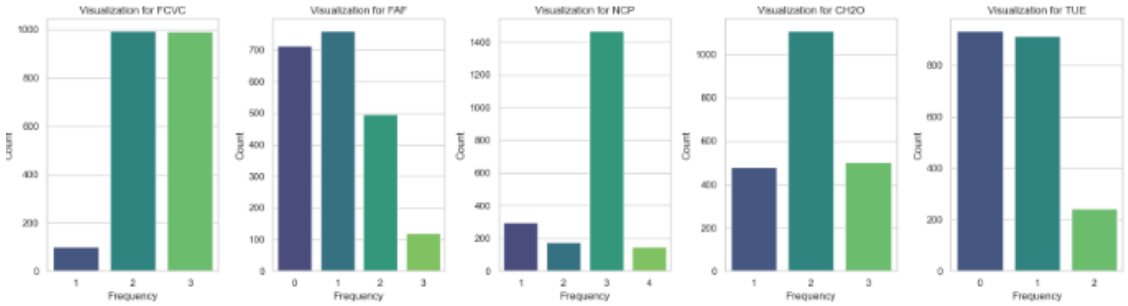



*Figure 22: Distribution for discrete data.*

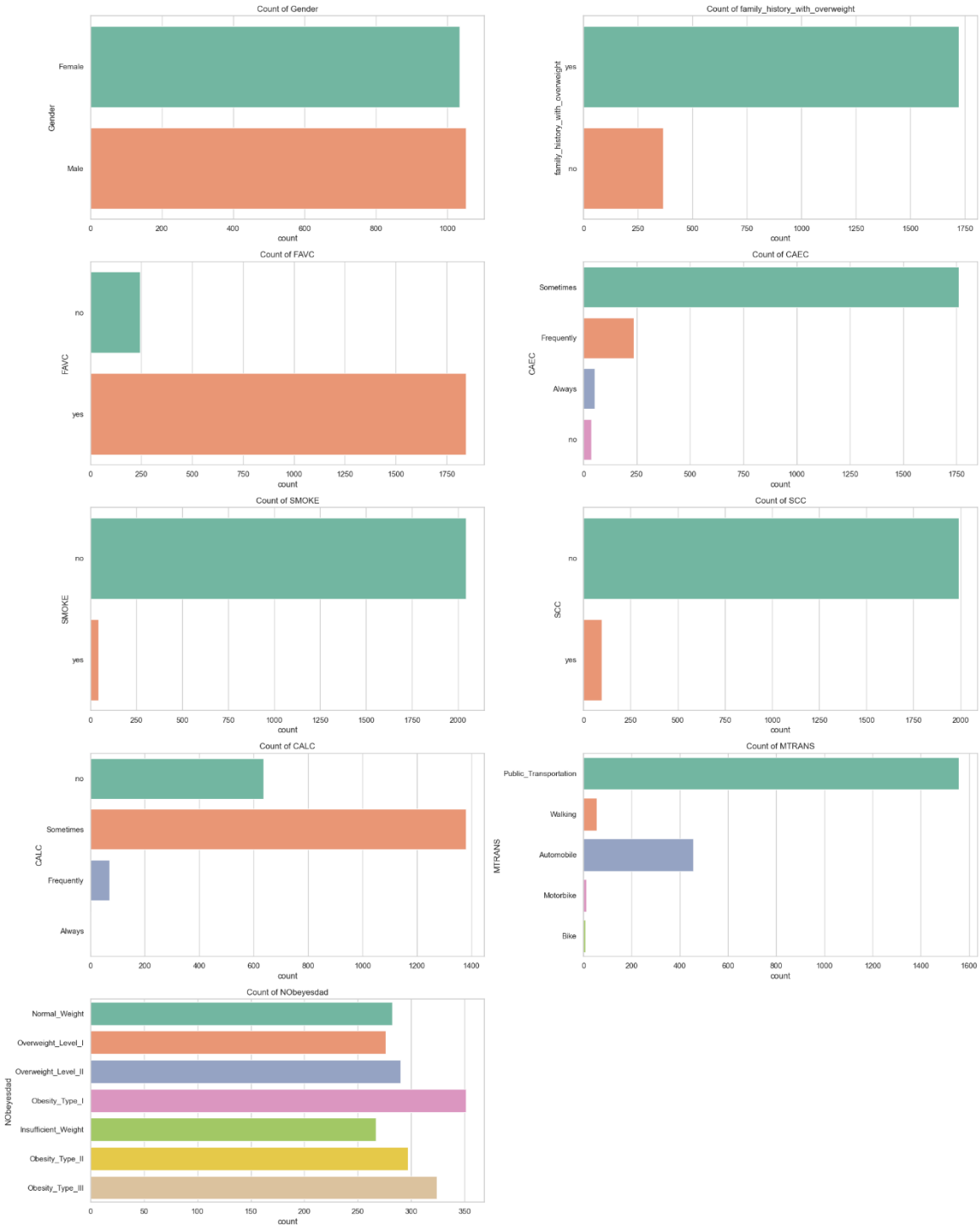
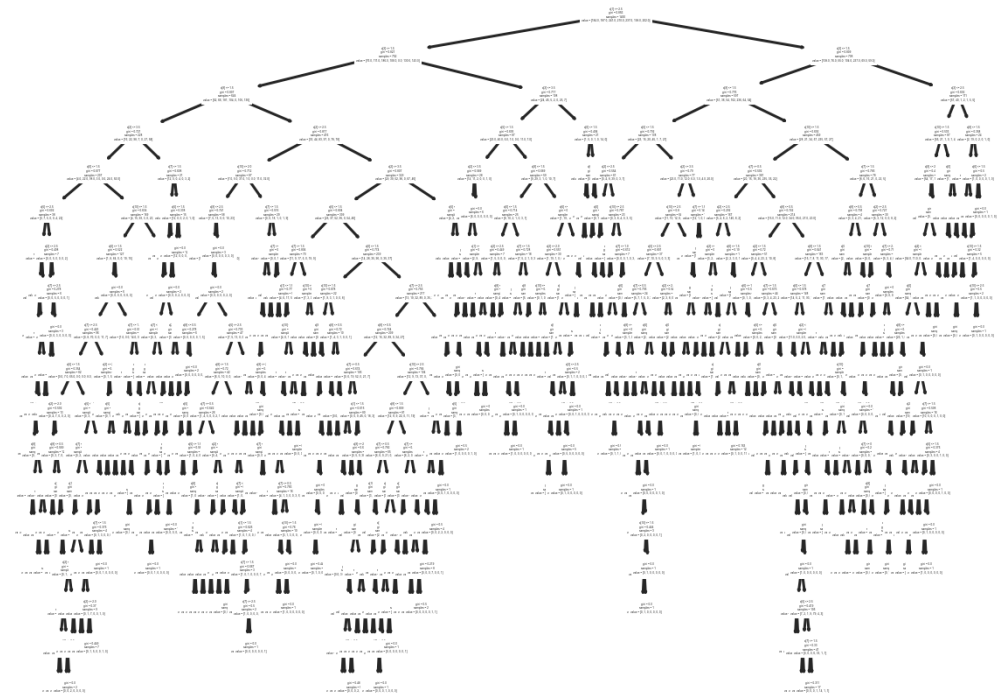



*Figure 23: Distribution for categorical data.*

## PART 2: Create Decision Tree Algorithm.

The purpose of this research is focused on how eating habits and lifestyles habits affect obesity. Therefore, all demographic factors such as Age, Height, weight, and family_history_with_obesity is removed to simplify the features selection process.



```
Number of tree nodes:  895
Model accuracy score with criterion gini index: 0.60
```

*Figure 24: Original Tree*

As can be seen from the figure 24, the original Tree has 895 number of tree nodes and accuracy score is 0.6, which is very complex and very hard to interpret. Therefore, using two parameters which is Max depth and Max leaf nodes to reduce the complexity of the model but keep the accuracy scores at acceptable level.

Below is the plot that visualizes the max depth and max leaf node over accuracy score.

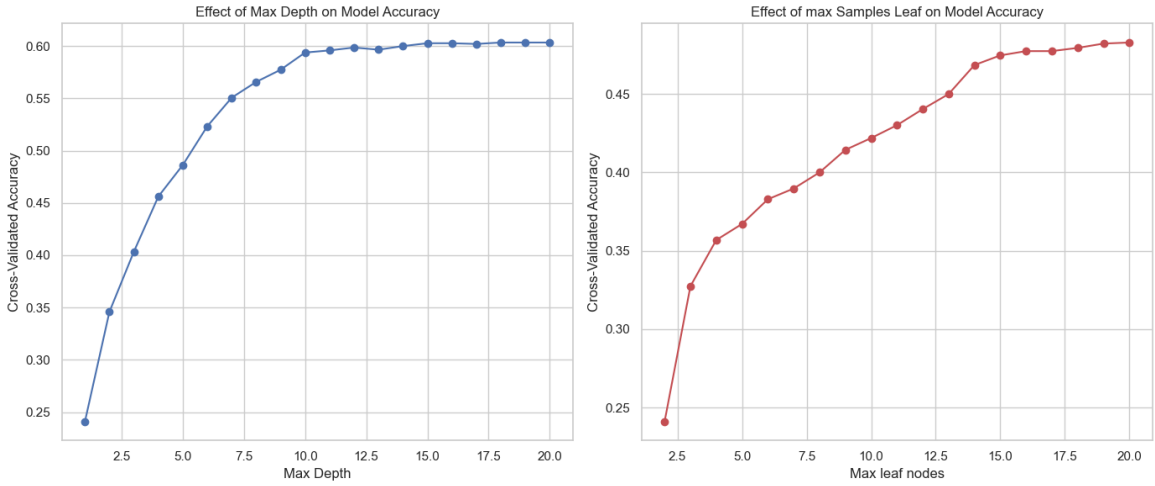



*Figure 25: Max depth and leaf node on Cross-validated Accuracy.*

Effect of max depth:

- As the max_depth of the decision tree increases, the model accuracy initially improves and then stabilizes or slightly decreases. This suggests that beyond a certain depth, the model may start overfitting the training data.

- The optimal max_depth of this model appears to be around 10, where the model achieves the highest cross-validated accuracy. Because beyond 10, the accuracy is stable.

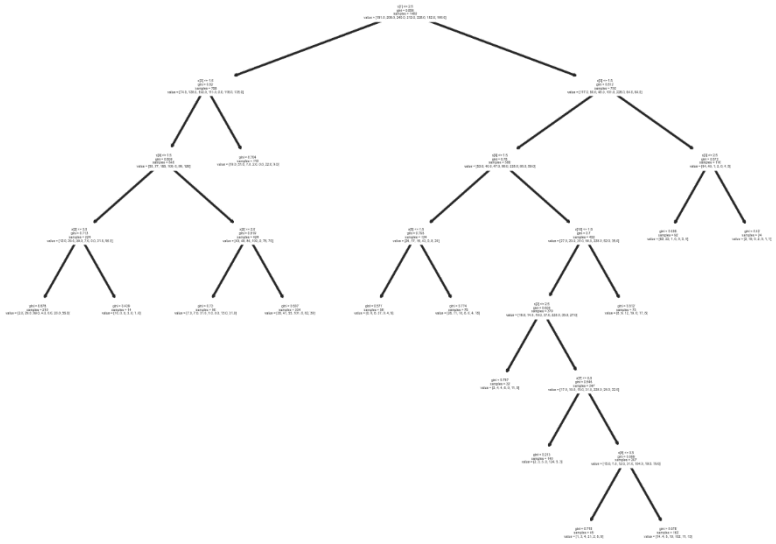
Effect of max leaf_node:

- The accuracy increases significantly with more leaf nodes up until around 10 leaf nodes, after which the increase in accuracy slows down, and the curve starts to plateau around 14 leaf nodes.

## PART 3: FINAL OPTIMISED CLASSIFICATION TREE AND DESCRIBE ITS STRUCTURE

Based on the analysis above, optimal max depth is 10 and optimal max leaf node is 14. Below is final Decision Tree.

Decision tree trained on all the dataset features using max depth=10 and max_leaf_nodes = 14



```
Number of tree nodes:  27
Model accuracy score with criterion gini index: 0.45
```

*Figure 26: Final Decision Tree and its scores.*

complexity and interpretability are effectively addressed by the decision tree model's simplification using max depth and leaf nodes. To better comprehend model decisions and lessen overfitting, the tree's complexity can be reduced (from 895 nodes to 27 nodes) while keeping acceptable accuracy levels (0.45).

## Part 4: Find the features important.

| | Feature | Importance |
|---|---|---|
| 3 | CAEC | 0.258 |
| 1 | FCVC | 0.192 |
| 9 | CALC | 0.171 |
| 2 | NCP | 0.130 |
| 10 | MTRANS | 0.088 |
| 8 | TUE | 0.062 |
| 7 | FAF | 0.050 |
| 5 | CH2O | 0.048 |
| 0 | FAVC | 0.000 |
| 4 | SMOKE | 0.000 |
| 6 | SCC | 0.000 |

*Figure 27: features importance summary.*

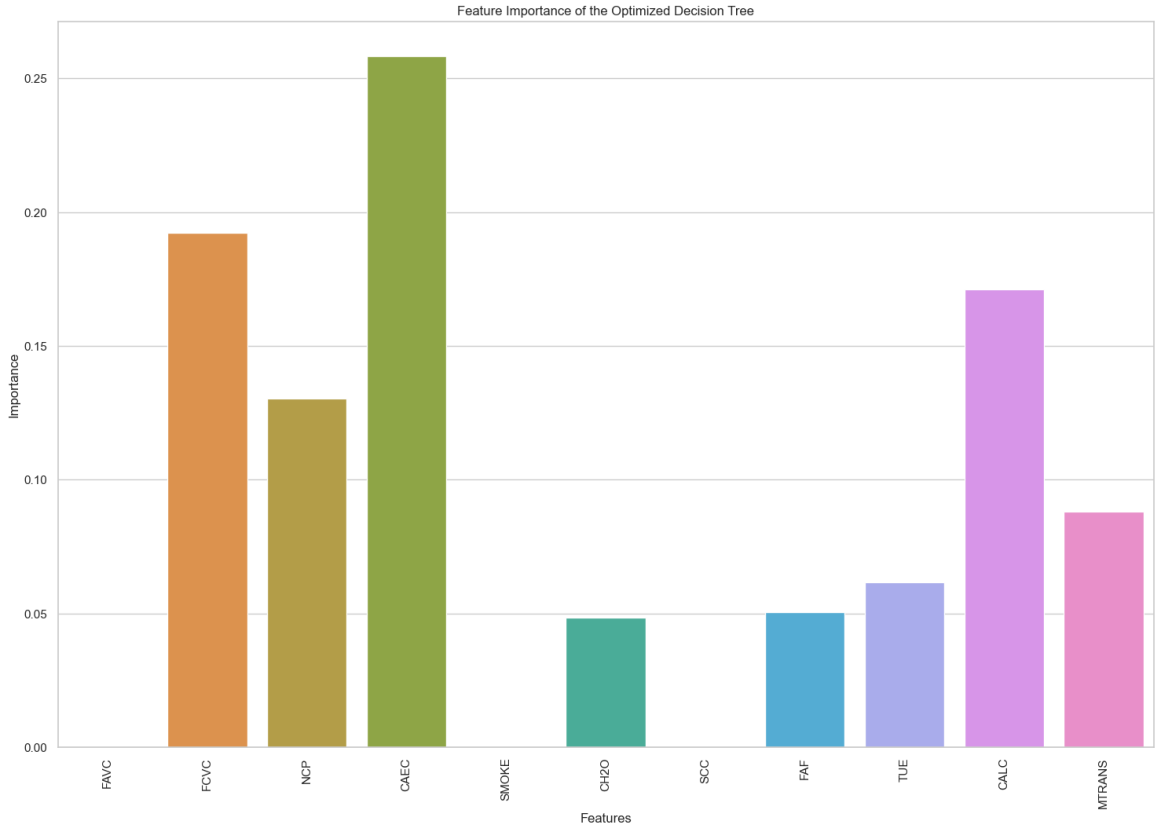



*Figure 28: features importance.*

Based on figure 27 and figure 28, he highest bar indicates that CAEC (over 25) is the most significant trait, with FCVC come at the second (around 0.19) and third is CALC, which approximately 0.17. In contrast, FAVC, SMOKE and SCC is completely 0.

## PART 5: GENERATE AND CAREFULLY EXAMINE THE CONFUSION MATRIX AND EXPLAIN FINDING

According to the figure 28, Obesity_Type_III is the category in which the model performs the best, with an F1-score of 0.90, a recall of 1.00, and a precision of 0.82, indicating good accuracy and completeness. In other word, Obesity_Type_III has very few positives are misclassified as negative and very few negatives misclassified as positives. The Overweight_Level_I class is where the model is least accurate, suggesting that there will be more false positives in this class. The macro and weighted averages for

precision, recall, and F1-score all point to a model's overall accuracy of 0.56, which indicates a modest level of performance overall.

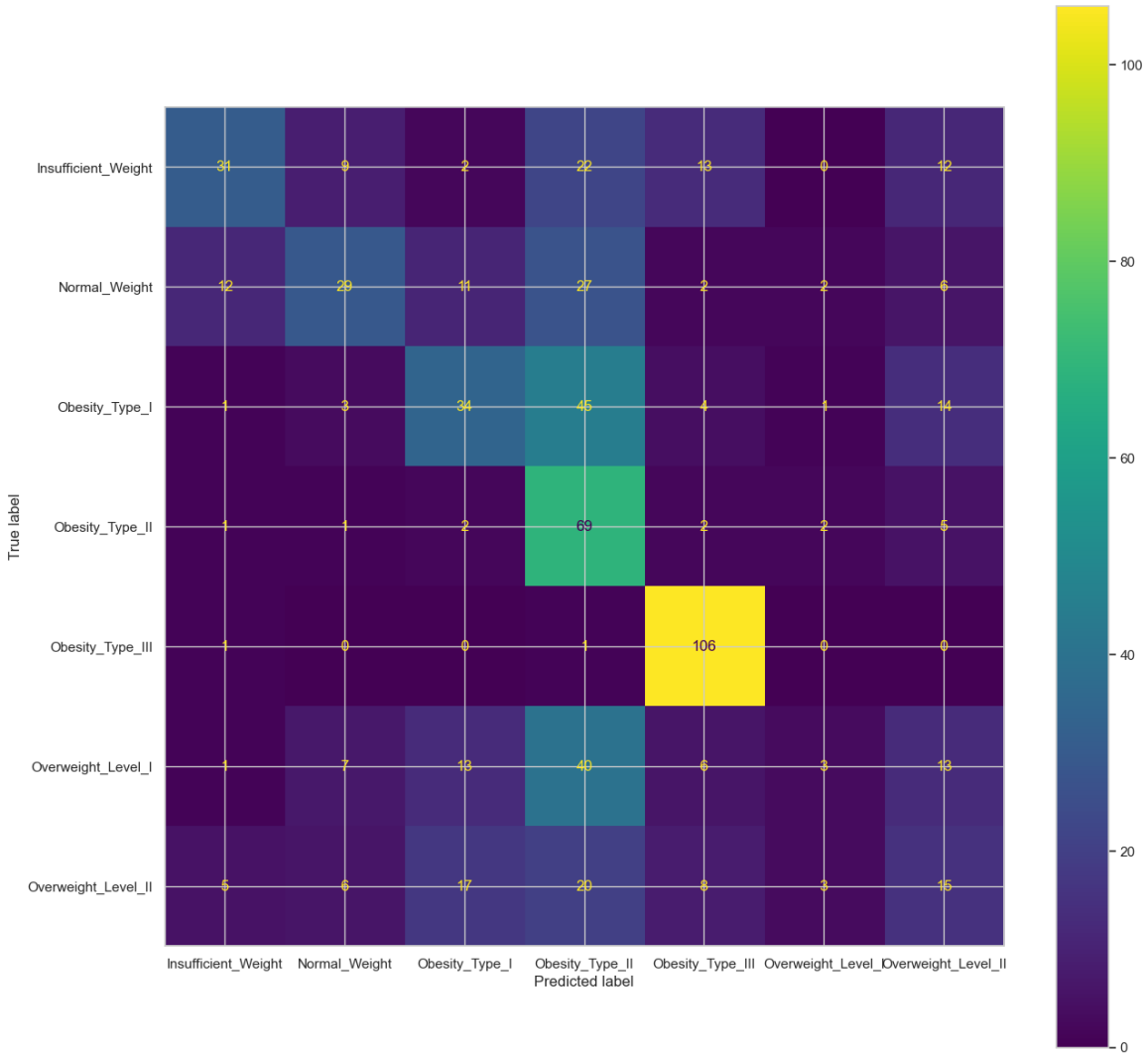



*Figure 29: Confusion Matrix*

From figure 29, can observe that:

Insufficient_Weight:
- True Positives: 31 (correctly classified as "Insufficient_Weight")
- Misclassified: 56 (sum of all non-diagonal values in the row)

Normal_Weight:
- True Positives: 29 (correctly classified as "Normal_Weight")
- Misclassified: 60 (sum of all non-diagonal values in the row)

Obesity_Type_I:
- True Positives: 34 (correctly classified as "Obesity_Type_I")
- Misclassified: 68 (sum of all non-diagonal values in the row)

Obesity_Type_II:
- True Positives: 69 (correctly classified as "Obesity_Type_II")
- Misclassified: 13 (sum of all non-diagonal values in the row)

Obesity_Type_III:
- True Positives: 106 (correctly classified as "Obesity_Type_III")
- Misclassified: 2 (sum of all non-diagonal values in the row)

Overweight_Level_I:
- True Positives: 3 (correctly classified as "Overweight_Level_I")

- Misclassified: 80 (sum of all non-diagonal values in the row)

Overweight_Level_II:

- True Positives: 15 (correctly classified as "Overweight_Level_II")
- Misclassified: 59 (sum of all non-diagonal values in the row)

Below is classification report.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Insufficient_Weight | 0.60 | 0.35 | 0.44 | 89 |
| Normal_Weight | 0.53 | 0.33 | 0.40 | 89 |
| Obesity_Type_I | 0.43 | 0.33 | 0.38 | 102 |
| Obesity_Type_II | 0.31 | 0.84 | 0.45 | 82 |
| Obesity_Type_III | 0.75 | 0.98 | 0.85 | 108 |
| Overweight_Level_I | 0.27 | 0.04 | 0.06 | 83 |
| Overweight_Level_II | 0.23 | 0.20 | 0.22 | 74 |
| accuracy | | | 0.46 | 627 |
| macro avg | 0.45 | 0.44 | 0.40 | 627 |
| weighted avg | 0.46 | 0.46 | 0.42 | 627 |

*Figure 30: Classification report.*

According to the figure 28, Obesity_Type_III is the category in which the model performs the best, with an F1-score of 0.85, a recall of 0.98, and a precision of 0.75, indicating good accuracy and completeness. In other word, Obesity_Type_III has very few positives are misclassified as negative and very few negatives misclassified as positives. The Overweight_Level_I class is where the model is least accurate, suggesting that there will be more false positives in this class. The macro and weighted averages for precision, recall, and F1-score all point to a model's overall accuracy of 0.46, which indicates a modest level of performance overall.

# Task 4: RESULT AND DISCUSSION

a) For noise handling and Preprocessing, correcting misclassified values by rounding off and converting them to integer to be appropriately survey questions. That is a hard part and might occurs mistake in handling process.

b) Performance Metrics and Model Evaluation:
   - Confusion Matrix Analysis (Figure 29): The model performs exceptionally well for 'Obesity_Type_III' with high F1-score, recall, and precision. This indicates that the model is particularly effective in identifying severe cases of obesity, which can be critical in clinical settings.
   - Underperformance in Overweight Categories: The lower performance metrics (recall, precision, F1-score) for 'Overweight_Level_I' suggest that the model struggles to differentiate between borderline cases. This could be due to overlapping characteristics between categories or insufficient distinguishing features.
   - Accuracy Scores: The accuracy score from the original model (0.6) compared to the final optimized model shows an improvement, though the overall performance still points to a modest level of accuracy. This could

be further investigated by looking at class distribution or by employing different algorithms or more sophisticated ensemble techniques.

In conclusion, to answer the question that generate at the beginning of the research 'What dietary habits and daily living habits are significantly affected obesity levels?' According to the result from Task 3, eating behaviours such as eating snacks in between meals (CAEC), eating vegetables frequently (FCVC), and consuming high-calorie foods frequently (FAVC) had a significant effect on obesity rates. In a similar vein, daily routines like screen time (TUE) and physical activity (FAF) are important. These results imply that treatments aiming at decreasing obesity should prioritise encouraging physical exercise, decreasing the frequency of snacks, increasing the intake of vegetables, and promoting good eating habits while limiting sedentary behaviours linked to technology use.

# References

*Just a moment...* (n.d.). Just a moment... https://www.sciencedirect.com/science/article/pii/S235234091930698 5?via%3Dihub

*Obesity and overweight*. (2024, March 1). World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Key%20facts,years%20and%20older)%20were%20overweig ht

*UCI machine learning repository*. (n.d.). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition