Project Title: **Estimation of Obesity Levels Using Machine Learning**

**Data Analysis and Visualization  
Semester Project Proposal**

**Session *2022-2026***



Department of Computer Science

Namal University Mianwali

01 June 2025

**Group Members:**

|  |  |
| --- | --- |
| Muhammad Asim Khan | NUM-BSCS-2022-14 |
| Aqeel Abbas Khan | NUM-BSCS-2022-38 |

**Submitted to:** Mr. Bilal

Table of Contents

[1. Introduction 4](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688944)

[2. Preprocessing: 7](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688945)

[2.1. Removing Null Values: 7](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688946)

[2.2. Dropping Duplicated Values: 7](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688947)

[2.3. Removing Outliers: 7](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688948)

[3. EDA: 8](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688949)

[3.1. Correlations between Features: 8](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688950)

[3.2. Relationship Between Different Features and Obesity of Full Dataset: 9](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688954)

[3.3. Male’s Data Visualization and Analysis: 16](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688955)

[3.4. Female’s Data Visualization and Analysis: 21](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688956)

[4. Model Training: 26](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688957)

[4.1. Decision Tree: 26](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688958)

[4.2. KNN: 27](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688959)

[4.3. Logistic Regression: 29](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688960)

[4.4. Random Forest: 30](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688961)

[4.5. Support Vector Machine: 32](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688962)

[5. Summary of Results’ Accuracy: 33](file:///C:\Users\lenovo\Desktop\report.docx#_Toc199688963)

# Introduction

Obesity is a critical global health challenge, with the World Health Organization reporting that over 1 billion people were living with obesity in 2022 ([WHO Obesity Factsheet](https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight)). It is linked to chronic diseases like diabetes and heart disease, making it essential to understand its contributing factors and predict risk for early intervention. This project proposes using a publicly available dataset to analyze obesity levels through machine learning, aiming to identify key predictors and develop accurate predictive models. By exploring demographic, dietary, and lifestyle factors, the study seeks to provide insights that could inform public health strategies and personalized interventions.

**Dataset Source and Description**

The dataset, titled "Estimation of Obesity Levels Based On Eating Habits and Physical Condition," is sourced from the UCI Machine Learning Repository ([UCI Dataset](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)). It comprises 2111 records and 17 attributes, collected from individuals in Mexico, Peru, and Colombia. The attributes include:

| **Attribute** | **Type** | **Description** |
| --- | --- | --- |
| Gender | Categorical | Male or Female |
| Age | Continuous | Age in years |
| Height | Continuous | Height in meters |
| Weight | Continuous | Weight in kilograms |
| family\_history\_with\_overweight | Binary | Family history of overweight (yes/no) |
| FAVC | Binary | Frequent consumption of high caloric food (yes/no) |
| FCVC | Integer | Frequency of vegetable consumption in meals |
| NCP | Continuous | Number of main meals daily |
| CAEC | Categorical | Consumption of food between meals |
| SMOKE | Binary | Smoking status (yes/no) |
| CH2O | Continuous | Daily water intake in liters |
| SCC | Binary | Monitoring of calorie consumption (yes/no) |
| FAF | Continuous | Frequency of physical activity |
| TUE | Integer | Time using technological devices in hours |
| CALC | Categorical | Alcohol consumption frequency |
| MTRANS | Categorical | Mode of transportation (e.g., walking, public transport) |
| NObeyesdad | Categorical | Obesity level (target variable: Insufficient Weight, Normal Weight, Overweight Level I/II, Obesity Type I/II/III) |

The dataset includes no missing values, with 77% of the data synthetically generated using the Weka tool and SMOTE filter, and 23% collected directly from users via a web platform ([Dataset Paper](https://www.sciencedirect.com/science/article/pii/S2352340919306985)). This composition supports robust analysis but requires caution in generalizing findings.

**Research Questions:**

* What are the most significant lifestyle factors that contribute to higher levels of obesity?
* How effective are combined lifestyle interventions (e.g., diet and physical activity) in reducing the predicted obesity level compared to single interventions?
* How does the distribution of obesity levels vary across demographic groups (e.g., age, gender)?
* Can machine learning models accurately classify individuals into different obesity levels based on their health and lifestyle attributes?

**Proposed Methodology**

The project will follow a structured approach to analyze the dataset and answer the research questions:

* **Data Preprocessing:** Categorical variables will be encoded, and numerical features (e.g., Age, Height) will be standardized to ensure model compatibility.
* **Exploratory Data Analysis (EDA):** Initial analysis will examine feature distributions, correlations, and relationships with obesity levels to identify patterns and potential issues like multicollinearity.
* **Model Development:** Multiple machine learning models will be implemented for multi-class classification to choose the best fit model.

**Potential Challenges and Mitigation Strategies**

Several challenges may arise during the project, each with a planned mitigation strategy:

| **Challenge** | **Description** | **Mitigation Strategy** |
| --- | --- | --- |
| Synthetic Data | 77% of the data is synthetic, potentially limiting real-world applicability. | Focus on dataset-specific patterns and acknowledge generalizability limitations in findings. |
| Class Imbalance | Uneven distribution of obesity levels may skew model performance. | Apply class weighting to balance classes during model training. |
| Feature Correlation | High correlation between features may cause multicollinearity. | Conduct correlation analysis and use feature selection to reduce redundant features. |

Additional considerations include the dataset’s regional focus (Mexico, Peru, Colombia), which may limit applicability to other populations, and ethical concerns around obesity-related stigmatization, which will be addressed by framing results sensitively.

**Expected Outcomes**

The project anticipates the following outcomes:

* **Identification of Key Factors:** Determining which demographic, dietary, or lifestyle factors most strongly predict obesity levels, potentially highlighting actionable areas like physical activity or dietary habits.
* **Accurate Predictive Models:** Developing machine learning models with high accuracy, comparable to studies achieving good accuracy.
* **Demographic Insights:** Uncovering variations in obesity levels by gender or age, informing targeted public health interventions.
* **Public Health Implications:** Providing insights that could guide personalized health recommendations and broader obesity prevention strategies.

These outcomes aim to contribute to the growing body of research on obesity prediction and support evidence-based health policies.

# Preprocessing:

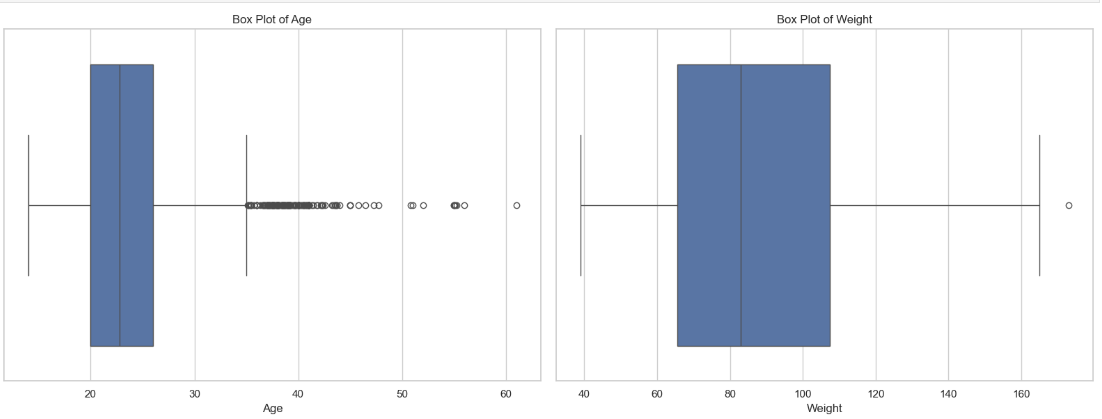
## Removing Null Values:

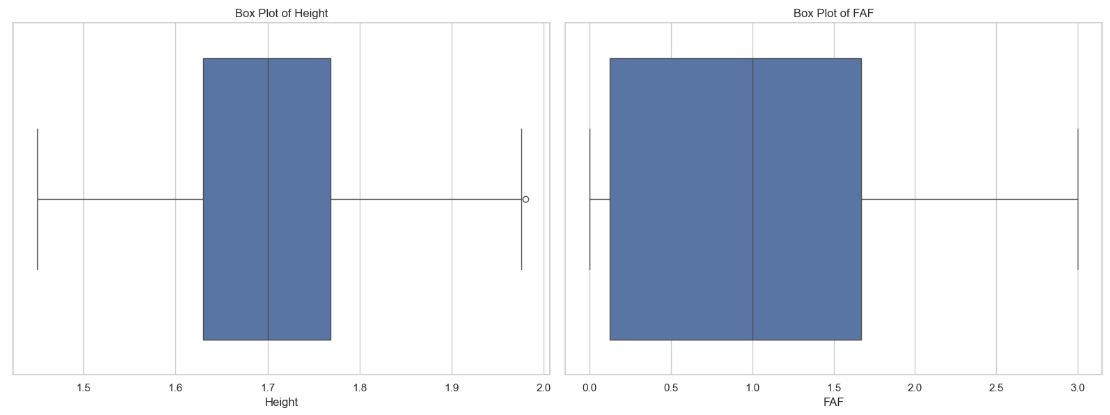
In this step we have to remove the null values but dataset under process has no null values. Therefore, it will remain same as it is.

## 2.2. Dropping Duplicated Values:

In this step we will remove the duplicated values of the dataset. In our dataset there are 24 duplicated values. Therefore, we have removed those values.

## Removing Outliers:

Let’s plot the box plot for finding outliers:  




In the above box plot it can be analyzed clearly that age contain too much outliers. Therefore, we will remove these outliers for best results.

# EDA:

In this methodology we will first visualize different features of data set through different plots and then we will analyze the dataset using these plots. After analyzing this dataset, we will be able to fir the suitable models for best accuracy.

## 3.1. Correlations between Features:

Let try to analyze the correlation of different other feature:

## Highly Correlated Feature Pairs:

1. **Height and Weight** (Correlation = **+0.46**)

Taller individuals tend to have higher body weight.

1. **Height and Physical Activity Frequency** (Correlation = **+0.29**)

Taller individuals tend to engage more in physical activities.

1. **Height and Number of Main Meals Per Day** (Correlation = **+0.24**)

Taller individuals may have slightly more main meals daily.

1. **Weight and Vegetable Consumption Frequency** (Correlation = **+0.21**)

Heavier individuals tend to eat vegetables slightly more frequently.

1. **(CH2O) Water Intake and Physical Activity Frequency** (Correlation = **+0.16**)

Individuals who drink more water are somewhat more physically active.

## Most Negative (Inverse) Correlated Feature:

1. **Age and Technology Usage Time** (Correlation = **-0.29**)

Older individuals spend less time using technological devices compared to younger individuals.

## Weakest (Almost No) Correlations:

1. **Age and Height** (Correlation = **-0.02**)

No meaningful relationship between a person's age and their height.

1. **Vegetable Consumption and Physical Activity** (Correlation = **+0.02**)

No strong relation between vegetable eating habits and physical activity frequency.

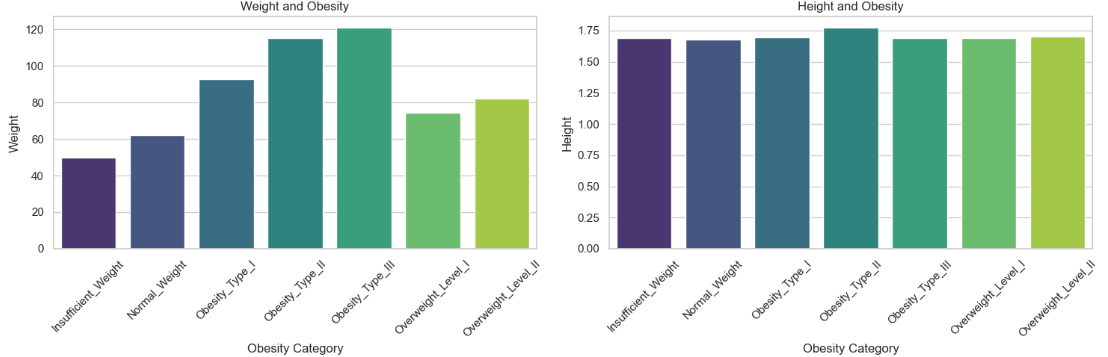
1. **Number of Meals and Technology Usage** (Correlation = **+0.04**)

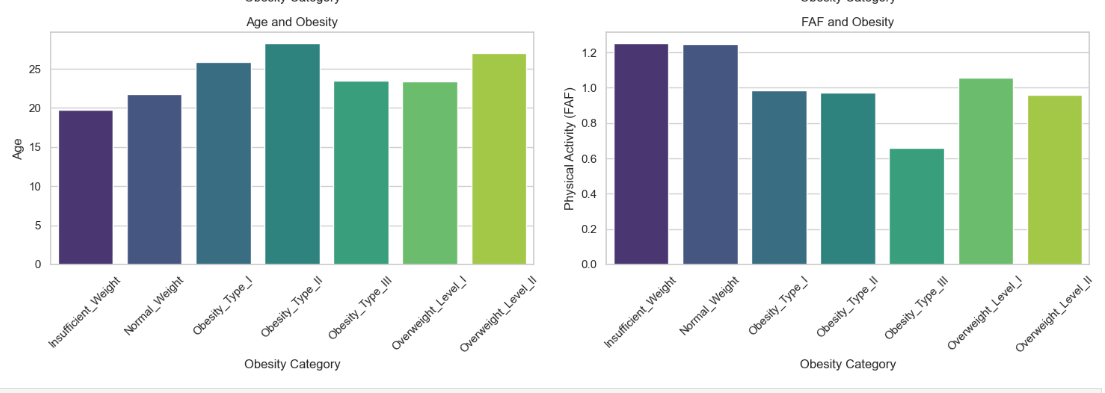
No strong relationship between how often someone eats and their technology usage time.

## 3.2. Relationship Between Different Features and Obesity of Full Dataset:

In this section we will try to analyze the relation and effect of other features on target variable.

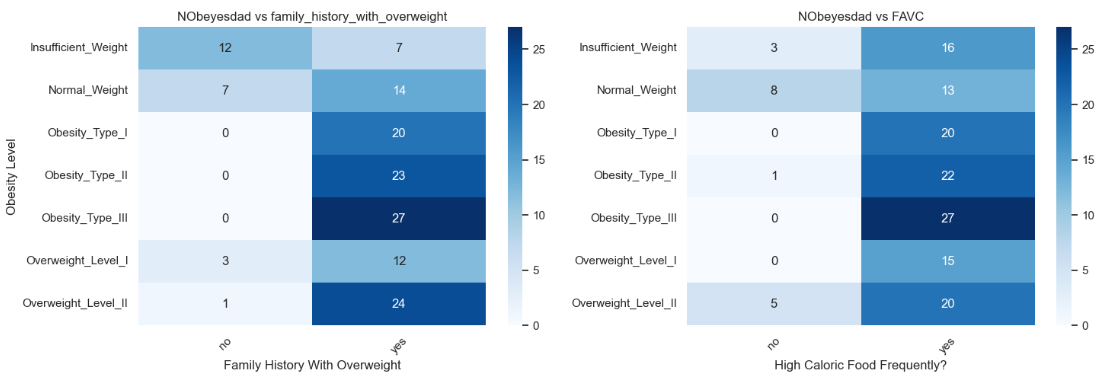
For that first we will plot bar graph on different features to analyze the impact of those features on obesity.

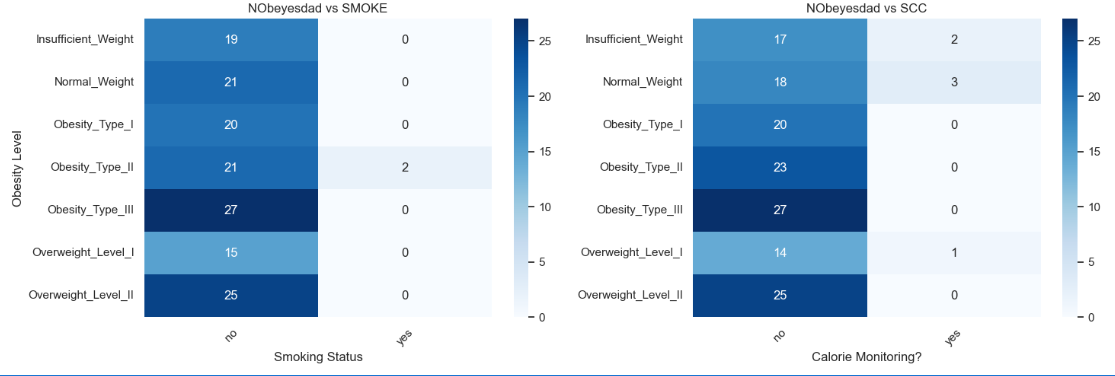




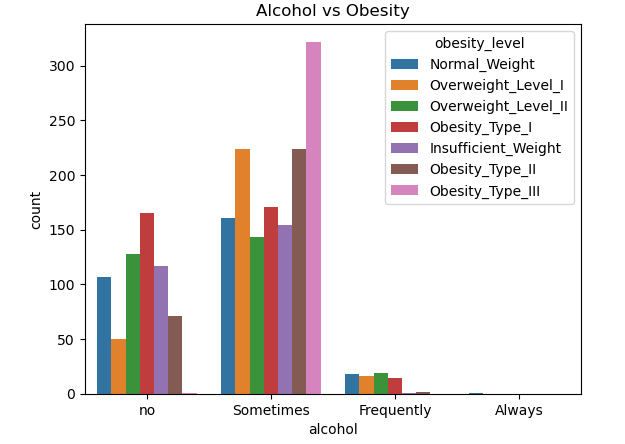
1. **Weight and Obesity**  
   As obesity category increases from **Insufficient Weight** to **Obesity\_Type\_III**, weight also increases significantly. Weight then decreases again slightly in **Overweight\_Level\_I** and **II**, but remains higher than normal weight categories.
2. **Height and Obesity**  
   Height remains relatively constant across all obesity categories with only minor fluctuations. There is no strong relationship between height and obesity in this dataset.
3. **Age and Obesity**  
   Age tends to increase with higher obesity categories up to **Obesity\_Type\_II**, but decreases slightly in **Obesity\_Type\_III** and **Overweight\_Level\_I** before rising again in **Overweight\_Level\_II**. Overall, higher obesity levels are associated with middle to older age groups.
4. **FAF (Physical Activity) and Obesity**  
   Higher physical activity levels (FAF) are associated with **Insufficient Weight** and **Normal Weight**. As obesity level increases, physical activity generally decreases, showing lower activity in **Obesity\_Type\_II** and **III**, but a slight increase again in **Overweight\_Level\_I**.

Now we will analyze the binary features to check their relation with obesity.





1. **NObeyesdad vs family\_history\_with\_overweight**  
   Individuals with a family history of overweight ("yes") have higher counts in obesity categories, especially **Obesity\_Type\_III** and **Overweight\_Level\_II**. Lack of family history shows higher counts in lower obesity levels or insufficient weight.
2. **NObeyesdad vs FAVC (High Caloric Food Frequently)**  
   People who frequently consume high-calorie food ("yes") are concentrated more in higher obesity types (Obesity\_Type\_II, III). Those who do not consume high-calorie food regularly ("no") are more represented in Insufficient\_Weight and Obesity\_Type\_I.
3. **NObeyesdad vs SMOKE (Smoking Status)**  
   Both smokers and non-smokers show similar distribution across obesity levels, but non-smokers ("no") slightly dominate higher obesity categories (Obesity\_Type\_III, Overweight\_Level\_II). Smoking doesn't appear to drastically skew obesity levels.
4. **NObeyesdad vs SCC (Calorie Monitoring)**  
   Individuals who monitor their calories ("yes") show slightly higher counts in lower obesity levels and insufficient weight. Those who do not monitor calories ("no") have greater numbers in higher obesity types, especially Obesity\_Type\_III and Overweight\_Level\_II.

Now analyzing obesity vs Alcohol  
**Graph:**  


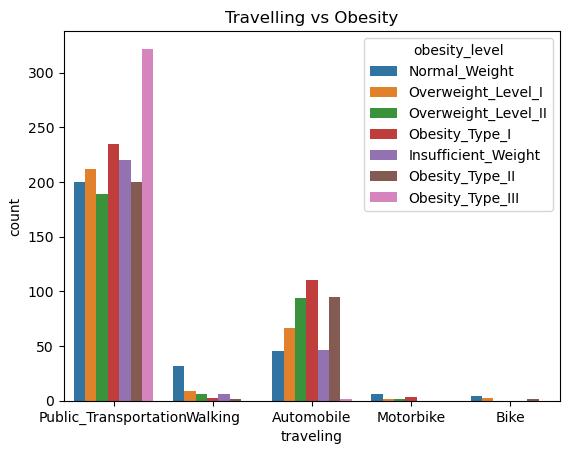
**Description:**

* The grouped bar chart visualizes the distribution of **obesity levels** across different **alcohol consumption categories**.
* The **x-axis** represents alcohol consumption categories (e.g., *low*, *medium*, *high*).
* The **y-axis** shows the count (frequency) of individuals within each alcohol category.
* Each bar group is divided by obesity categories (e.g., *non-obese*, *moderately obese*, *highly obese*), represented by different colors (legend).

**Observations**:

1. First it has been observed that most of the people are those who use alcohol sometime or do use it.
2. Those who not use lies in category of obesity i.e. insufficient, normal or obesity 1.
3. Those who use alcohol sometime lies in the category of obesity i.e. overweight, obesity level 2 and obesity level 3.

Now analyzing obesity vs Traveling  
**Graph:**



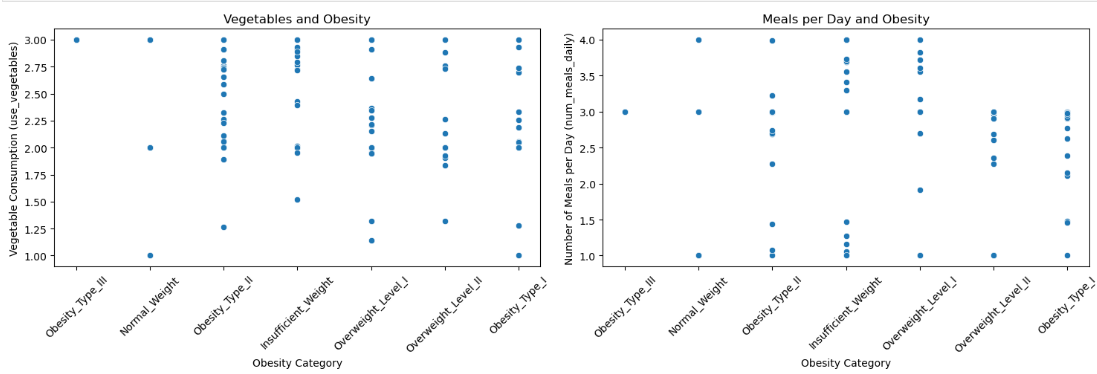
**Description:**

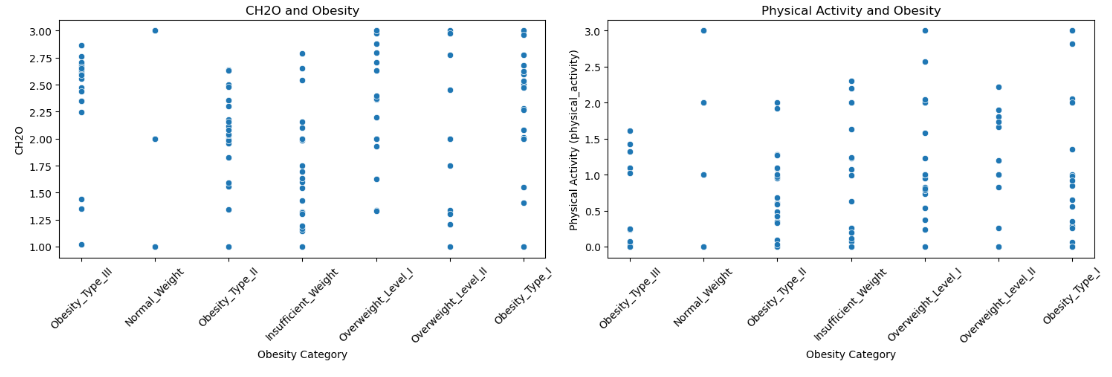
* The grouped bar chart visualizes the distribution of **obesity levels** across various **travel frequency categories**.
* The **x-axis** displays the categories of travel frequency (e.g., rarely travels, sometimes travels, frequently travels).
* The **y-axis** represents the count of individuals within each travel frequency category.
* Each bar group is subdivided into obesity categories (e.g., non-obese, moderately obese, highly obese), with different colors representing each obesity level as indicated in the legend.

**Observation**

* First observation is that this dataset who mostly uses public transport. Then contain those people who uses automobile and other catagories have very low data.
* Those people who uses public transport are effected by high obesity i.e. obesity 3 at large level.
* Those who walks have normal weight.
* Those who uses their auto mobile also lies between obesity level1 to overweight2. It means that automobile usage also effecting the obesity of people positively.

Now visualizing and observing features having float datatype;





**Description:**

The figure comprises **four scatter plots**, each displaying the distribution of different lifestyle factors across various **obesity categories**.

#### **1. Vegetables Consumption and Obesity**

* This plot illustrates the relationship between **vegetable consumption frequency** (on a 1–3 scale) and **obesity categories**.
* Vegetable consumption levels are fairly **evenly spread** across all obesity categories, indicating **no strong pattern** associating vegetable intake directly with obesity status.
* Individuals with **higher vegetable consumption (level 3)** appear across all obesity groups but people of obesity level 3 and overweight 1 are more then other categories.

#### **2. Meals per Day and Obesity**

* This plot shows the association between the **number of meals consumed per day** and **obesity categories**.
* Individuals consuming **3 or more meals per day** are present across overweight or insufficient weight.
* Conversely, individuals in the **normal weight** or **underweight** categories show **more variation**, with some consuming fewer meals (1–2 meals/day).

#### **3. CH2O (Carbohydrate Intake) and Obesity**

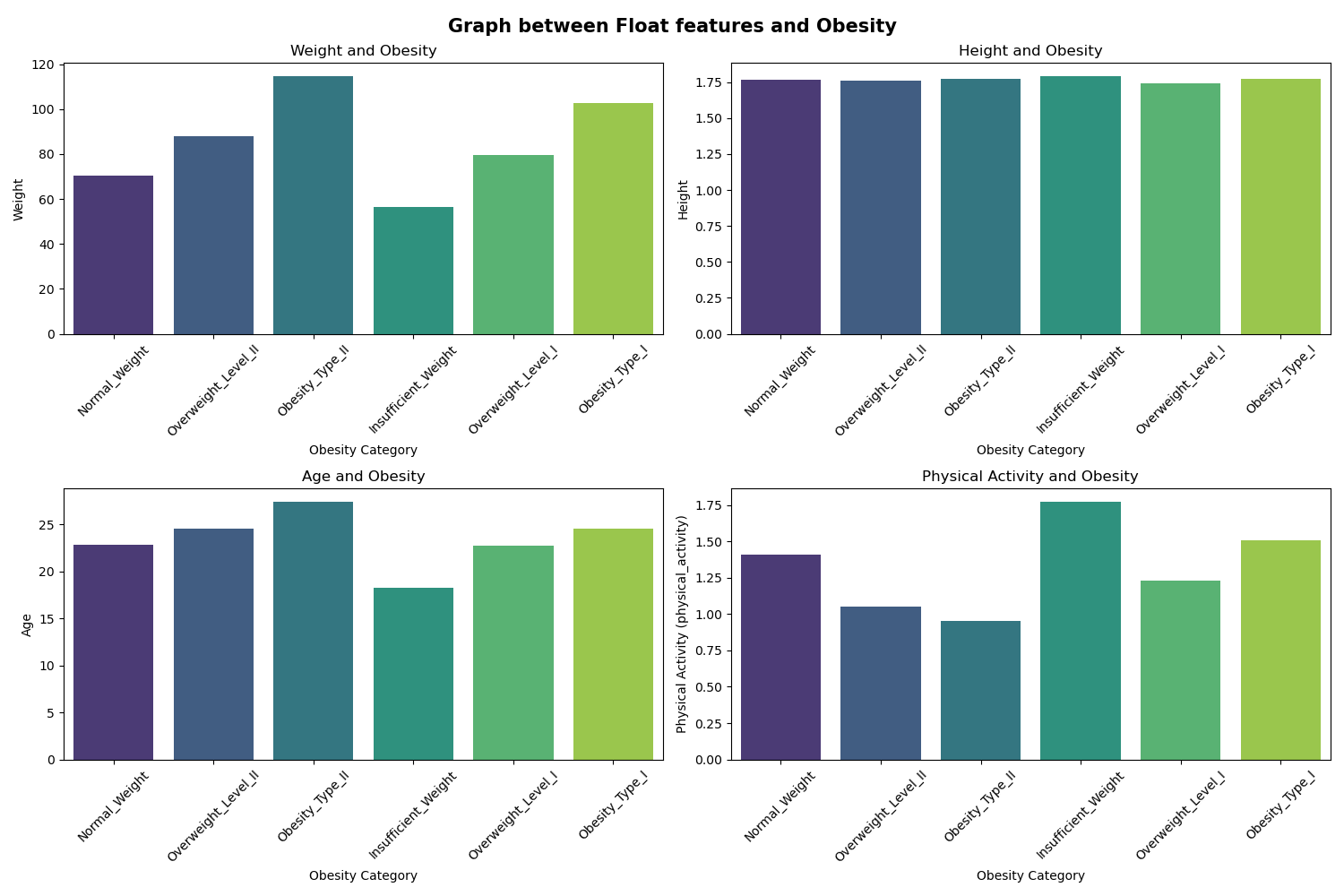
* This plot visualizes **carbohydrate consumption levels** (on a 1–3 scale) in relation to **obesity categories**.
* Carbohydrate intake appears **distributed across all obesity levels**, with **no clear concentration** of low or high carbohydrate intake in a specific obesity group.
* Similar to vegetable consumption, **carbohydrate intake** alone does not show an obvious pattern differentiating obesity categories.

#### **4. Physical Activity and Obesity**

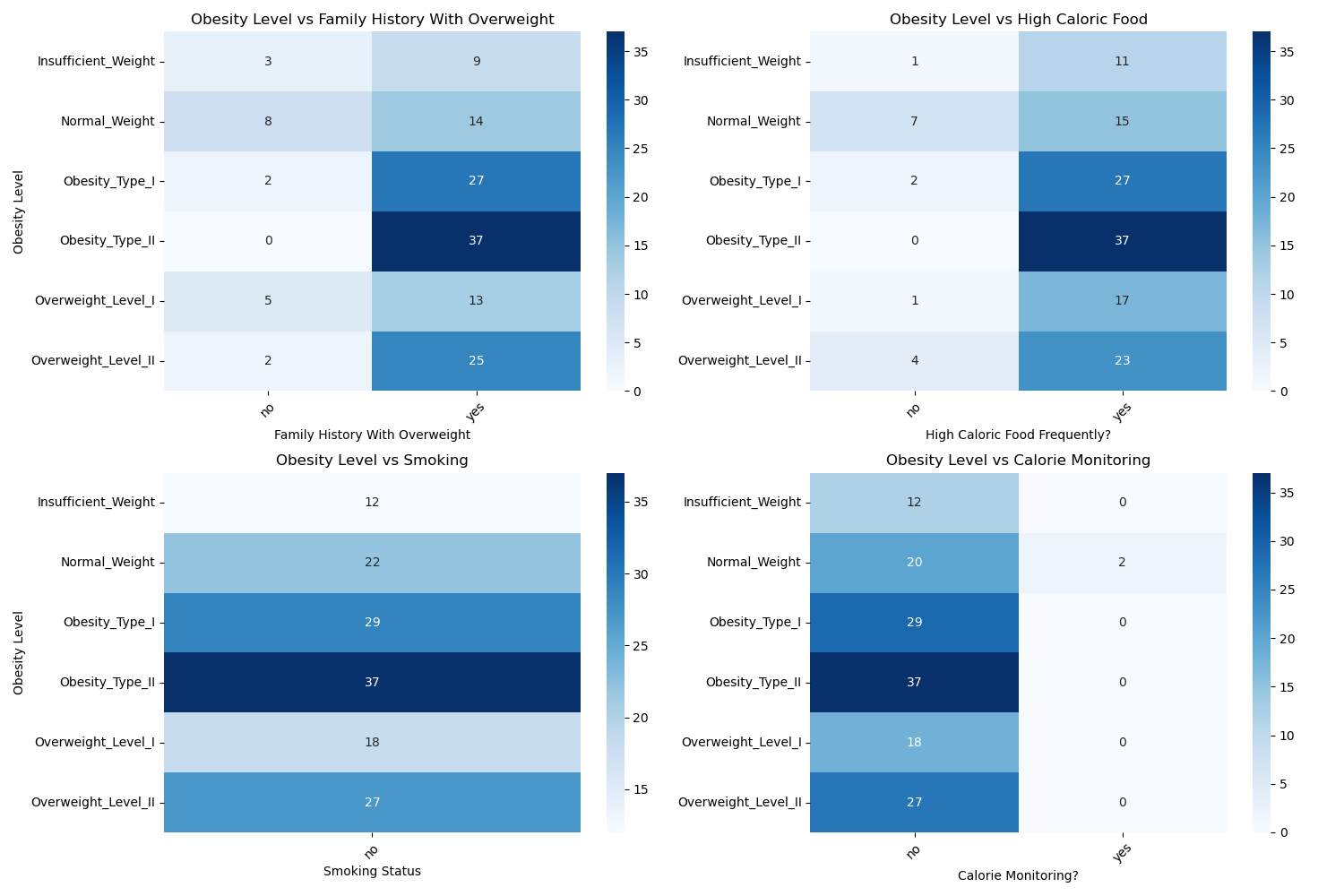
* This plot depicts the relationship between **physical activity frequency (days per week)** and **obesity categories**.
* Individuals engaging in **no or low physical activity (0–1 days/week)** are more commonly associated with **higher obesity levels** (Overweight and Obese categories).
* Conversely, individuals with **higher physical activity levels (2–3 days/week)** appear more frequently in the **normal weight** and **underweight** categories.
* This suggests a **negative association** between physical activity frequency and obesity — **higher physical activity** is linked with **lower obesity prevalence**.

## Male’s Data Visualization and Analysis:

Bar Graph between some floating features and obesity levels:

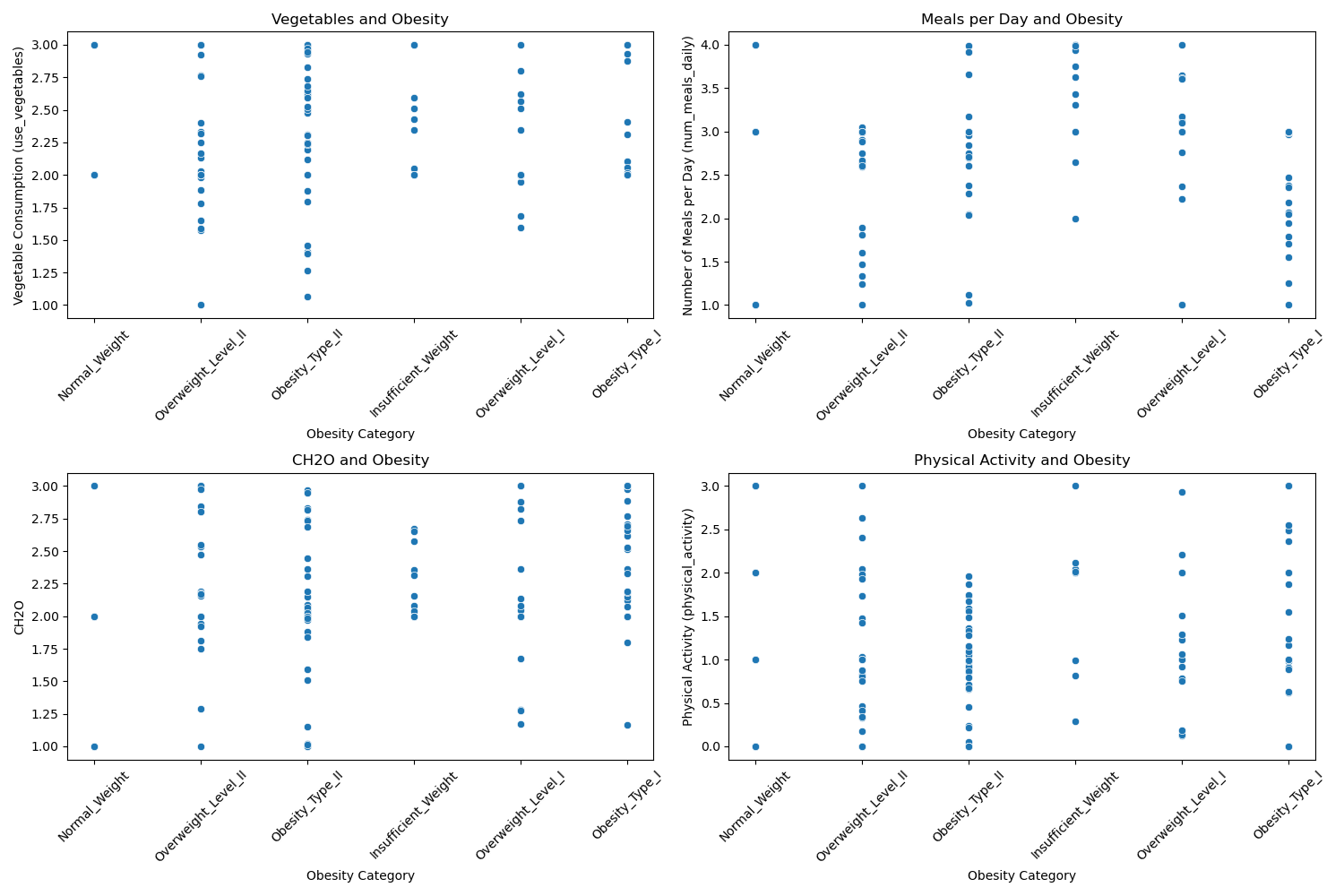
**Graph:**  
  
**Observations:**

* In weight and obesity graph it can be analyzed that people having weight more than 60 are the victim of the obesity or over weight.
* Height and obesity plot shows that there is no effective relation between height and obesity.
* Age and obesity graph clearly shows that people under 18 years have insufficient weight.
* Physical activity and obesity graph visualize that the people who do more physical activities are healthy i.e. insufficient weight or normal in weight.

**Heat plots between Binary Features and Obesity types feature:**  
  


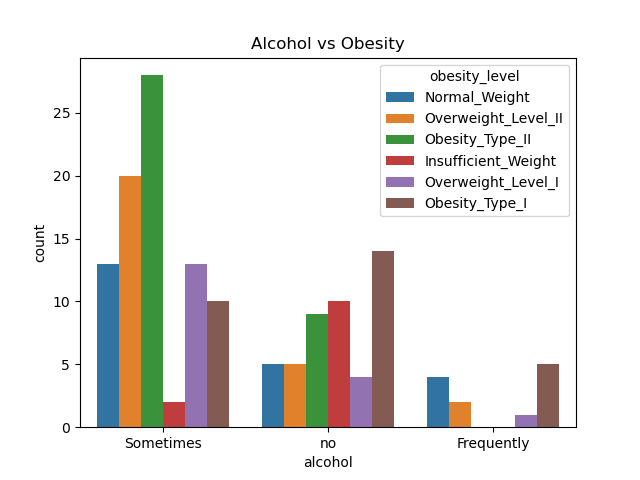
**Observation:**

* The plot of **obesity vs family history of overweight** initially shows that most of the people in this data set are the people who family history shows that their family have some over weighted people. Secondly, plot shows that people who family member has family history ‘yes’ are also victim of obesity or overweight.
* **Obesity vs High chloric food graph** shows that people who user high chloric foods are over weighted people or the victim of obesity.
* **Obesity vs Smoking** plot shows that people of this data set does not smoke.
* **Obesity vs Calories graph** shows that people who do not monitor calories are over weighted people or the victim of obesity.



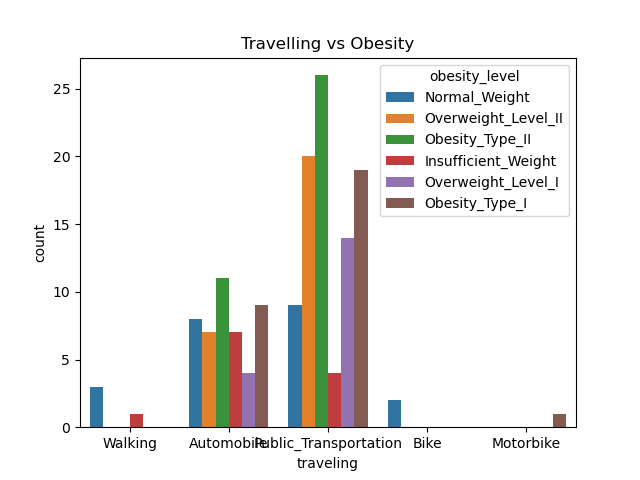
**Description:**

* **Vegetables and Obesity**: People in the higher obesity categories tend to have more variation in vegetable consumption, indicating inconsistent healthy eating habits among obese individuals.
* **Meals per Day and Obesity**: Individuals in obese categories (especially Obesity\_Type\_III) tend to have more meals per day, suggesting a possible link between frequent eating and obesity.
* **CH2O and Obesity**: The carbohydrate (CH2O) intake appears fairly distributed across all obesity levels, implying that CH2O intake alone may not directly correlate with obesity status in this dataset.
* **Physical Activity and Obesity**: Higher obesity levels are associated with a concentration of individuals having low physical activity, showing an inverse relationship between physical activity and obesity.



**Observations:**

* This graph shows that majority of people drinks alcohol some time or never. Some people uses alcohol frequently.
* People who drinks alcohol sometime have obesity type 2 or overweight. But people with insufficient weight are too less than other categories.
* people who never uses alcohol have insufficient weight or lies in obesity level 1 or 2.
* Frequently use of alcohol have no greater impact on the obesity.

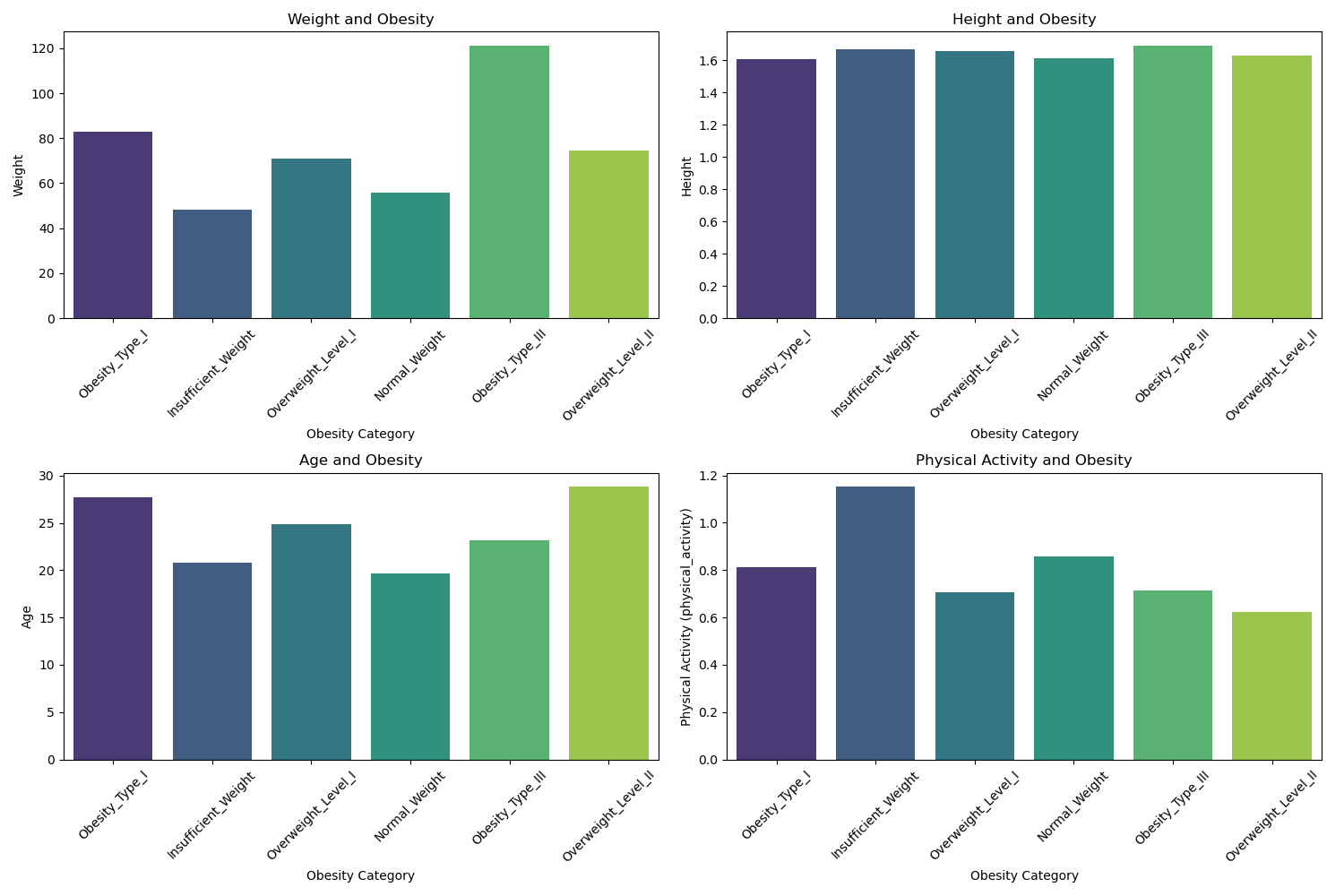


**Description:**

* People who walks are heathy means that they are normal in weight or insufficient weight.
* people who uses their automobiles are uses public transport are victim of obesity or overweight.
* People who uses their bikes are also normal in weight.

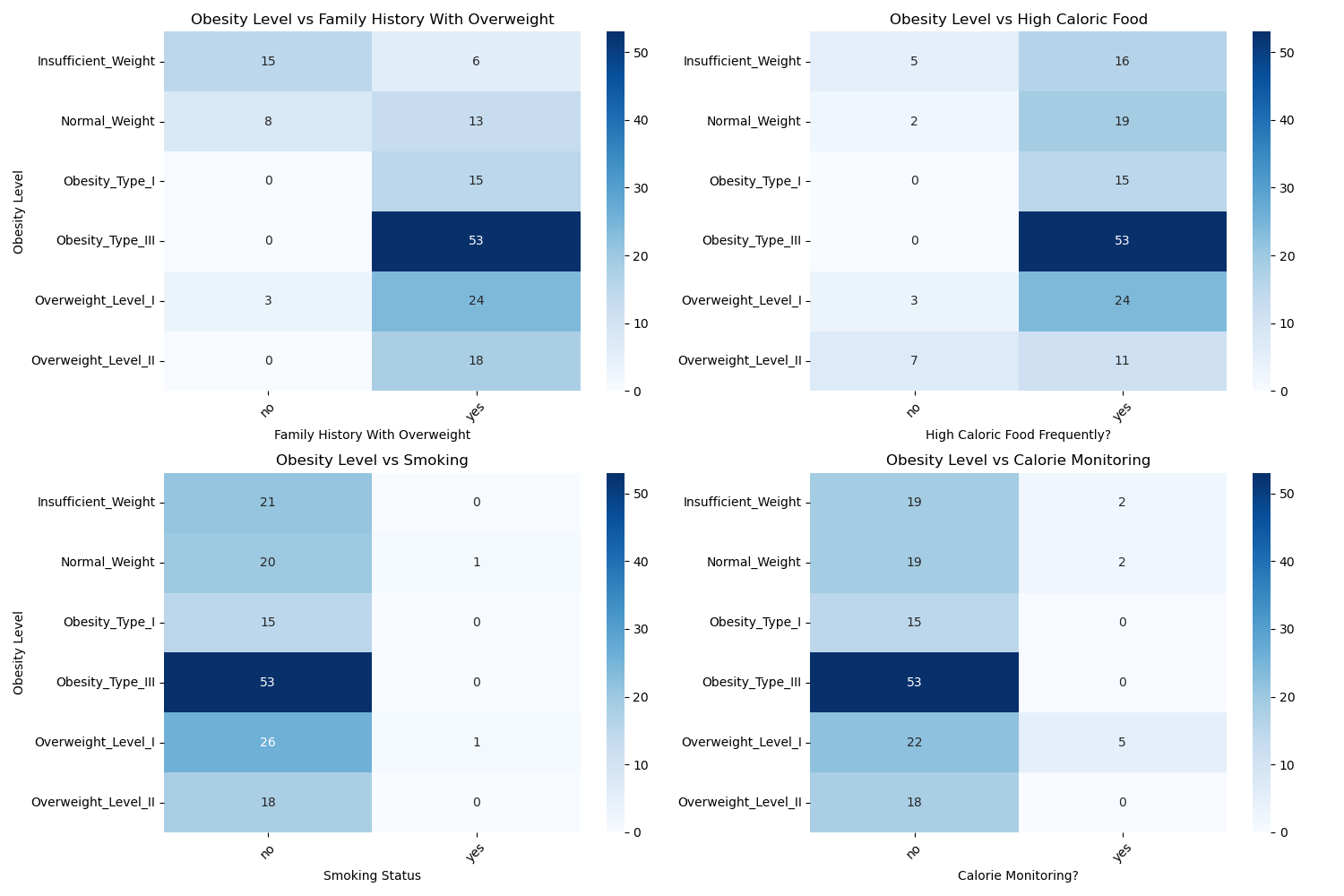
## Female’s Data Visualization and Analysis:

Bar Graph between some floating features and obesity levels:

****

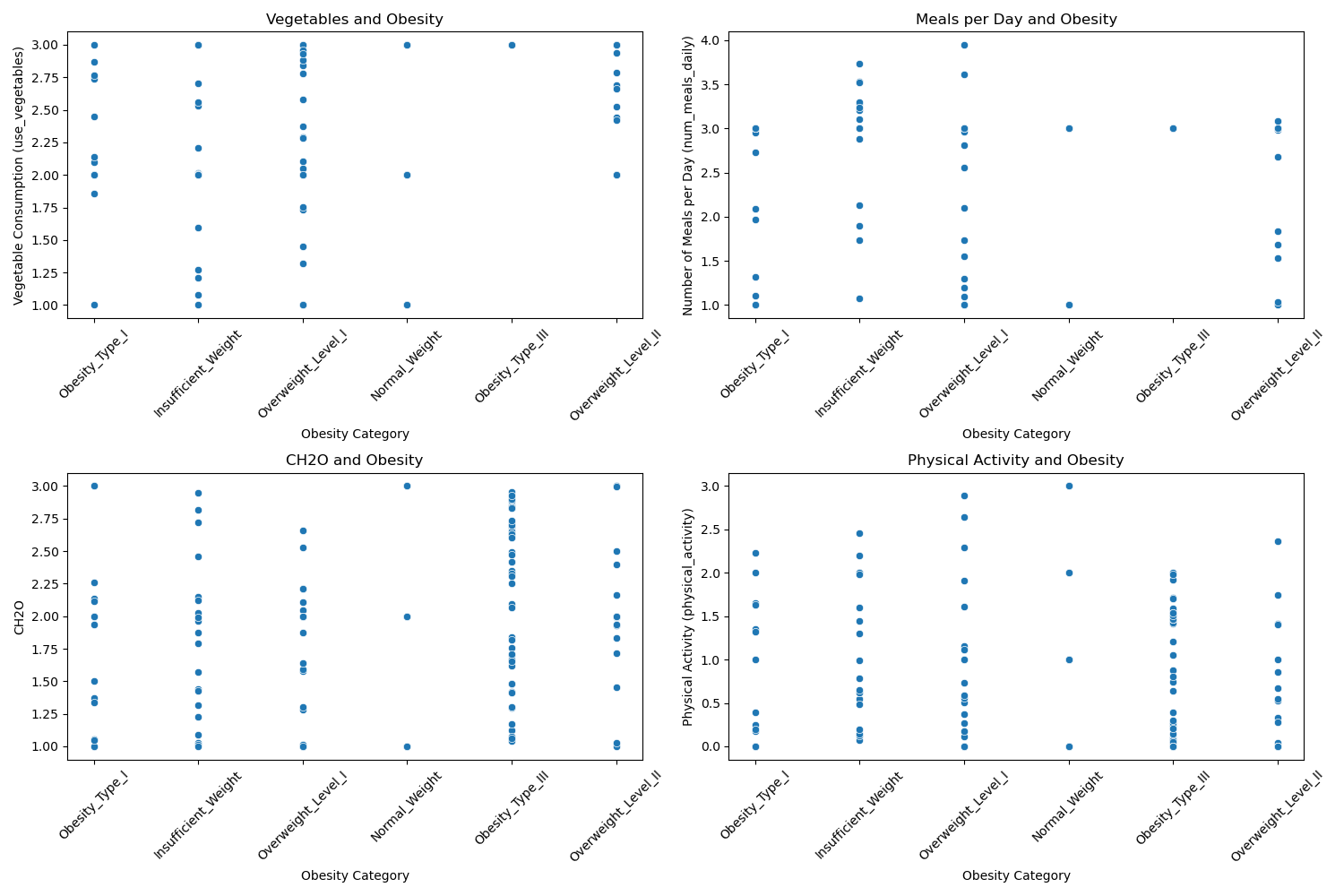
**Observations:**

* In **weight and obesity** graph it can be analyzed that females having weight more than 60 are the victim of the obesity or over weight.
* **Height and obesity** plot shows that there is no effective relation between height and obesity.
* **Age and obesity** graph clearly shows that females under 20 years have insufficient weight or normal weight.
* **Physical activity and obesity** graph visualize that the females who do more physical activities are healthy i.e. insufficient weight.

****

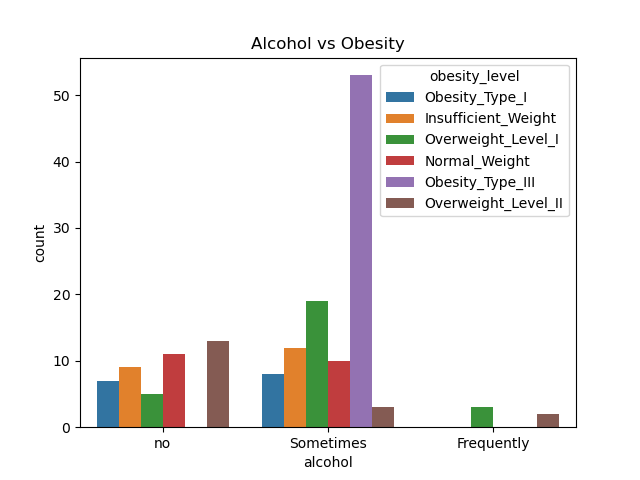
**Observations:**

* The plot of obesity vs family history of overweight initially shows that most of the females in this data set are the people who family history shows that their family have some over weighted females. Secondly, plot shows that females who family member has family history ‘yes’ are mostly victim of obesity level 3.
* Obesity vs High chloric food graph shows that females who uses high chloric foods are over weighted people or the victim of obesity level 3.
* Obesity vs Smoking plot shows that females who do not smoke mostly lies in the category of obesity level 3 and after that in overweight level 1.
* Obesity vs Calories graph shows that females who do not monitor calories are the victim of obesity level 3.

****

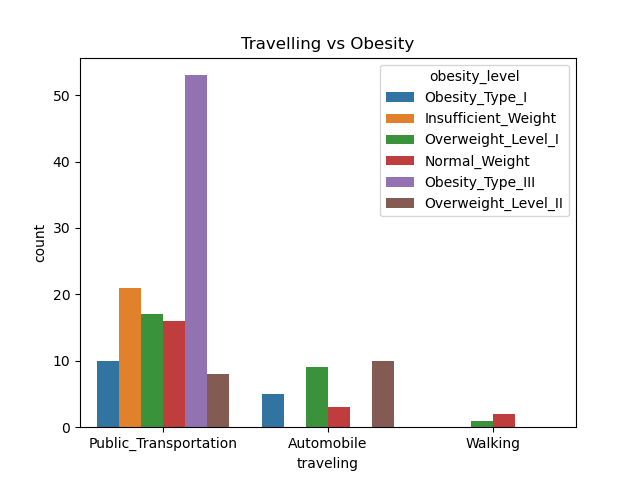
**Observations:**

* **Vegetables and Obesity**: There is no clear pattern indicating a strong relationship between the frequency of vegetable consumption and obesity category, as individuals across all obesity levels seem to consume vegetables at varying frequencies.
* **Meals per Day and Obesity**: People with higher obesity levels tend to have fewer meals per day, suggesting a possible negative correlation between number of meals and obesity severity.
* **CH2O and Obesity**: The amount of carbohydrate consumption (CH2O) appears to be spread across all obesity categories, with no obvious trend, indicating that CH2O intake may not be a strong standalone predictor of obesity.
* **Physical Activity and Obesity**: Individuals with lower obesity levels seem to engage more frequently in physical activity, while higher obesity levels correlate with lower physical activity, suggesting a negative relationship between physical activity and obesity.

****

**Observations:**

* Female who do not uses alcohol have no strong relation with obesity.
* Females who uses alcohol sometimes are the victim of obesity i.e. they are over weighted.
* Other categories of alcohol usage have no effective impact on obesity.

****

**Observations:**

* Most of the females use public transport.
* Public transport has a greater impact on the obesity of females. Mostly female who usepublic transport lies in the category of obesity level 3.
* Plot shows that over weighted females uses automobiles more than others.

# Model Training:

After analyzing the complete dataset, now turn comes for training models. Here we will train different models and then will finalize that which one is best one according to its accuracy. So let try different classification models:

## Decision Tree:

Decision Tree uses a tree-like structure to make decisions based on feature values. It is intuitive and interpretable. However, it may over fit if not properly pruned.

**Accuracy**:

The accuracy of this model is 96.99%

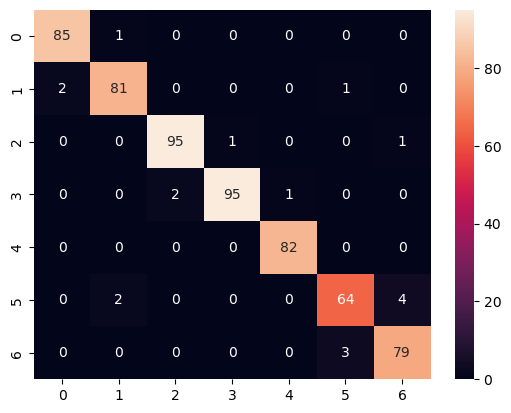
**Confusion Matrix:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 85 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 81 | 0 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 95 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 2 | 95 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 82 | 0 | 0 |
| 5 | 0 | 2 | 0 | 0 | 0 | 64 | 4 |
| 6 | 0 | 0 | 0 | 0 | 0 | 3 | 79 |

**Confusion Matrix Analysis:**

* Class 0: 85 correct, 1 misclassified.
* Class 1: 81 correct, 3 misclassified.
* Class 2: 95 correct, 2 misclassified.
* Class 3: 95 correct, 3 misclassified.
* Class 4: Perfect classification with 82 correct.
* Class 5: 64 correct, 6 misclassified.
* Class 6: 79 correct, 3 misclassified.

**Heat Map for Confusion Matrix:**



**Heat Map Analysis:**

Most values lie strongly along the diagonal, indicating excellent classification. Very few off-diagonal values suggest the model has high precision and recall.

## KNN:

KNN classifies samples based on the most common class among their nearest neighbors. It’s simple and works well with low-dimensional data.

**Accuracy:**

The accuracy of KNN model is 84.14%

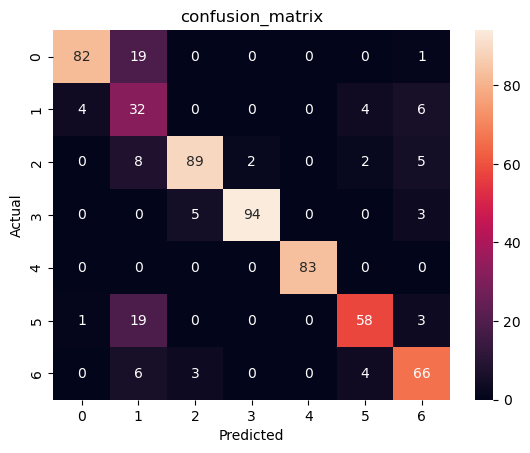
**Confusion matrix:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 82 | 19 | 0 | 0 | 0 | 0 | 1 |
| 1 | 4 | 32 | 0 | 0 | 0 | 4 | 6 |
| 2 | 0 | 8 | 89 | 2 | 0 | 2 | 5 |
| 3 | 0 | 0 | 5 | 94 | 0 | 0 | 3 |
| 4 | 0 | 0 | 0 | 0 | 83 | 0 | 0 |
| 5 | 1 | 19 | 0 | 0 | 0 | 58 | 3 |
| 6 | 0 | 6 | 3 | 0 | 0 | 4 | 66 |

**Confusion Matrix Analysis:**

* High misclassification in Class 0 and 1 (e.g., 19 misclassifications for Class 0).
* Class 5 and 6 show noticeable confusion with nearby classes.
* Class 4 is well classified (83 correct).

**Heat Map of Confusion Matrix:**



**Heat Map Analysis:**

Diagonal elements are still strong but more diluted. Off-diagonal cells like (0,1), (5,1), and (6,2) show misclassification, reflecting poorer class separation compared to other models.

## Logistic Regression:

Logistic Regression models linear boundaries between classes. It’s efficient and interpretable, but limited in handling non-linear patterns.

**Accuracy:**

The accuracy of logistic model is 89.97%

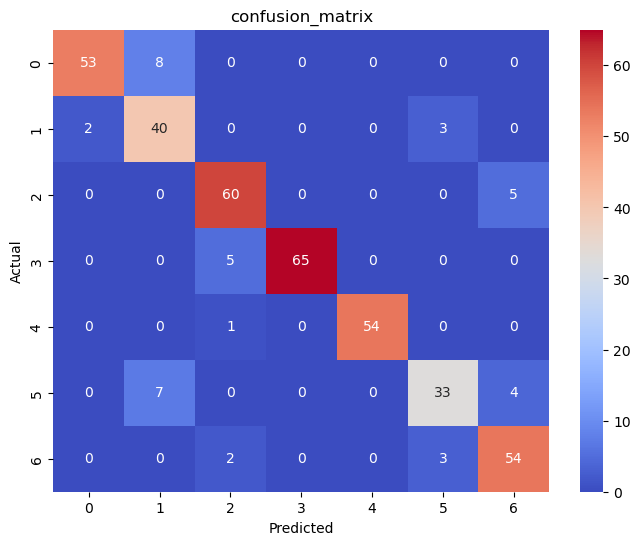
**Confusion Matrix:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 53 | 8 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 40 | 0 | 0 | 0 | 3 | 0 |
| 2 | 0 | 0 | 60 | 0 | 0 | 0 | 5 |
| 3 | 0 | 0 | 5 | 65 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 54 | 0 | 0 |
| 5 | 0 | 7 | 0 | 0 | 0 | 33 | 4 |
| 6 | 0 | 0 | 2 | 0 | 0 | 3 | 54 |

**Confusion Matrix Analysis:**

* Class 0 and 1 have moderate misclassifications.
* Class 2 and 3 are relatively well predicted.
* Class 5 and 6 show confusion with other classes (e.g., 7 misclassifications in Class 5).

**Heat Map of Confusion Matrix:**



**Heat Map Analysis:**

The diagonal dominates, but noticeable off-diagonal errors, particularly for Class 5 and Class 6, lower the overall performance compared to tree-based models.

## Random Forest:

Random Forest is an ensemble of decision trees. It reduces overfitting and increases generalization by aggregating predictions from multiple trees.

**Accuracy**

The accuracy rate of Random forest is 97.24%

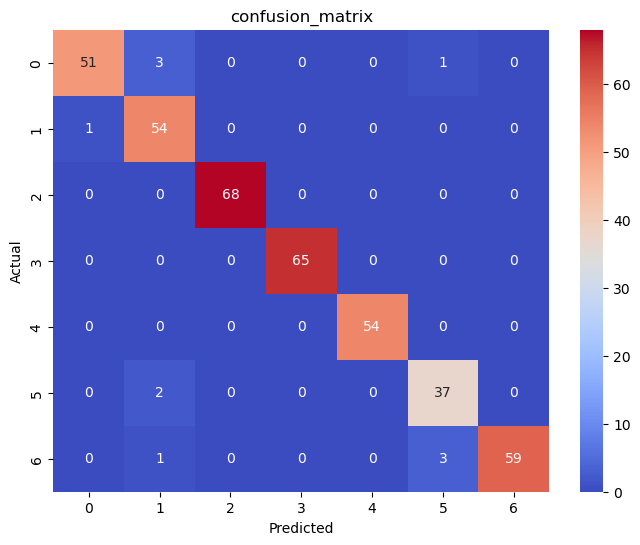
**Confusion Matrix:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 51 | 3 | 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 54 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 68 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 65 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 54 | 0 | 0 |
| 5 | 0 | 2 | 0 | 0 | 0 | 37 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 3 | 59 |

**Confusion Matrix Analysis:**

* Near-perfect classification across all classes.
* Minor confusion in Class 0 and 5.
* Class 2, 3, and 4 have perfect predictions.

**Heat Map of Confusion matrix:**



**Heat Map Analysis:**

The heat map is almost completely diagonal. Strong dark diagonal blocks indicate high confidence predictions. Very few misclassifications overall.

## Support Vector Machine:

SVM finds an optimal boundary (hyperplane) that separates classes with maximum margin. It works well with high-dimensional and non-linear data when used with kernels.

**Accuracy:**

The accuracy of Support Vector Analysis is 97.32%

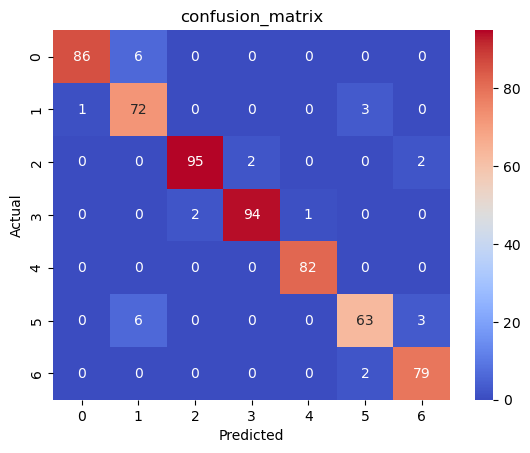
**Confusion Matrix:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 86 | 6 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 72 | 0 | 0 | 0 | 3 | 0 |
| 2 | 0 | 0 | 95 | 2 | 0 | 0 | 2 |
| 3 | 0 | 0 | 2 | 94 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 82 | 0 | 0 |
| 5 | 0 | 6 | 0 | 0 | 0 | 63 | 3 |
| 6 | 0 | 0 | 0 | 0 | 0 | 2 | 79 |

**Confusion Matrix Analysis:**

* Strong performance across all classes.
* Only slight misclassifications in Class 1 and 5.
* Classes 2, 3, 4, and 6 are nearly perfect.

**Heat Map of Confusion Matrix:**



**Heat Map Analysis:**

Very dark diagonal line with lighter surrounding areas. Shows excellent class separation. Only minimal misclassification between similar classes (like Class 5,6 and 1).

# Summary of Results’ Accuracy:

Now we will compare the results’ accuracy of these all classification model that are applied here:

|  |  |
| --- | --- |
| **Model Name** | **Accuracy (%)** |
| Support Vector Machine (SVM) | 97.32 |
| Random Forest | 97.24 |
| Decision Tree | 96.99 |
| Logistic Regression | 89.97 |
| K-Nearest Neighbors (KNN) | 84.14 |

It is cleared from the accuracy table that “Support Vector Machine” have greatest accuracy. Therefore, SVM model is the Best fitted model on our “Obesity Dataset”.