BANA 212

Data and Programming for Analytics

Obesity Analysis: Analyzing Correlation and Predictive Models for Obesity

Group B16

Yiwei Lu

Ziqi Zhang

Zaiheng Shen

Wan-Lun Tsai

Chia-Chien Chang

1. Introduction

1.1 Background

Obesity is a complex and pressing public health issue with significant health implications. It is associated with a higher risk of chronic diseases like heart disease, type 2 diabetes, and certain cancers.[1] The financial burden of obesity is also substantial. To better understand and address this issue, we will explore the factors contributing to obesity and develop predictive models using machine learning techniques. This will help identify individuals at risk and inform targeted interventions to prevent or manage obesity.

1.2 Key Questions

In response to the issue of obesity, we propose the following key questions for our report. We aim to address these questions by finding appropriate tools and methodologies during the data analysis process, guided by the following objectives:

A. Can machine learning (ML) techniques be used to predict the risk of developing obesity?
B. What are the most significant factors of obesity, and how can they be addressed?
C. How do eating habits and physical activity affect the risk of obesity?
D. What role do genetic and family factors play in obesity predisposition?

2. Exploratory Data Analysis

2.1 Data Understanding

2.1.1 Data Collection

The dataset we will be using for this analysis is titled **"Estimation of Obesity Levels Based on Eating Habits and Physical Condition."** The data encompasses information collected from

[1]From the internet: Controlling the global obesity epidemic, *World Health Organization* Website.https://www.who.int/activities/controlling-the-global-obesity-epidemic

individuals in Mexico, Peru, and Colombia, offering a diverse representation of eating habits and physical conditions within these regions.[2]

2.1.2 Data availability & Variables description

The Dataset was donated to the UCI repository on August 26, 2019. It comprises **2,111** records and **17** attributes, which include variables related to demographic details, dietary patterns, physical activity levels, and indicators of obesity. This dataset Provides valuable insights into the relationship between eating habits, physical condition, and obesity levels in these countries.

Feature	Description
Gender	Gender
Age	Age
Height	Height
Weight	Weight
family_history_with_overweight	Family history of 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?
CALC	How often do you drink alcohol?
MTRANS	Which transportation do you usually use?
NObeyesdad	Obesity level

2.1.3 Attribute classification

Attributes Related to Eating Habits:

- FAVC: Frequent consumption of high-caloric food
- FCVC: Frequency of consumption of vegetables

[2] From the internet: Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). *UCI Machine Learning Repository.* https://doi.org/10.24432/C5H31Z.

- NCP: Number of main meals
- CAEC: Consumption of food between meals
- CH2O: Consumption of water daily
- CALC: Consumption of alcohol

Attributes Related to Physical Condition

- SCC: Calorie consumption monitoring
- FAF: Physical activity frequency
- TUE: Time using technology devices
- MTRANS: Transportation used

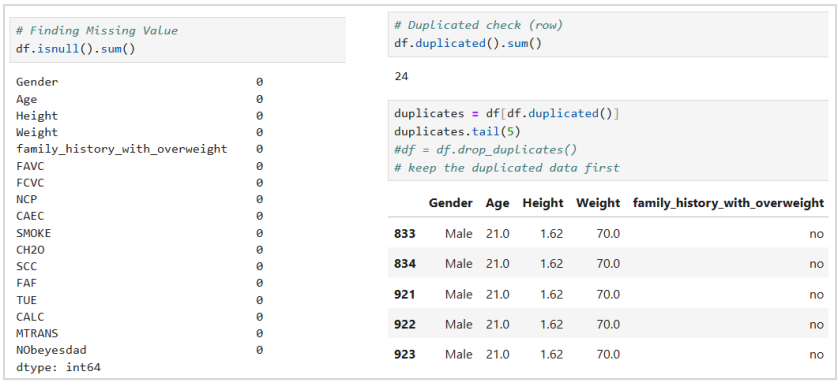
Variables

- Gender
- Age
- Height
- Weight

2.2 Data Cleaning

The next step in addressing the data is to identify any missing values and duplicates. According to Python's output, the dataset is complete with no missing values. However, on the right side of the graph, we found 24 duplicate rows.

Action: We decided not to remove the duplicates because this dataset originates from a web-based survey where respondents anonymously answered each question. These duplicates might represent valid and independent responses**.**



```
# Finding Missing Value
df.isnull().sum()
```

```
Gender                            0
Age                               0
Height                            0
Weight                            0
family_history_with_overweight    0
FAVC                              0
FCVC                              0
NCP                               0
CAEC                              0
SMOKE                             0
CH2O                              0
SCC                               0
FAF                               0
TUE                               0
CALC                              0
MTRANS                            0
NObeyesdad                        0
dtype: int64
```

```
# Duplicated check (row)
df.duplicated().sum()
```

```
24
```

```
duplicates = df[df.duplicated()]
duplicates.tail(5)
#df = df.drop_duplicates()
# keep the duplicated data first
```

	Gender	Age	Height	Weight	family_history_with_overweight
833	Male	21.0	1.62	70.0	no
834	Male	21.0	1.62	70.0	no
921	Male	21.0	1.62	70.0	no
922	Male	21.0	1.62	70.0	no
923	Male	21.0	1.62	70.0	no

2.3 Handling Outliers

To ensure the accuracy of our analysis, it is important to find and handle outliers. After we ran the result in Python, we observed that the “Age” and “NCP”(How many main meals for daily) columns contain a significant number of outliers.

While outliers can potentially affect certain statistical models or analysis results, we have decided to retain these data points for now instead of removing them. This decision stems from the understanding that these outliers are not the result of data entry errors but represent individuals or groups of people with unique characteristics. It probably can provide insights into specific groups or behaviors that differ from normal notions.

If necessary, we will revisit these variables for further data cleaning or modeling adjustments in later stages.

```
df_outliers = df[numerical_columns].copy()
```

```
def detect_all_outliers(data, columns):
    outlier_summary = {}
    for column in columns:
        Q1 = data[column].quantile(0.25)
        Q3 = data[column].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        outliers = data[(data[column] < lower_bound) | (data[column] > upper_bound)]
        outlier_summary[column] = len(outliers)
    return outlier_summary

numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns
outlier_counts = detect_all_outliers(df, numeric_columns)
print(outlier_counts)
```

```
{'Age': 168, 'Height': 1, 'Weight': 1, 'FCVC': 0, 'NCP': 579, 'CH2O': 0, 'FAF': 0, 'TUE': 0}
```

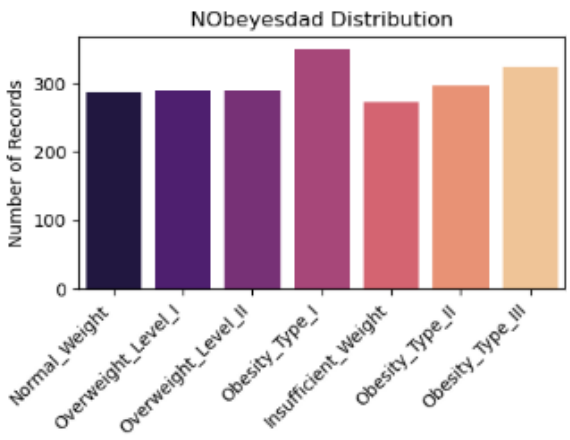
2.4 Data Exploration & Descriptive Statistics

In this stage, we are going to explore the dataset to gain deeper insights into the distribution of variables. This not only helps us to understand the variable but also identify potential relationships between features. This may help us to understand the situation, clarify the problem, and finally find the right tool to take further analysis.

Target variable

Analysis obesity is our current issue, therefore our target variable is the column name NObeyesdad. The histogram demonstrates different classes of categories. On the other hand, we notice that each of the categories is quite evenly distributed.

	NObeyesdad	count
0	Insufficient_Weight	272
1	Normal_Weight	287
2	Obesity_Type_I	351
3	Obesity_Type_II	297
4	Obesity_Type_III	324
5	Overweight_Level_I	290
6	Overweight_Level_II	290



In addition, each classification is based on BIM and then categorizes each obesity level range. Instructions are as follows

BMI Classification[3]

- Underweight: BMI less than 18.5
- Normal weight: BMI between 18.5 and 24.9
- Overweight: BMI between 25.0 and 29.9
- Overweight_Level_I: BMI between 25.0 and 27.4
- Overweight_Level_II: BMI between 27.5 and 29.9
- Obesity
    - Obesity Type I: BMI between 30.0 and 34.9
    - Obesity Type II: BMI between 35.0 and 39.9
    - Obesity Type III: BMI higher than 40

[3] From the internet: "Adult BMI Categories," *Centers for Disease Control and Prevention website*, March 19, 2024, https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

Categorical Variable:

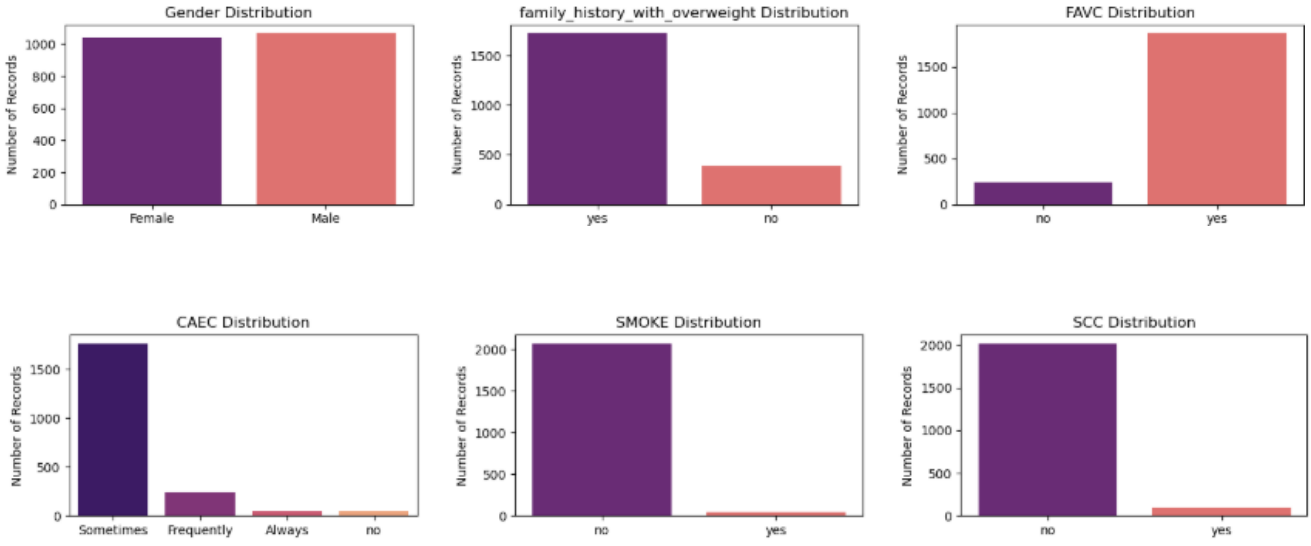
We use the Python code below to quickly identify the categorical variables. And we output some visualizations for each variable distribution.

```
categorical_columns = df.select_dtypes(include=['object', 'category']).columns.tolist()
print("Categorical columns:", categorical_columns)
```

```
Categorical columns: ['Gender', 'family_history_with_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS', 'NObeyesda
d']
```

Based on the graph, we observe the following:

1. Gender: The distribution between females and males is evenly distributed.
2. Family_history_with_overweight: There are 1,726 people who have a family history of being overweight.
3. FAVC(Frequent consumption of high-calorie foods): 1,866 people report eating high-calorie foods.
4. SOMK: 2,067 people are non-smokers, while 44 people smoke.



Numerical Variable:

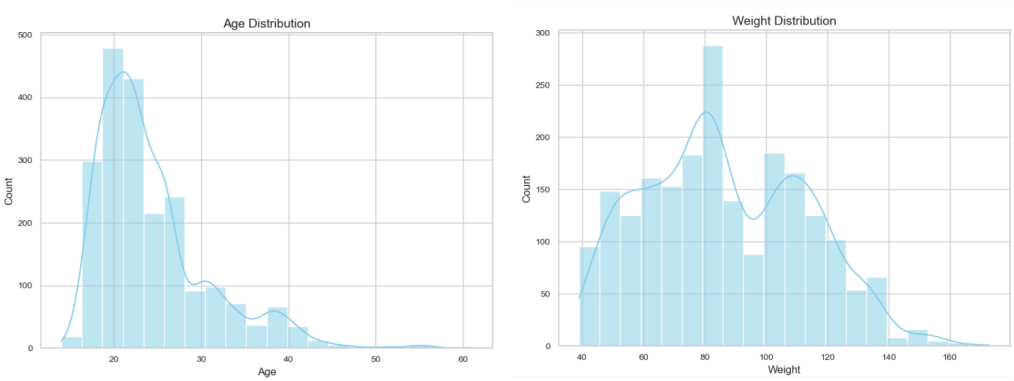
After identifying the numerical variables, we used a heatmap to understand the correction between each variable.

```
numerical_columns = df.select_dtypes(include=['int64', 'float64']).columns.tolist()
print("Numerical columns:", numerical_columns)
```

```
Numerical columns: ['Age', 'Height', 'Weight', 'FCVC', 'NCP', 'CH2O', 'FAF', 'TUE']
```

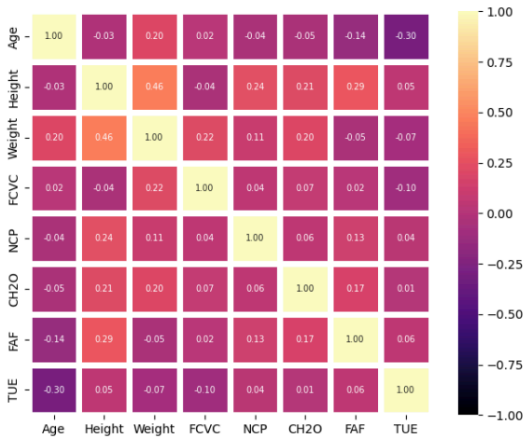
Distribution of Age and Weight:

The age distribution in this dataset is primarily around the younger generation, within the 20 to 30-year-old range.



Correction between each numerical variable:

1. The correlation coefficient between Height and Weight is close to 0.46, showing a moderate positive correlation, indicating that the taller the height, the heavier the weight.
2. The correlations between Age and other variables are very low, even some of them are negative such as NCP, and CH2O. Additionally, the correlations like Weight and FCVC are all lower than 0.05, showing that age has almost nothing to do with these variables.
3. TUE (using time of tech devices) also has very low correlations with other variables, with almost no direct correlation.
4. The correlation between CH2O (water consumption) and NCP (dietary calories) is slightly higher (about 0.24), possibly indicating a link between water consumption and dietary calories. Beside, FCVC(0.07 |vegetables consumption) and FAF(0.17| physical activity) have high correction with CH2O



2.5 Data Visualization

We included both categorical and numerical variables in the heatmap to observe the correlations between each other.

1. Correlation between **family history of obesity** and **body weight**

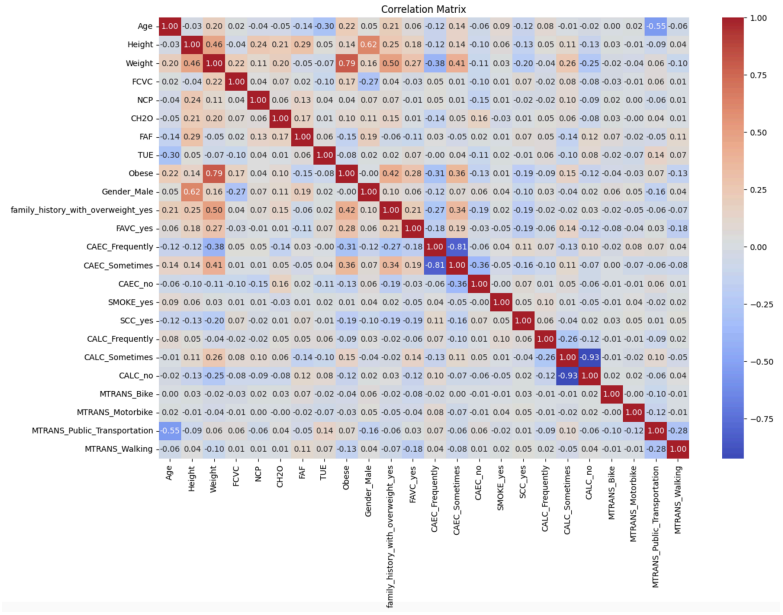
   The correlation coefficient between family history of obesity and weight is 0.39, shows that people with a family history of obesity are more likely to be heavier.

2. The relationship between **healthy eating behaviors** and **obesity**
   - **FAVC (frequency of high fat)** showed a 0.27 correlation with weight, indicating that dietary habits have a certain impact on weight.
   - The correlation coefficient between **NCP (daily number of meals)** and Weight is 0.11, indicating an impact but a weak correlation.

3. Negative correlation between **exercise** and **weight**

   FAF (physical activity frequency) shows a negative correlation (-0.05) with Weight. Although the correlation is not strong, it can be seen that the higher the frequency of exercise, the slight trend of weight loss.

4. There is no significant relationship between SMOKE (smoking behavior) and obesity.



2.6 Data Summarization

1. Outliers or missing values:
    - No missing values were found in the dataset.
    - 24 duplicate rows were identified but retained as they are considered important information.
    - Outliers (Age 168, NCP 579): These will be revisited during data cleaning or modeling adjustments.
2. The data is evenly distributed
    - Gender: Evenly distributed across the dataset.
    - Age: Mainly distributed among young people aged 20-30 years old
    - Target variable NObeyesdad: Shows an approximately balanced distribution.

3. Data Engineering

After analyzing the dataset and identifying the key questions, we approach this as a classification problem. Therefore we are going to use machine learning to address the following objectives:

A. **Predicting Obesity Risk**: A combination of Decision Tree, and Logistic Regression models will be employed to predict the likelihood of individuals developing obesity.
B. **Identifying Significant Factors of Obesity:** Using K-Nearest Neighbors (KNN) to classify them into different clusters. And explore to select the cluster to find the key factor that contributes to obesity classification.
C. **Analyzing Eating Habits and Physical Activity: Logistic Regression** will be used to evaluate the impact of eating habits and physical activity on obesity. Most of the values are yes/no (binary), so should handle values converting first.
D. **Exploring Genetic Contributions:** This one will be a **Linear Regression** and scatter plot to see how family history and obesity develop.

3.1 Decision Tree

Confusion matrix:

The confusion matrix shows the prediction results for each category.

- Insufficient_Weight: The diagonal values show that 87 samples were correctly classified.
- Obesity_Type_I: In the diagonal values 109 samples were correctly classified.

Classification Report:

Precision, which represents how many samples of a certain category are predicted correctly.

- Obesity_Type_III: Has the highest accuracy of 99%, showing almost no error in predicting this category.
- Normal_Weight: Has the lowest accuracy however it still has a good accuracy of 88%, indicating a slightly higher error in predicting this category.

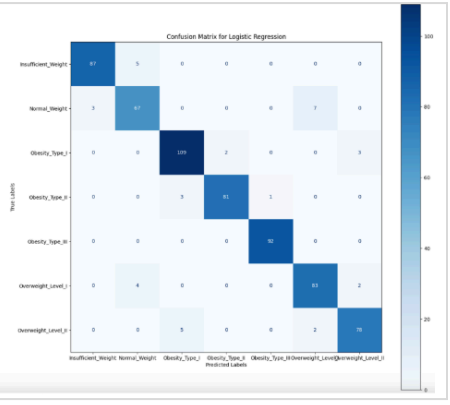
Recall, which represents how many samples that actually belong to a certain category are correctly predicted.

- Obesity_Type_III: The recall rate of Obesity_Type_III is 100%, indicating that all samples of this category are correctly predicted.
- Normal_Weight: Normal_Weight has relatively low recall, meaning these categories are easily confused.

F1-Score is a balance between precision and recall.

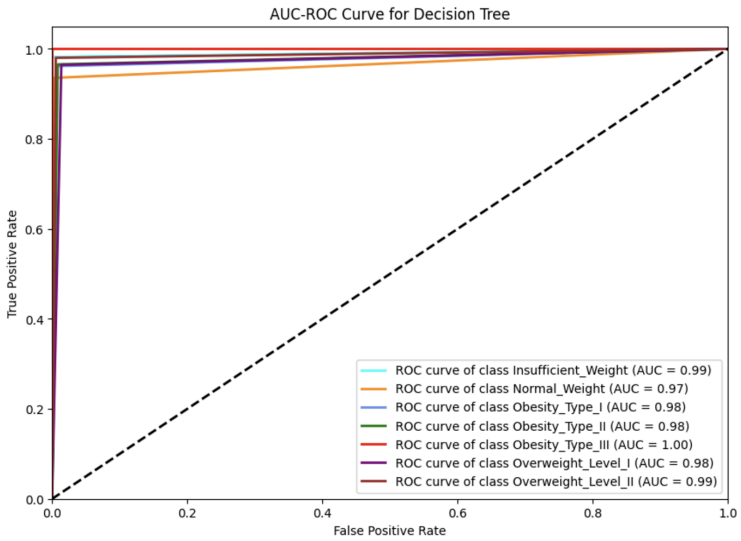
The F1-Score for all categories is close to or above 90%, showing that the model performs well balanced across these categories.

```
Accuracy: 0.9416403785488959
Confusion Matrix:
[[ 87   5   0   0   0   0   0]
 [  3  67   0   0   0   7   0]
 [  0   0 109   2   0   0   3]
 [  0   0   3  81   1   0   0]
 [  0   0   0   0  92   0   0]
 [  0   4   0   0   0  83   2]
 [  0   0   5   0   0   2  78]]
Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.97      0.95      0.96        92
      Normal_Weight       0.88      0.87      0.88        77
     Obesity_Type_I       0.93      0.96      0.94       114
    Obesity_Type_II       0.98      0.95      0.96        85
   Obesity_Type_III       0.99      1.00      0.99        92
 Overweight_Level_I       0.90      0.93      0.92        89
Overweight_Level_II       0.94      0.92      0.93        85

           accuracy                           0.94       634
          macro avg       0.94      0.94      0.94       634
       weighted avg       0.94      0.94      0.94       634
```



The AUC values of all categories are in the range of 0.97~1.00, indicating that the model has a very strong ability to distinguish each category.

- Insufficient Weight (AUC = 0.99): Very close to perfect.
- Obesity_Type_II (AUC = 1.00): Achieve theoretically perfect discrimination.
- Other categories had similarly high performance.



3.2 Random Forest

The accuracy of this model is 93%, this shows that the model can correctly predict about 93% of the samples, showing good classification results.

<u>Classification Report:</u>

Precision

- Obesity_Type_II, Obesity_Type_III: Accuracy reaches 100%
- Normal_Weight: 73% accuracy, indicating some mispredictions

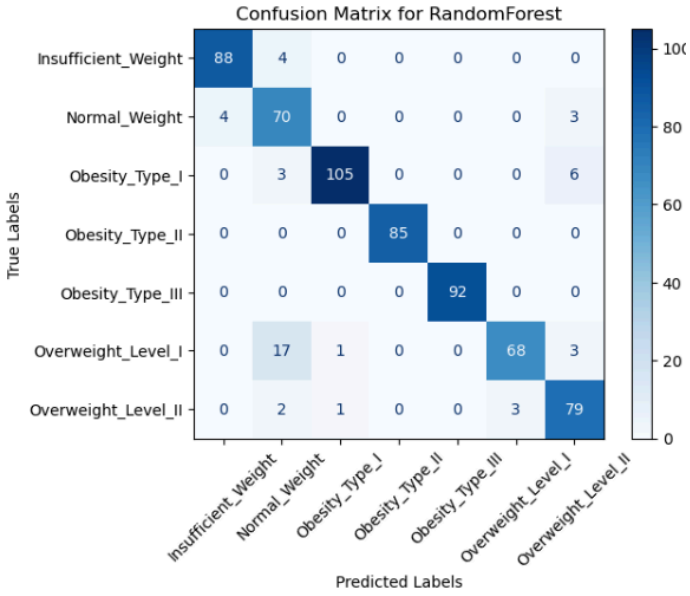
Recall

- Overweight_Level_I: Has the lowest recall rate, which is 76%, indicating that some samples were not predicted correctly.

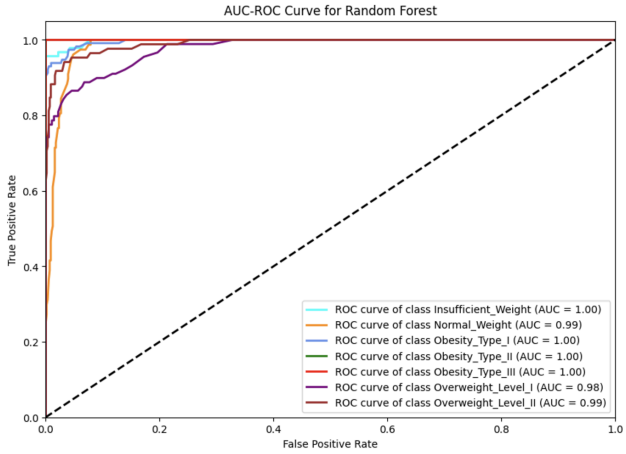
F1-Score:

- The F1 scores for all categories are above 80%, indicating that the model performs well in most categories.

```
Accuracy: 0.9258675078864353
Confusion Matrix:
[[ 88   4   0   0   0   0   0]
 [  4  70   0   0   0   0   3]
 [  0   3 105   0   0   0   6]
 [  0   0   0  85   0   0   0]
 [  0   0   0   0  92   0   0]
 [  0  17   1   0   0  68   3]
 [  0   2   1   0   0   3  79]]
Random Forest:
                     precision    recall  f1-score   support

Insufficient_Weight       0.96      0.96      0.96        92
      Normal_Weight       0.73      0.91      0.81        77
     Obesity_Type_I       0.98      0.92      0.95       114
    Obesity_Type_II       1.00      1.00      1.00        85
   Obesity_Type_III       1.00      1.00      1.00        92
 Overweight_Level_I       0.96      0.76      0.85        89
Overweight_Level_II       0.87      0.93      0.90        85

           accuracy                           0.93       634
          macro avg       0.93      0.93      0.92       634
       weighted avg       0.93      0.93      0.93       634
```



The overall performance of the AUC value of Random Forest is very high, close to 1.00 (perfect classification) for almost every category, indicating that the Random Forest model is very effective for this classification task.



According to the AUC value in the graph:

Insufficient Weight (AUC = 1.00) and Obesity_Type_II (AUC = 1.00): The classification ability of these two categories reaches theoretical perfection (the model can completely distinguish these categories from other categories).

Obesity_Type_I (AUC = 0.99) and Overweight_Level_I (AUC = 0.99): Classification ability for these categories is also close to perfect, but slightly less than perfectly correct.

Normal Weight (AUC = 0.99) and Overweight_Level_II (AUC = 0.98): Classification ability is still very high, but slightly weaker than other categories.

Overall, the AUCs of all categories are in the range of 0.98~1.00, indicating that the model's ability to distinguish each category is almost flawless.

Comparison between Decision Tree and Random Forest

- Random Forest has a more stable performance, especially when dealing with complex data, and can usually effectively avoid overfitting.
- The AUC value of Random Forest is slightly higher than that of Decision Tree, indicating its superior classification ability on this dataset.

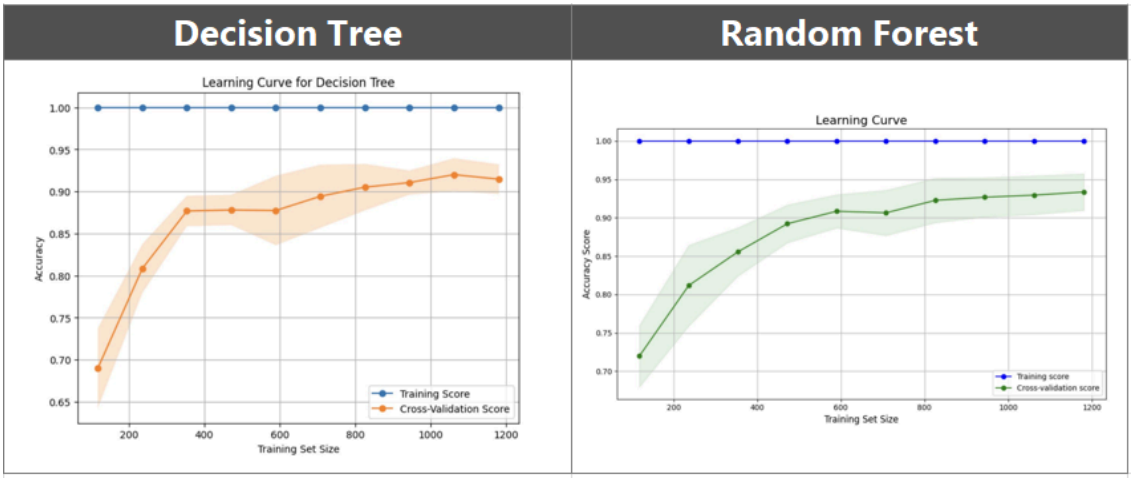
The accuracy of the Decision Tree model is 94.16%, while the Random Forest model achieves 92.59%. Although these accuracy scores appear high, there is a possibility of overfitting. To further evaluate the models, we performed cross-validation.

	Decision Tree				Random Forest			
	precision	recall	f1-score	support	precision	recall	f1-score	support
Insufficient_Weight	0.97	0.95	0.96	92	0.96	0.96	0.96	92
Normal_Weight	0.88	0.87	0.88	77	0.73	0.91	0.81	77
Obesity_Type_I	0.93	0.96	0.94	114	0.98	0.92	0.95	114
Obesity_Type_II	0.98	0.95	0.96	85	1	1	1	85
Obesity_Type_III	0.99	1	0.99	92	1	1	1	92
Overweight_Level_I	0.9	0.93	0.92	89	0.96	0.76	0.85	89
Overweight_Level_II	0.94	0.92	0.93	85	0.87	0.93	0.9	85
accuracy			0.94	634			0.93	634
macro avg	0.94	0.94	0.94	634	0.93	0.93	0.92	634
weighted avg	0.94	0.94	0.94	634	0.93	0.93	0.93	634
Accuracy	94.16%				92.59%			

The learning curves for both the Decision Tree and Random Forest models suggest a high likelihood of overfitting, as explained below:

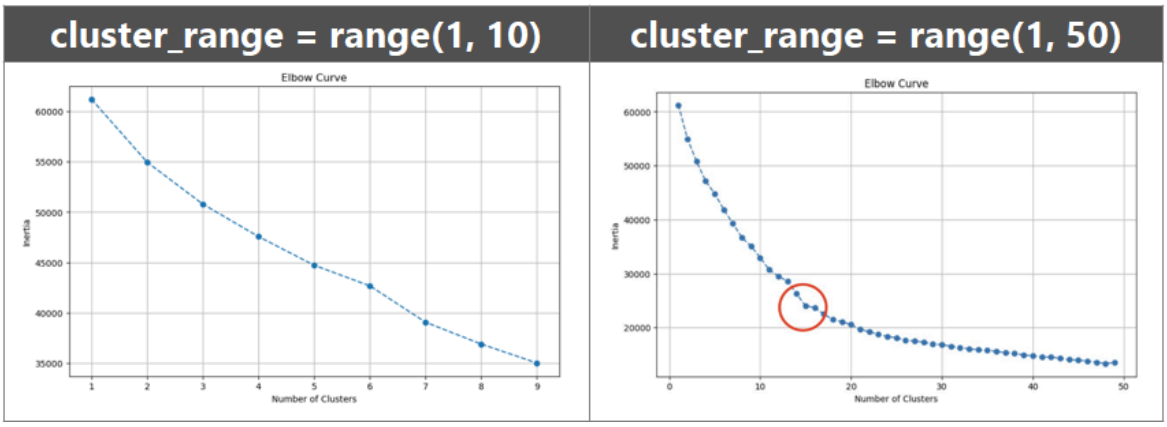
Shaded area:

- Represents the standard deviation range of cross-validation results across different folds, highlighting the variability in model performance.
- A wider orange-shaded area indicates greater fluctuation in performance across different data partitions, suggesting the model's performance is less stable. Conversely, a narrower shaded area indicates more consistent performance.
- Both of them are overfitting but the model performance of Random Forest is more stable compared with Decision Tree.

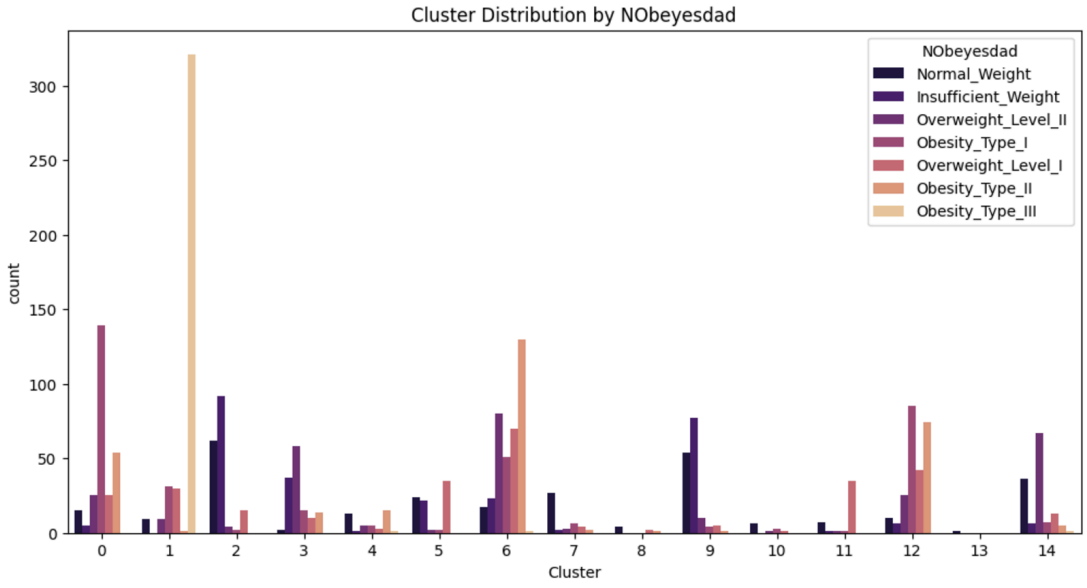


3.3 K Nearest Neighbor

In our KNN model, we initially used 9 clusters (cluster_range = range(1, 10)) but couldn't find a clear elbow point from the elbow curve.



To further explore, we extended the iterations to 49. After analyzing the results, we observed that the elbow curve stabilized at 15, indicating that the rate of change in within-cluster variance significantly slowed down at this point. Based on this observation, we decided to use 15 clusters for our model, as it represents a more optimal clustering solution.



After applying K-Means clustering, we use the histogram to visualize the distribution of NObeyesdad categories within selected clusters.

3.3.1 Obesity

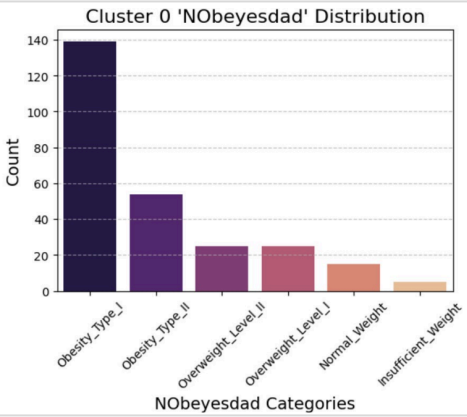
For obesity, we are going to focus on three clusters: Cluster 0, Cluster 1, and Cluster 6. We classify these three into obesity-related. This allowed us to examine how obesity is distributed across these clusters. Below are the insights of our observations for each cluster.

3.3.1.1 Cluster Identification

Cluster 0 | Obesity Type I

The majority category of BMI classification in Cluster0 is Obesity Type I, with 139 people. This category corresponds with a BMI between 30.0 and 34.9, indicating that Cluster 0 predominantly represents individuals with mild obesity.

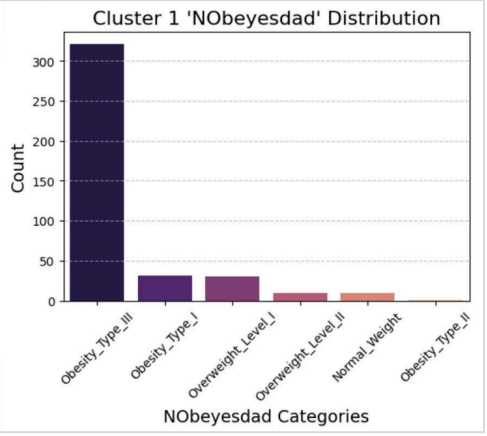
```
Cluster 0 'NObeyesdad' distribution:
NObeyesdad
Obesity_Type_I         139
Obesity_Type_II         54
Overweight_Level_II     25
Overweight_Level_I      25
Normal_Weight           15
Insufficient_Weight      5
Name: count, dtype: int64
```



Cluster 1 | Obesity Type III

The largest group within Group 1 is obesity type III, accounting for 80% of the population in this group. This is the most obese of the category, with a BMI of 40.0 or higher, showing that Cluster 1 is primarily composed of severely obese individuals. We may be able to identify the key factors and correlations that cause obesity.

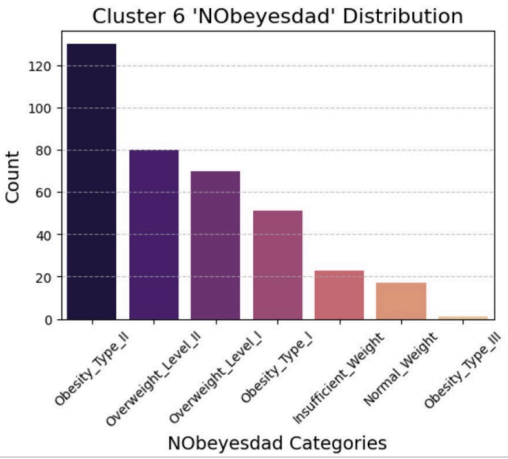
```
Cluster 1 'NObeyesdad' distribution:
NObeyesdad
Obesity_Type_III       321
Obesity_Type_I          31
Overweight_Level_I      30
Overweight_Level_II      9
Normal_Weight            9
Obesity_Type_II          1
Name: count, dtype: int64
```



Cluster 6 | Obesity Type II

In Cluster 6 the distribution is more diverse than the previous two clusters. Obesity Type II (130 people) occupies the majority of obesity categories. Next comes Overweight Level II and I. The presence of both obesity and overweight categories accounted for 88% of the total in this cluster. The remaining 11% consist of individuals with normal and insufficient weight. Within this cluster, we can next look further at what happens between obese and non-obese people.

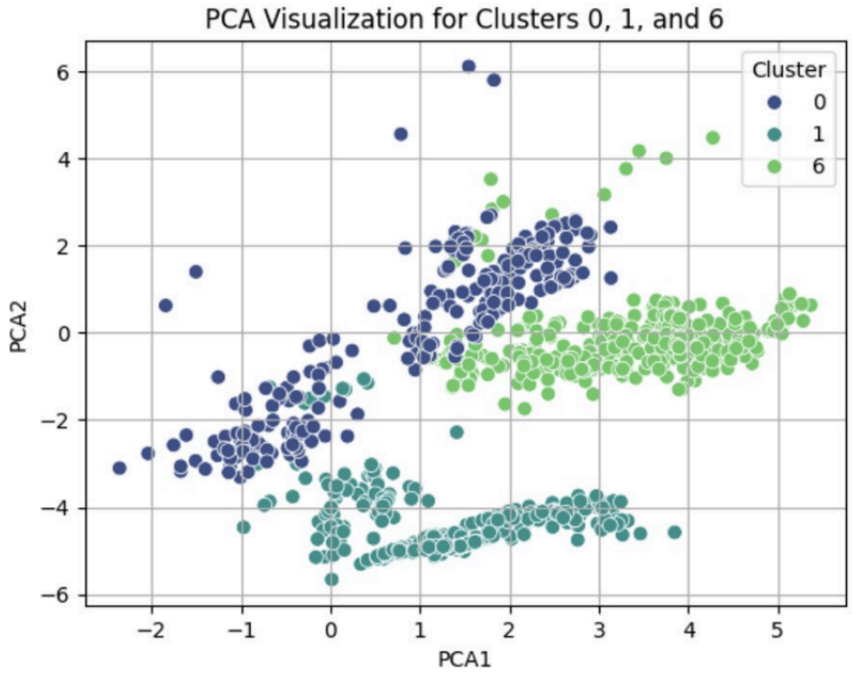
```
Cluster 6 'NObeyesdad' distribution:
NObeyesdad
Obesity_Type_II        130
Overweight_Level_II     80
Overweight_Level_I      70
Obesity_Type_I          51
Insufficient_Weight     23
Normal_Weight           17
Obesity_Type_III         1
Name: count, dtype: int64
```



3.3.1.2 PCV Visualization

Next, we use **PCV visualization** to observe the distribution of Clusters 0, 1, and 6 together. As shown in the figure below, we notice that Cluster 0 and Cluster 1 partially overlap. Similarly,

Cluster 0 and Cluster 6 also overlap. From this result, we can infer that Cluster 0 and Cluster 1 are related, as are Cluster 0 and Cluster 6. However, the relation between Cluster 1 and 6 appears independent. Therefore, our next step is to analyze the correlation between their variables and obesity.



3.3.1.3 ANOVA (F-Statistic, P-Value)

In order to understand the data distribution and variable relationship in each cluster, we use the statistical method ANOVA to verify the significance to observe the F-Statistic and P-Value.

> **F-Statistic**: If the F-value is very high, meaning that the difference in age distribution between NObeyesdad types is very significant.
>
> **P-Value:** While P-value is extremely small (much less than 0.05), indicating that age has a significant effect on different NObeyesdad types.

Below is the result of the ANOVA for each Cluster listing the value of the F-Statistic and P-Value for each variable. We are able to go through each variable and compare it with each

cluster. There are some values showing NaN, which means that when performing an ANOVA analysis, there may be missing values or too little variation in the data for that variable, preventing F and P values from being calculated.

ANOVA Results for Cluster 0:

	Variable	F-Statistic	P-Value
0	Age	30.903716	1.391911e-24
1	Height	4.234017	1.023716e-03
2	Weight	118.274701	1.465830e-64
3	FCVC	37.479402	8.476306e-29
4	NCP	9.388964	3.116930e-08
5	CH2O	41.554611	2.908690e-31
6	FAF	1.309404	2.603375e-01
7	TUE	6.531682	9.708903e-06
8	Gender_Female	13.430745	1.214803e-11
9	Gender_Male	13.430745	1.214803e-11
10	family_history_with_overweight_no	NaN	NaN
11	family_history_with_overweight_yes	NaN	NaN
12	FAVC_no	NaN	NaN
13	FAVC_yes	NaN	NaN
14	CAEC_Always	NaN	NaN
15	CAEC_Frequently	NaN	NaN
16	CAEC_Sometimes	NaN	NaN
17	CAEC_no	NaN	NaN
18	SMOKE_no	NaN	NaN
19	SMOKE_yes	NaN	NaN
20	SCC_no	NaN	NaN
21	SCC_yes	NaN	NaN
22	CALC_Always	NaN	NaN
23	CALC_Frequently	6.342828	1.424126e-05
24	CALC_Sometimes	NaN	NaN
25	CALC_no	6.342828	1.424126e-05
26	MTRANS_Automobile	NaN	NaN
27	MTRANS_Bike	NaN	NaN
28	MTRANS_Motorbike	NaN	NaN
29	MTRANS_Public_Transportation	12.814248	3.928768e-11
30	MTRANS_Walking	12.814248	3.928768e-11

ANOVA Results for Cluster 1:

	Variable	F-Statistic	P-Value
0	Age	6.972520	2.946528e-06
1	Height	4.729034	3.300501e-04
2	Weight	129.816783	4.404039e-81
3	FCVC	151.587499	1.491291e-89
4	NCP	46.501103	9.573508e-38
5	CH2O	3.990404	1.525366e-03
6	FAF	7.324152	1.402393e-06
7	TUE	16.033151	2.145944e-14
8	Gender_Female	NaN	NaN
9	Gender_Male	NaN	NaN
10	family_history_with_overweight_no	NaN	NaN
11	family_history_with_overweight_yes	NaN	NaN
12	FAVC_no	NaN	NaN
13	FAVC_yes	NaN	NaN
14	CAEC_Always	NaN	NaN
15	CAEC_Frequently	NaN	NaN
16	CAEC_Sometimes	NaN	NaN
17	CAEC_no	NaN	NaN
18	SMOKE_no	NaN	NaN
19	SMOKE_yes	NaN	NaN
20	SCC_no	NaN	NaN
21	SCC_yes	NaN	NaN
22	CALC_Always	NaN	NaN
23	CALC_Frequently	29.877416	9.467028e-26
24	CALC_Sometimes	29.877416	9.467028e-26
25	CALC_no	NaN	NaN
26	MTRANS_Automobile	NaN	NaN
27	MTRANS_Bike	NaN	NaN
28	MTRANS_Motorbike	NaN	NaN
29	MTRANS_Public_Transportation	9.653367	1.040901e-08
30	MTRANS_Walking	9.653367	1.040901e-08

To interpret it more clearly, we filter P-value <0.05. The table provided below shows the significant variables. Explaining as below

Comparison in Cluster 0 and Cluster1

Common variables: ['Height', 'Weight', 'TUE', 'CALC_Frequently', 'MTRANS_Walking', 'NCP', 'MTRANS_Public_Transportation', 'FCVC', 'Age', 'CH2O']

- These common variables suggest consistent patterns across both clusters, emphasizing the importance of age, body measurements, water intake, and transportation habits in understanding obesity-related outcomes.

Unique Variable:

Cluster0: ['Gender_Male', 'Gender_Female', 'CALC_no']

Cluster1: ['FAF', 'CALC_Sometimes']

- The statistical significance of **gender-related factors** (based on F-statistic and p-value) confirms that **gender** is a key influencer in Cluster 0.
- **FAF** (frequency of physical activity) highlights the importance of **exercise** in this cluster.
- **CALC_Sometimes** (occasional alcohol consumption) suggests that moderate drinking habits are more prevalent or impactful in Cluster 1.

Cluster0

Significant Variables for Cluster 0 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	30.903716	1.391911e-24
1	Height	4.234017	1.023716e-03
2	Weight	118.274701	1.465830e-64
3	FCVC	37.479402	8.476306e-29
4	NCP	9.388964	3.116930e-08
5	CH2O	41.554611	2.908690e-31
7	TUE	6.531682	9.708903e-06
8	Gender_Female	13.430745	1.214803e-11
9	Gender_Male	13.430745	1.214803e-11
23	CALC_Frequently	6.342828	1.424126e-05
25	CALC_no	6.342828	1.424126e-05
29	MTRANS_Public_Transportation	12.814248	3.928768e-11
30	MTRANS_Walking	12.814248	3.928768e-11

Cluster1

Significant Variables for Cluster 1 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	6.972520	2.946528e-06
1	Height	4.729034	3.300501e-04
2	Weight	129.816783	4.404039e-81
3	FCVC	151.587499	1.491291e-89
4	NCP	46.501103	9.573508e-38
5	CH2O	3.990404	1.525366e-03
6	FAF	7.324152	1.402393e-06
7	TUE	16.033151	2.145944e-14
23	CALC_Frequently	29.877416	9.467028e-26
24	CALC_Sometimes	29.877416	9.467028e-26
29	MTRANS_Public_Transportation	9.653367	1.040901e-08
30	MTRANS_Walking	9.653367	1.040901e-08

<u>Comparison in Cluster 0 and Cluster6</u>

Common variables: ['Height', 'Weight', 'TUE', 'MTRANS_Walking', 'NCP', 'MTRANS_Public_Transportation', 'FCVC', 'Age', 'CH2O']

Age, Height, Weight, CH2O, and TUE are particularly noteworthy, suggesting they have broad predictive power in obesity-related outcomes across both clusters.

<u>Unique Variable:</u>

Cluster 0 : ['Gender_Male', 'Gender_Female', 'CALC_no', 'CALC_Frequently']

Cluster 6 : ['FAF', 'CAEC_Frequently', 'CAEC_Sometimes']

- Cluster 0 demonstrates that alcohol consumption patterns (CALC_no, CALC_Frequently) suggest that demographic factors and lifestyle choices around alcohol play a significant role.
- Cluster 6 shows the physical activity (FAF) and snacking behavior (CAEC_Frequently, CAEC_Sometimes) highlight the exercise and eating habits in defining this cluster.

Cluster0

Significant Variables for Cluster 0 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	30.903716	1.391911e-24
1	Height	4.234017	1.023716e-03
2	Weight	118.274701	1.465830e-64
3	FCVC	37.479402	8.476306e-29
4	NCP	9.388964	3.116930e-08
5	CH2O	41.554611	2.908690e-31
7	TUE	6.531682	9.708903e-06
8	Gender_Female	13.430745	1.214803e-11
9	Gender_Male	13.430745	1.214803e-11
23	CALC_Frequently	6.342828	1.424126e-05
25	CALC_no	6.342828	1.424126e-05
29	MTRANS_Public_Transportation	12.814248	3.928768e-11
30	MTRANS_Walking	12.814248	3.928768e-11

Cluster6

Significant Variables for Cluster 6 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	34.049303	1.306668e-32
1	Height	18.865552	3.829173e-19
2	Weight	961.174122	3.598123e-220
3	FCVC	2.228991	3.987319e-02
4	NCP	19.117558	2.195173e-19
5	CH2O	3.146194	5.085702e-03
6	FAF	6.245212	2.930055e-06
7	TUE	20.406472	1.305367e-20
15	CAEC_Frequently	inf	0.000000e+00
16	CAEC_Sometimes	inf	0.000000e+00
29	MTRANS_Public_Transportation	6.301723	2.551604e-06
30	MTRANS_Walking	6.301723	2.551604e-06

Comparison in Cluster 1 and Cluster6

Common variables: ['Height', 'Weight', 'TUE', 'MTRANS_Walking', 'NCP', 'MTRANS_Public_Transportation', 'FCVC', 'Age', 'CH2O']

Unique Variable:

Cluster 1 : [‘'Gender_Male', 'Gender_Female', 'CALC_no', 'CALC_Frequently']

Cluster 6 : [‘'FAF', 'CAEC_Frequently', 'CAEC_Sometimes']

- In Cluster 1, alcohol consumption varies widely, with more people either drinking frequently or not at all (CALC_no, CALC_Frequently). This suggests lifestyle choices related to alcohol are a key differentiator.

- People in **Cluster 6** engage in more **physical activity(FAF)** and tend to eat **between meals (CAEC)** more often. This indicates a focus on **physical activity and snacking**, which sets this cluster apart from Cluster 1.

This comparison highlights how **alcohol consumption** in Cluster 1 contrasts with **active lifestyles and snacking habits** in Cluster 6.

3.3.1.4 Revised Summary:

Across all three clusters, common factors such as **Height**, **Weight**, **TUE** (time spent using technology), **MTRANS_Walking**, **NCP** (number of meals), **MTRANS_Public_Transportation**, **FCVC** (frequency of vegetable consumption), **Age**, and **CH2O** (water intake) consistently show importance in predicting obesity-related outcomes.

The key differences that drive obesity within each cluster are as follows:

- **Cluster 0**: Influenced primarily by **demographic factors**, notably **gender** and **alcohol abstinence**.
- **Cluster 1**: Shaped by **lifestyle behaviors**, with an emphasis on **physical activity** and **moderate alcohol consumption**.
- **Cluster 6**: Defined by **behavioral factors**, particularly **physical activity** and **snacking patterns**.

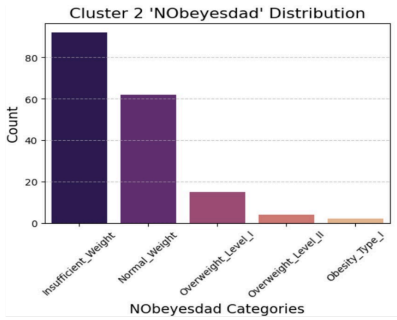
3.3.2 Non-Obesity

For non-obesity, we analyzed Cluster 2 and Cluster 9. They have fewer obesity cases compared to other clusters. This can help us understand the characteristics and factors of those who are in normal weight. These insights may offer valuable suggestions for helping individuals prevent or manage obesity effectively.

This can be seen from the two charts below. In both Clusters, Insufficient Weight & Normal Weight account for more than 80% of the clusters in which they are distributed. Therefore, further analysis can help us understand the eating habits of non-obese people and help us provide suggestions for people who are obese and want to lose weight.

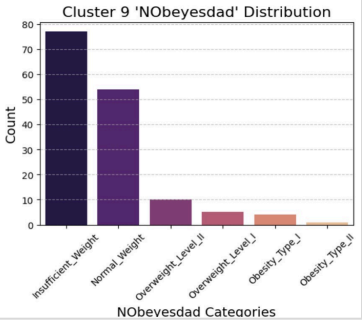
Cluster 2 | Insufficient Weight & Normal Weight

```
Cluster 2 'NObeyesdad' distribution:
NObeyesdad
Insufficient_Weight    92
Normal_Weight          62
Overweight_Level_I     15
Overweight_Level_II     4
Obesity_Type_I          2
Name: count, dtype: int64
```



Cluster 9 | Insufficient Weight & Normal Weight

```
Cluster 9 'NObeyesdad' distribution:
NObeyesdad
Insufficient_Weight    77
Normal_Weight          54
Overweight_Level_II    10
Overweight_Level_I      5
Obesity_Type_I          4
Obesity_Type_II         1
Name: count, dtype: int64
```



<u>Comparison in Cluster 2 and Cluster9</u>

Common variables: ['Height', 'Weight', 'TUE', 'CALC_Frequently', 'MTRANS_Walking', 'NCP', 'MTRANS_Public_Transportation', 'FCVC', 'Age', 'CH2O"]

Unique Variable:

Cluster 2 : [‘'Gender_Male', 'Gender_Female', 'CALC_no']

Cluster 9 : ['FAF', 'CALC_Sometimes']

- In Cluster 2 factors like gender and alcohol abstinence (CALC_no) may indicate that these demographic and lifestyle choices contribute to maintaining a lower weight.
- Cluster 9 presence of physical activity (FAF) and occasional alcohol consumption (CALC_Sometimes) suggests that regular exercise and moderation in alcohol intake might play a protective role in preventing obesity.

Cluster2

Significant Variables for Cluster 2 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	23.105987	2.906257e-15
1	Height	2.202304	7.079205e-02
2	Weight	90.483644	4.569255e-41
3	FCVC	0.139478	9.673751e-01
4	NCP	0.571828	6.834337e-01
5	CH2O	0.821166	5.133192e-01
7	TUE	1.791104	1.328249e-01
8	Gender_Female	7.537935	1.296108e-05
9	Gender_Male	7.537935	1.296108e-05
23	CALC_Frequently	3.580920	7.850151e-03
25	CALC_no	7.854230	7.835538e-06
29	MTRANS_Public_Transportation	4.937433	8.631081e-04
30	MTRANS_Walking	1.395400	2.376236e-01

Cluster9

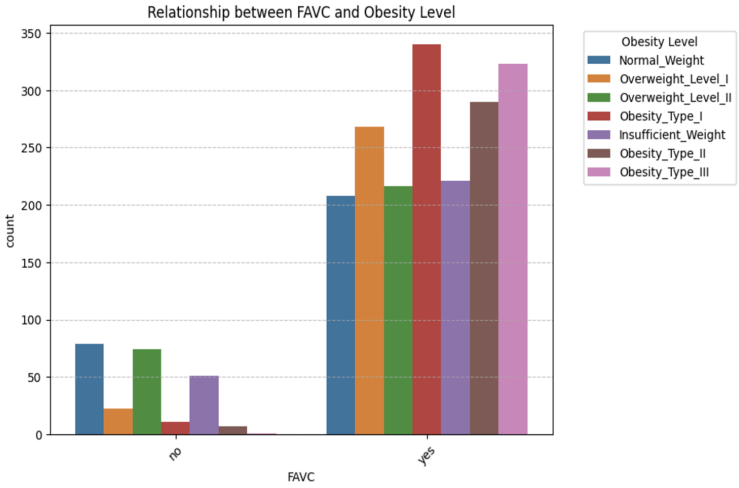
Significant Variables for Cluster 9 (p-value < 0.05):

	Variable	F-Statistic	P-Value
0	Age	1.692201	1.400743e-01
1	Height	1.052649	3.892663e-01
2	Weight	80.929443	3.347720e-40
3	FCVC	3.373630	6.509923e-03
4	NCP	0.158807	9.770318e-01
5	CH2O	1.441548	2.128391e-01
6	FAF	1.647097	1.512238e-01
7	TUE	1.515125	1.886066e-01
23	CALC_Frequently	4.002889	1.972667e-03
24	CALC_Sometimes	2.975850	1.376077e-02
29	MTRANS_Public_Transportation	3.421010	5.952153e-03
30	MTRANS_Walking	0.976095	4.345271e-01

3.4 Logistic Regression

3.4.1 Eating Habits with Obesity

3.4.1.1 FAVC - Frequent consumption of high-caloric food.



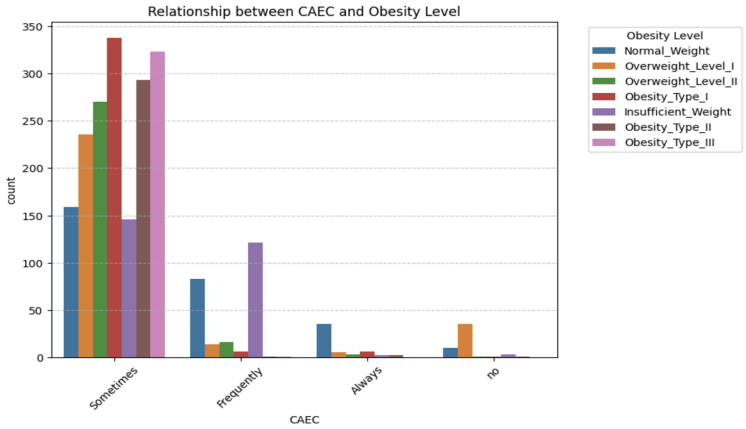
Based on the (3.4.1.1) bar chart analysis, we observed that the majority of individuals classified as obese, including those in Obesity_Type_I, Obesity_Type_II, and Obesity_Type_III categories, consistently selected "Yes" for the variable **"Frequent consumption of high-caloric food" (FAVC).** However, a notable number of individuals in the "Normal Weight" and "Insufficient Weight" categories also selected "Yes."

This finding highlights a strong relationship between frequent consumption of high-caloric food and obesity, but it also suggests that high-caloric food consumption is not exclusive to obese

individuals. Other factors, such as metabolism, physical activity, or overall dietary patterns, may play a role in mitigating its impact on weight for individuals with normal or insufficient weight.

3.4.1.2 CAEC - Consumption of Food Between Meals

Based on the (3.4.1.2) bar chart analysis, we observe that the majority of individuals across all weight categories, including obese individuals (Obesity_Type_I, Obesity_Type_II, and Obesity_Type_III), selected "Sometimes" for the variable **"Consumption of food between meals" (CAEC)**. This indicates that occasional snacking between meals is a common habit across all groups, regardless of weight status.



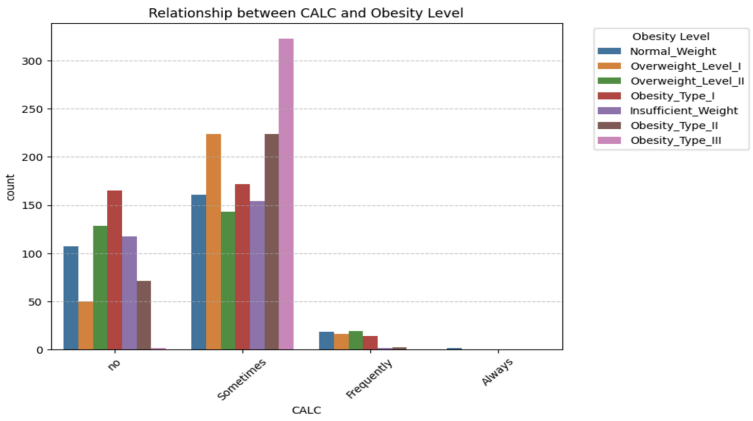
Interestingly, for the "Frequently" and "Always" options, the largest number of individuals selecting these categories belong to the "Normal Weight" and "Insufficient Weight" groups, rather than the obese categories.

This suggests that frequent snacking is not exclusive to individuals with higher obesity levels. Instead, this behavior is observed in lighter-weight groups as well, which points to other factors such as metabolism, physical activity, or the quality of snacks consumed that may mitigate the effects of frequent snacking on weight.

3.4.1.3 CALC - Consumption of Alcohol

Based on the (1.1.3) bar chart analysis, we observed that the majority of individuals across all weight categories, including those in Obesity_Type_I, Obesity_Type_II, and Obesity_Type_III

categories, consistently selected “Sometimes” for the variable **“Alcohol consumption” (CALC)**.

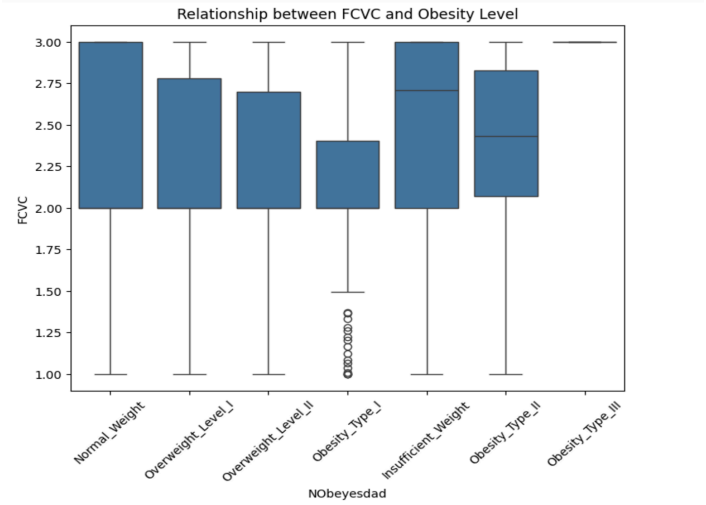


However, a notable number of individuals in the “Normal Weight” and “Insufficient Weight” categories also selected “Sometimes,” and the highest counts for this option were observed in the Obesity_Type_III category.

This finding highlights a common tendency toward occasional alcohol consumption across all weight categories. However, it also suggests that alcohol consumption is not exclusive to individuals with obesity. Other factors, such as the type of alcohol consumed, lifestyle habits, or overall caloric intake, may play a role in moderating its impact on weight, particularly for individuals with normal or insufficient weight.

3.4.1.4 FCVC - Frequency of consumption of vegetables

The (1.1.4) boxplot shows the relationship between the **Frequency of Consumption of Vegetables (FCVC)** and different obesity levels:



In most groups, the median frequency of vegetable consumption is between 2.5 and 3 times per day, suggesting that eating vegetables regularly is a common habit, regardless of weight status.

Obesity_Type_I has the lowest median vegetable consumption among all categories.

There are outliers showing that some individuals in this group eat vegetables as little as once a day, which could indicate poor eating habits in part of the group.

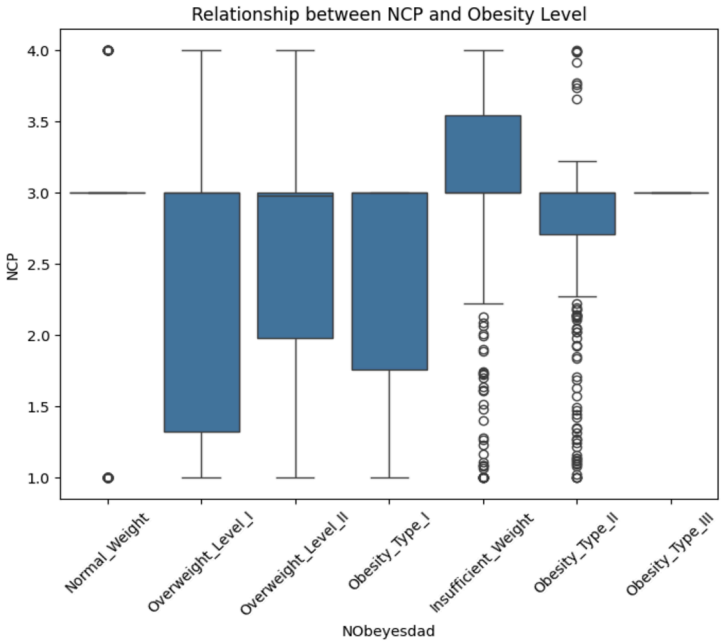
Normal Weight Groups and Insuffication Weight Groups groups have a median vegetable consumption of close to 3 times per day and less variation compared to other groups. This consistency suggests that frequent vegetable consumption may be linked to maintaining a healthy weight.

The Obesity_Type_III group also has a high median vegetable consumption, similar to the Normal Weight and Insufficient Weight groups. However, there is more variation in this group, indicating that while some individuals eat vegetables frequently, others may have less consistent eating habits.

3.4.1.5 NCP - Number of Main Meals

The (1.1.5) box plot shows the relationship between the **Number of Main Meals (NCP)** and different obesity levels. The median number of main meals for most groups is about 3 meals per day, and this is especially clear for individuals with Normal Weight. Almost all of them report

eating 3 meals a day, with very little variation. This shows that having a consistent routine of three main meals might be linked to maintaining a normal weight.



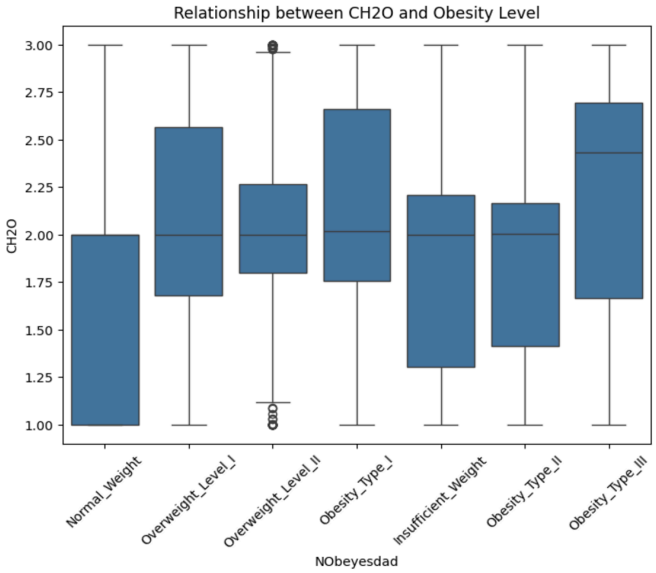
Some groups, like Obesity_Type_II and Insufficient Weight, show a lot of variation in the number of meals, with many people reporting fewer than 2 meals per day. Overweight_Level_I and Overweight_Level_II groups also have some variation but mostly stick to 3 meals a day. The Obesity_Type_I group is similar but with more irregular patterns, with some individuals eating as few as 1.5 meals. Insufficient Weight has the most inconsistent eating habits, which may make it harder to maintain a healthy weight. Meanwhile, Obesity_Type_II and Obesity_Type_III groups also show inconsistency, especially in Obesity_Type_II, where irregular meal routines might contribute to obesity.

Consistency in Meal Patterns: Individuals with Normal Weight tend to have a steady meal pattern of 3 main meals per day, while those with obesity or insufficient weight show more irregularities and variability in their meal routines.

Outliers in Obesity_Type_II and Insufficient Weight: The significant number of outliers in these groups highlights irregular eating habits, which may be linked to unhealthy weight levels, either due to overcompensation or insufficient caloric intake.

3.4.1.6 CH2O - Consumption of Water Daily

The (1.1.6) box plot shows the relationship between the **Consumption of Water Daily** (CH2O) and different obesity levels.



The median daily water consumption across all groups ranges between 2 and 2.5 liters. Especially the normal weight group. The main distribution is in 1 - 2 liters.

The Obesity_Type_I and Obesity_Type_II groups have a median water intake of about 2 liters, but their water consumption varies more compared to the Normal Weight and Overweight groups, showing less consistent hydration habits. The Obesity_Type_III group has the highest median intake, close to 2.5 liters, but with a wide range, indicating mixed drinking patterns.

For the Insufficient Weight group, the median is also around 2 liters, but with higher variability, as some individuals drink as little as 1 liter daily. This inconsistency in water intake, particularly in the Insufficient Weight group and among those drinking less than 2 liters, may reflect irregular dietary habits affecting their weight.

3.4.1.7 Logistic Regression Results

```
Optimization terminated successfully.
         Current function value: 0.551024
         Iterations 26
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Obese   No. Observations:                 2111
Model:                          Logit   Df Residuals:                     2100
Method:                           MLE   Df Model:                           10
Date:                Thu, 28 Nov 2024   Pseudo R-squ.:                  0.2014
Time:                        22:28:15   Log-Likelihood:                -1163.2
converged:                       True   LL-Null:                       -1456.6
Covariance Type:            nonrobust   LLR p-value:                1.173e-119
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const            -34.9714        nan        nan        nan        nan        nan
FCVC               0.7641      0.098      7.778      0.000      0.572      0.957
NCP                0.0151      0.068      0.220      0.826     -0.119      0.149
CH2O               0.2765      0.085      3.238      0.001      0.109      0.444
FAVC_yes           2.3299      0.253      9.225      0.000      1.835      2.825
CAEC_Frequently   -1.7164      0.542     -3.169      0.002     -2.778     -0.655
CAEC_Sometimes     1.6604      0.404      4.110      0.000      0.869      2.452
CAEC_no           -1.6573      0.834     -1.987      0.047     -3.292     -0.023
CALC_Frequently   28.0718        nan        nan        nan        nan        nan
CALC_Sometimes    28.9713        nan        nan        nan        nan        nan
CALC_no           28.6873        nan        nan        nan        nan        nan
==============================================================================
```

Combining the previous graphs and regression results, the most significant factor influencing obesity is **FAVC_yes (frequent consumption of high-calorie foods)**. From the histogram, it is clear that the majority of obese individuals selected “Yes” for this variable. The regression coefficient of 2.3299 is highly significant (p-value = 0.000), indicating that regular intake of high-calorie foods greatly increases the risk of obesity.

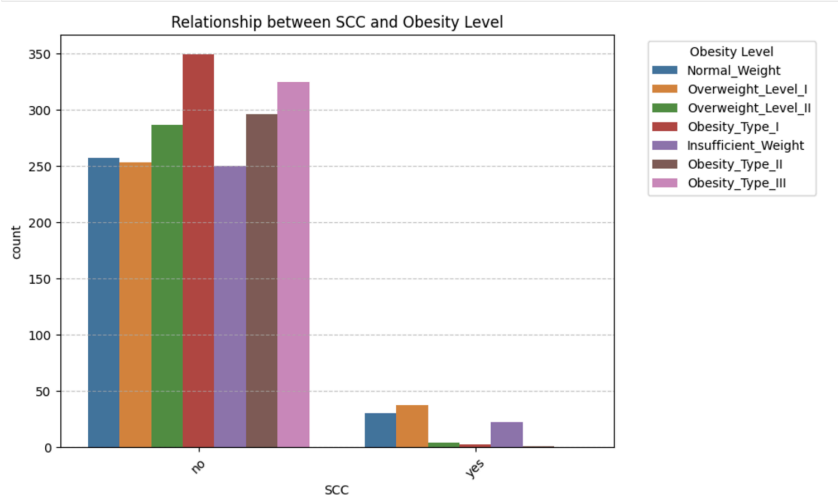
Additionally, the regression results show positive correlations between obesity and **FCVC (frequency of vegetable intake)** as well as **CH2O (water consumption)**, with coefficients of 0.7641 and 0.2765, respectively. From the graphs, we can observe that obese individuals generally report a higher frequency of vegetable and water intake. However, this may reflect their overall higher food and drink consumption or other dietary patterns, rather than vegetables or water directly contributing to obesity.

On the other hand, **CAEC_Frequently (frequent snacking)** is negatively correlated with obesity, with a coefficient of -1.72. The graphs show that while some obese individuals report frequent snacking, most select “Sometimes.” This suggests that the relationship between snacking and obesity may be influenced by the total amount or type of food consumed during

snacks. These findings highlight the complexity of dietary habits and indicate that focusing on a single factor is insufficient to fully explain the causes of obesity.

3.4.2 Physical Activities with Obesity

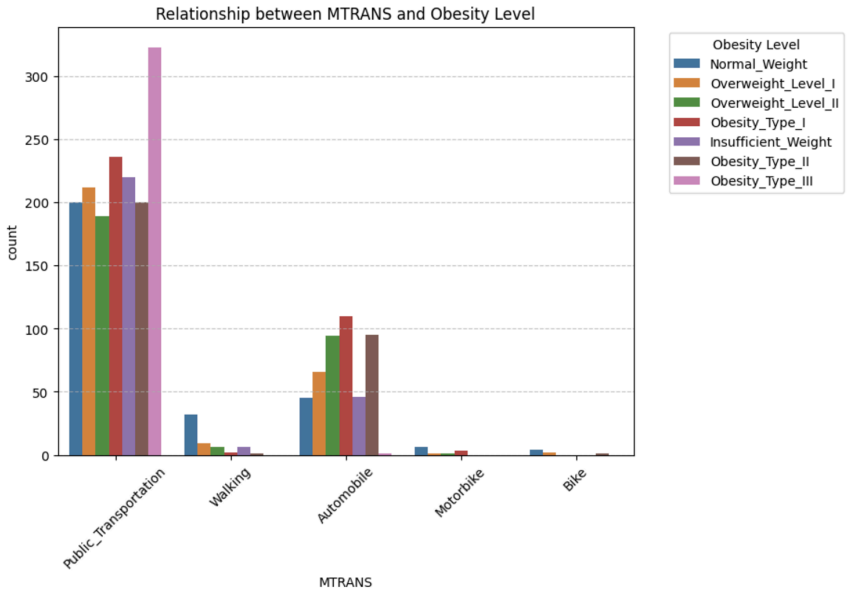
3.4.2.1 SCC - Calorie consumption monitoring



The insight we can tell from Figure 1.2.1 that the relationship between **calorie consumption Monitoring (SCC)**, most people do not monitor their calorie intake regardless of their weight status, especially in the Obesity_Type_III (severe obesity) group, where the largest number of people choose "no monitoring". This shows that the more obese people are, the less likely they are to actively monitor their calories. The number of people who choose "yes" is very small, only slightly more in the Overweight_Level_I and Obesity_Type_I groups, which may reflect that they have a certain awareness of weight management. Overall, calorie monitoring is not a common behavior for most people, which may be a key factor in the persistence of obesity.

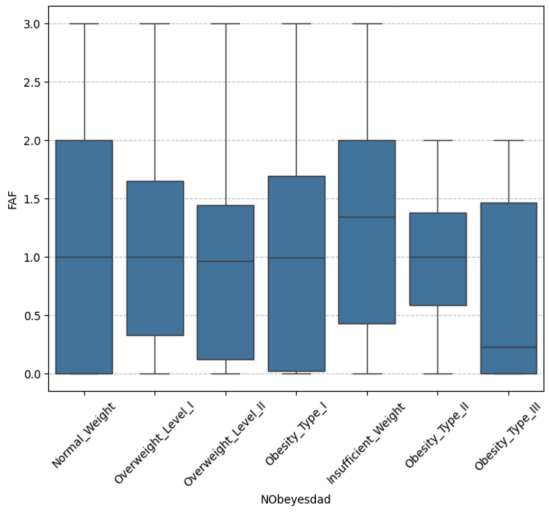
3.4.2.2 MTRANS - Transportation Used

The insight we can tell from Figure 3.4.2.2 is that the relationship between **Transportation Used (MTRANS)**, public transportation is the most common way of transportation in all weight categories, especially in the Obesity_Type_III (severely obese) group, where the usage rate is the highest.



However, the way of transportation that requires physical activity, such as walking and cycling, is rarely chosen in all weight categories, reflecting a trend of lack of active activities. In contrast, the Overweight_Level_I, Overweight_Level_II, and Obesity_Type_I groups are more inclined to use non-active means of transportation, such as cars. This reliance on modes of transportation that lack physical activity may lead to a decrease in overall activity levels, thereby affecting obesity rates.

3.4.2.3 FAF - Physical Activity Frequency

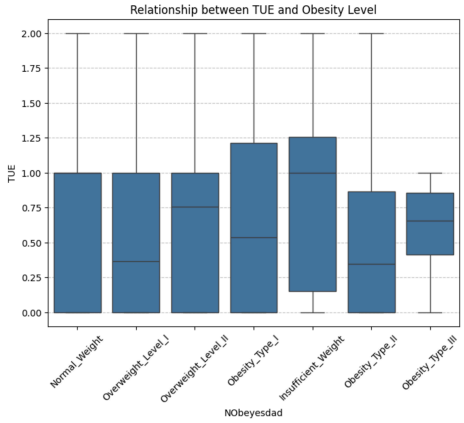


The chart shows that **Frequency of Physical Activity (FAF)** is related to weight levels. The Normal Weight group has a higher activity frequency, with a median close to 2 and an even distribution, suggesting regular physical activity helps maintain normal weight. In contrast, the Obesity_Type_II and Obesity_Type_III (severe obesity) groups have a lower median FAF and a wider range, indicating lower and more varied activity levels.

The Insufficient Weight group also has a high median FAF, but the distribution is wider, likely due to other factors causing underweight. Obese individuals, especially in severe obesity categories, may have limited activity due to health issues, worsening their condition.
Overall, low physical activity frequency may be an important factor contributing to obesity.

3.4.2.4 TUE - Time Using Technology Devices

Figure (1.2.4) shows that while daily electronic device usage time (TUE) varies across different weight groups, there doesn't appear to be a clear direct relationship with obesity.



The medians for the Normal Weight, Overweight_Level_I, and Obesity_Type_I groups are all around 1 hour, with similar distribution ranges. For the Obesity_Type_II and Obesity_Type_III groups, the median is slightly lower, at about 0.75 hours. However, this difference may be more influenced by individual lifestyle habits or other factors rather than being directly caused by obesity.

3.4.2.5 Chai and ANOVA Validation Result

```
Chi-Square Test for SCC:
Chi2 = 123.02389868912441, p-value = 3.773175792377203e-24, Degrees of Freedom = 6
------------------------------
Chi-Square Test for MTRANS:
Chi2 = 292.59394813167995, p-value = 5.177915203835779e-48, Degrees of Freedom = 24
------------------------------
```

SCC (Calorie Consumption Monitoring):

Chi2 = 123.02, p-value = 3.77e-24 (very significant).

There is a clear link between calorie monitoring and obesity. From the picture, most obese people, especially in the Obesity_Type_III group, do not monitor their calories. This could be due to low awareness of weight management or lifestyle habits. Calorie monitoring might play an important role in managing obesity.

MTRANS (Transportation Used):

Chi2 = 292.59, p-value = 5.18e-48 (very significant).

Different transportation modes are closely related to obesity. The picture shows that most people in the Obesity_Type_III group use public transportation while walking or cycling is rare. The way people travel may affect their physical activity levels, which can influence obesity.

```
ANOVA Test for FAF:
F-Statistic = 17.4842004293805, p-value = 7.653252995698972e-20
------------------------------
ANOVA Test for TUE:
F-Statistic = 7.876655737080669, p-value = 2.0687816228130554e-08
------------------------------
```

FAF (Frequency of Physical Activity):

F-Statistic = 17.48, p-value = 7.65e-20 (very significant).

The frequency of physical activity varies noticeably across different obesity levels. From the previous figure, it's clear that people with Normal Weight are more physically active compared to obese individuals. In particular, Obesity_Type_II and Obesity_Type_III show the lowest levels of activity. This suggests that low physical activity frequency could be a key factor contributing to obesity.

TUE (Time Using Technology Devices):

F-Statistic = 7.88, p-value = 2.07e-8 (significant).

The data shows some differences in electronic device usage time across obesity levels, with Obesity_Type_II and Obesity_Type_III groups spending less time on devices compared to Normal Weight and mildly obese groups. Although statistical analysis indicates a correlation, the figures do not show a clear or strong practical association between electronic device use and obesity. This suggests that other lifestyle factors may play a more significant role in influencing obesity levels.

3.5 Linear Regression

We would like to know whether family history (genetics) affects obesity by using Linear Regression and a scatter plot to gain insights.

Since **NObeyesdad** represents a range of BMI values, we first create a column named BMI and then calculate the corresponding values. Second, for the **family_history_with_overweight** part, we convert the data into binary (1 for 'yes', 0 for 'no') format. After splitting the data into training and testing sets, we built the linear regression model and visualized the result by using a scatter plot.

The graph on the right side shows the output of an OLS Regression. The BMI is the dependent variable, and family_history_encoded (binary variable) is the independent variable.

P-value (0.000): $P < 0.05$ indicates that the relationship between family history and BMI is statistically significant.

The scatter plot below shows the linear relationship between **family_history_with_overweight** and **BMI.** The red line represents the regression line. The regression line slopes up, indicating a positive relationship between having a family history of overweight and a higher BMI.

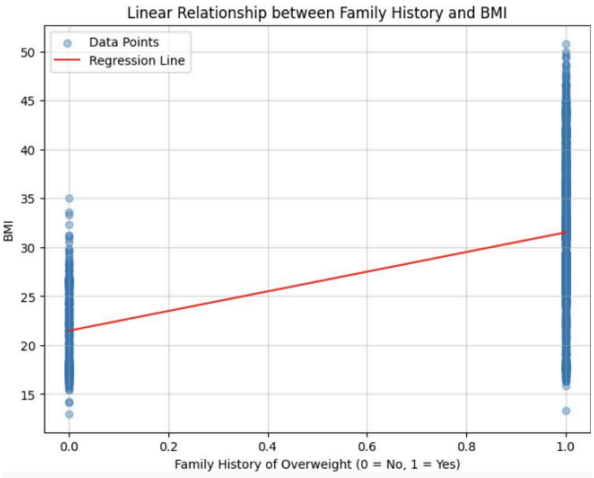
OLS Regression Results

Dep. Variable:	BMI	R-squared:	0.234
Model:	OLS	Adj. R-squared:	0.234
Method:	Least Squares	F-statistic:	516.3
Date:	Thu, 28 Nov 2024	Prob (F-statistic):	6.34e-100
Time:	23:52:04	Log-Likelihood:	-5675.5
No. Observations:	1688	AIC:	1.135e+04
Df Residuals:	1686	BIC:	1.137e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.4811	0.400	53.700	0.000	20.697	22.266
family_history_encoded	10.0415	0.442	22.722	0.000	9.175	10.908

Omnibus:	24.700	Durbin-Watson:	2.100
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14.816
Skew:	0.020	Prob(JB):	0.000606
Kurtosis:	2.543	Cond. No.	4.51

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



4. Takeaways

A. Can machine learning (ML) techniques be used to predict the risk of developing obesity?

Yes, machine learning techniques can effectively predict obesity risk. The Decision Tree model achieved 94.16% accuracy, while Random Forest reached 92.59%, successfully classifying individuals into obesity categories, with Obesity_Type_III showing 100% recall. K-Nearest Neighbors (KNN) was used to identify key factors contributing to obesity through clustering, and Logistic Regression analyzed how eating habits and physical activity impact obesity. Linear Regression showed a clear positive link between family history and BMI. Although the Decision Tree model showed some overfitting, these methods provide reliable tools for predicting obesity risk and understanding its key factors.

B. What are the most significant factors of obesity, and how can they be addressed?

From 3.3 Based on the classification results from the Cluster KNN algorithm across 15 clusters, with 5 clusters showing distinct characteristics selected for analysis, we can conclude the following:

Key factors influencing obesity include time spent using electronic devices (TUE), **Transportation used Public Transportation (MTRANS_Public_Transportation)**, the **frequency of daily meals (NCP)**, **vegetable intake frequency (FCVC)**, and **water intake**

(CH2O). These factors are consistently significant in predicting obesity across all groups and represent common points contributing to different obesity levels.
However, it is important to note that electronic device usage (TUE) is prevalent in the low-weight group (Cluster 2 and Cluster 9), indicating that while these factors are shared, their impact on weight outcomes may vary. Given that the survey primarily includes individuals aged 20-30, where electronic device usage is common, this factor cannot be directly linked to obesity and will not be used as a reference in our analysis.

The classification of the clusters shows a clear connection to obesity types: Cluster 0 is primarily associated with Obesity Type I, making up nearly 60% of the group. This cluster is characterized by both no alcohol consumption (CALC_no) and frequent drinking (CALC_Frequently), highlighting that drinking behavior and demographic factors play a key role. Cluster 1 is mostly concentrated in Obesity Type III, accounting for almost 90%. Despite having high-frequency physical activity (FAF) as a characteristic, occasional drinking (CALC_Sometimes) and other lifestyle habits significantly contribute to severe obesity. Cluster 6 mainly reflects Obesity Type II, with about 44% of individuals in this category. This cluster is defined by high-frequency physical activity (FAF) and both frequent and occasional snack intake (CAEC_Frequently, CAEC_Sometimes), showing how snack frequency and physical activity interact and impact weight management. These patterns illustrate the varied influence of lifestyle and dietary habits on different obesity types.

The populations of Cluster 2 and Cluster 9 were mainly normal weight or underweight, with 154 of 175 people (about 88%) in Cluster 2 and 131 of 151 people (about 86%) in Cluster 9. The main feature of Cluster 2 was **no alcohol consumption (CALC_no)**, which may help maintain a lower weight by reducing extra caloric intake; Cluster 9 was characterized by high frequency of physical activity (FAF) and occasional alcohol consumption (CALC_Sometimes), which combined helped maintain a normal weight by burning calories and moderately controlling alcohol caloric intake. These lifestyles were key factors in their maintenance of normal or underweight.

C. How do eating habits and physical activity affect the risk of obesity?

From 3.4 Eating habits and physical activity have a clear impact on obesity. Eating high-calorie foods (FAVC) is strongly linked to obesity, but those with lighter weight may rely on factors like faster metabolism and staying active (FAF). Most obese individuals, especially in Obesity_Type_III, rarely monitor their calorie intake (SCC). Additionally, walking or cycling as transportation is uncommon in all groups, reducing opportunities for physical activity. Keeping regular meal routines (NCP) and staying physically active are important for maintaining a healthy weight, showing how daily habits influence obesity.

D. What role do genetic and family factors play in obesity predisposition?

From 3.5 the results clearly show a strong linear relationship between family history of overweight and BMI. Individuals with a family history of overweight (family_history_encoded = 1) tend to have a BMI that is, on average, 10.04 units higher compared to those without such a history. This significant difference, supported by a very low p-value (0.000), highlights how family history strongly influences BMI, making it an important factor in understanding obesity risk.

5. Conclusion

This study shows that machine learning (ML) techniques can effectively predict obesity risk and identify factors that contribute to it. Models such as Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Linear Regression were used to analyze obesity from different perspectives, including risk prediction, key influencing factors, and the roles of genetics, eating habits, and physical activity.

The findings reveal a strong link between frequent high-calorie food consumption and obesity, the importance of regular meal patterns and physical activity for maintaining a healthy weight,

and the significant impact of family history on BMI. Clustering analysis highlighted differences across weight categories, showing how avoiding alcohol and staying active help lighter-weight groups, while irregular eating and snacking contribute to obesity in others.

In conclusion, obesity is influenced by multiple factors, including diet, physical activity, genetics, and lifestyle habits. Machine learning not only predicts obesity risk with high accuracy but also provides practical insights for interventions. These results highlight the need for personalized strategies that combine healthy eating, regular exercise, and awareness of genetic risks to better manage and prevent obesity.

6. References

1. Controlling the global obesity epidemic. *World Health Organization*. June 2022. https://www.who.int/activities/controlling-the-global-obesity-epidemic

2. Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. 2019. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5H31Z.

3. Adult BMI Categories. *Centers for Disease Control and Prevention website*, March 19, 2024, https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html