# Estimation of Obesity Level based on eating Habits and physical condition.

## Introduction

This report examines an obesity dataset and its associated variables. The dataset includes attributes such as frequency of high-calorie food consumption (FAVC), frequency of vegetable consumption (FCVC), number of main meals (NCP), consumption of food between meals (CAEC), daily water consumption (CH20), alcohol consumption (CALC), calories consumption monitoring (SCC), physical activity frequency (FAF), time spent using technology devices (TUE), mode of transportation used (MTRANS), gender, age, height, and weight. The dataset was labelled, and a NObesity class variable with categories like Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III was established.

The primary goal of this investigation is to investigate the link between various variables and the target variable NObesity, as well as to develop prediction models to identify the obesity type based on the supplied data.

This data was used to estimate obesity levels in adults aged 14 to 61 from Mexico, Peru, and Colombia, with a variety of eating habits and physical conditions. The information was analyzed after the data was obtained utilizing an online platform with a survey in which anonymous people answered each question.

The following characteristics are associated with eating habits:

* Frequent intake of high caloric meals (FAVC)
* Frequent consumption of vegetables (FCVC)
* Number of primary meals (NCP)
* Consumption of food between meals (CAEC)
* Daily water consumption (CH20)
* Alcohol consumption (CALC).
* Calories consumption monitoring (SCC)
* Physical activity frequency (FAF)
* Time utilizing technology devices.
* (TUE) Transportation used.
* (MTRANS) are the physical condition parameters.
* Other characteristics retrieved were Gender, Age, Height, and Weight.

Finally, all data was labelled, and the class variable NObesity was formed using the BMI values shown below:

* Underweight less than 18.5 lbs.
* Normal 18.5 to 24.9 lbs.
* Overweight 25.0 to 29.9 lbs.
* Obesity I 30.0 to 34.9 lbs.
* Obesity II 35.0 to 39.9 lbs.
* Obesity III greater than 40 lbs.
* Because the data contains both numerical and continuous information, it may be analyzed using algorithms for classification, prediction, segmentation, and association.
* A screenshot of a graph

  Description automatically generated

### Dataset Analysis

Handling missing numbers and outliers was part of the data wrangling process. The dropna() method was used to remove missing values before analyzing the remaining data. The IQR approach was used to find and eliminate outliers in the 'Height' and 'Weight' columns.

#### EDA

EDA was performed to gain insights into the distribution and relationships between variables. Histograms, scatter plots, box plots, and count plots were used to visualize the data.

A graph of different colored bars

Description automatically generated

This graph shows that all of the output categories were given identical weightage while there is no data collection inconsistency in the data

A screenshot of a chart

Description automatically generated

This graph depicts the relationship between all of the numerical variables. As we can see, there is no multicollinearity in the data since the characteristics are not intercorrelated.

A graph of a distribution plot

Description automatically generated

This graph depicts the height distribution in the data. The average height spans from 1.5 to 1.9 metres, with the majority of persons standing between 1.7 and 1.75 metres tall.

1. Height Vs Nobeyesdad : A graph with blue dots

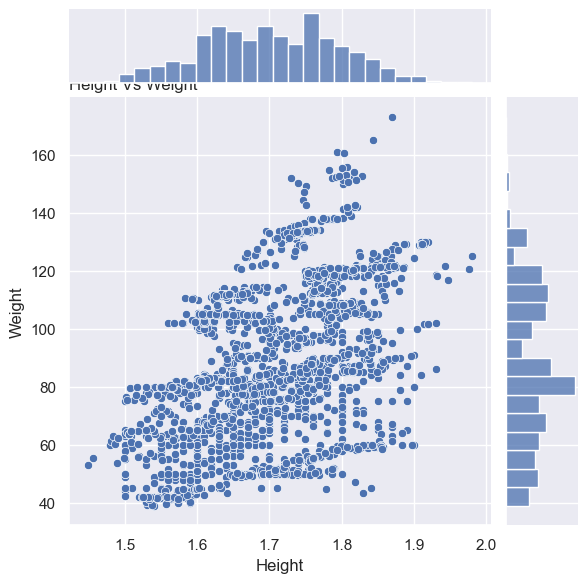
   Description automatically generated with medium confidence

According to this graph, the distribution of height is even across all obesity groups.

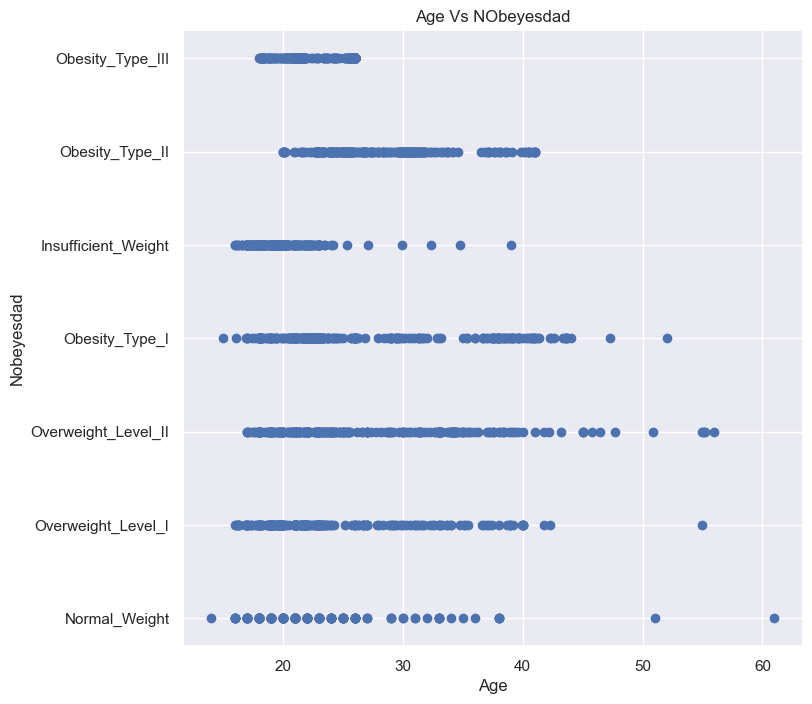
1. Weight Vs NObeysdad: A graph with blue dots

   Description automatically generated

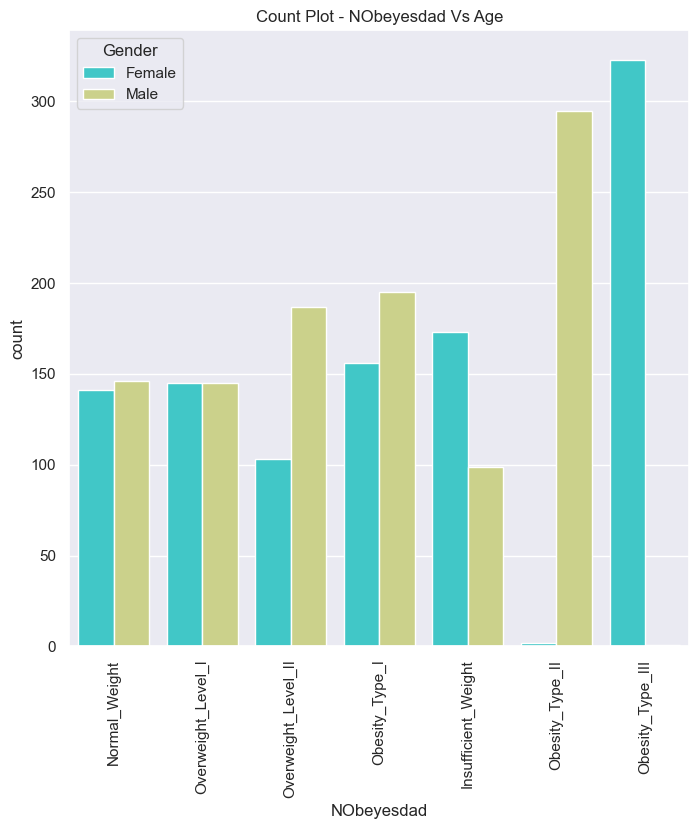
Weight distribution is unequal between obesity categories, which is noticeable since even little weight changes impact obesity. It also displays the weight ranges for each group.



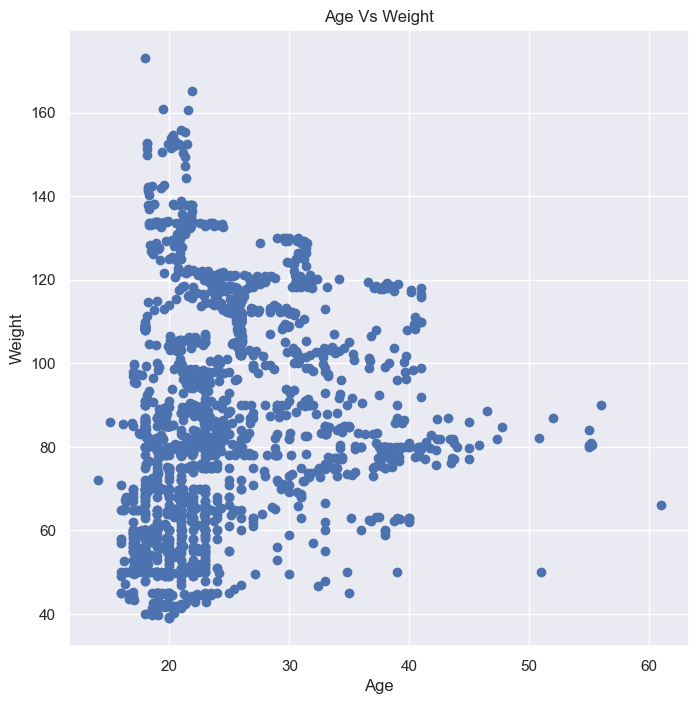
This graph depicts the height vs. weight dispersion. There is a thick dispersion where the height is 1.5 to 1.7 and the weight is 60 to 80.



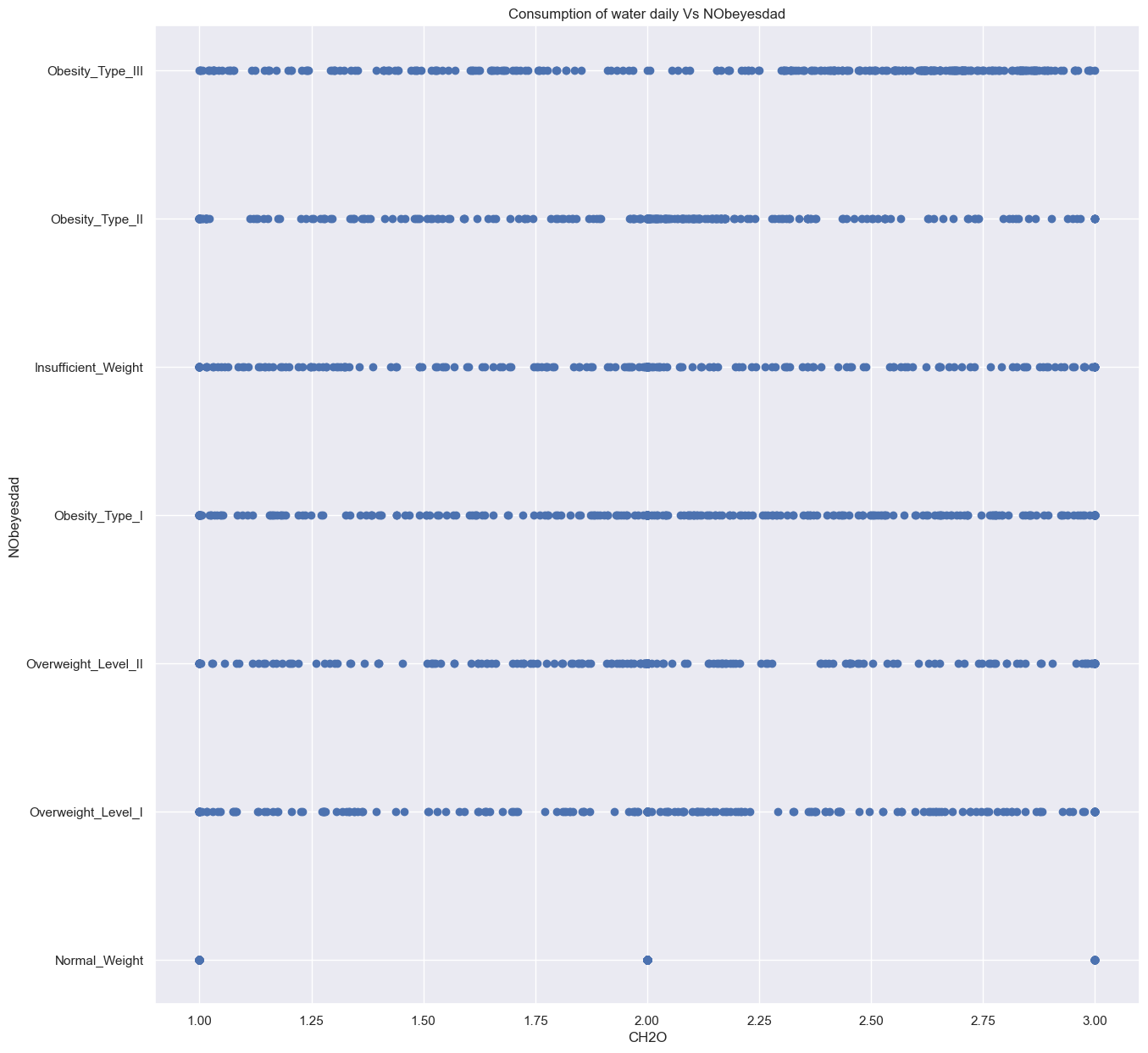
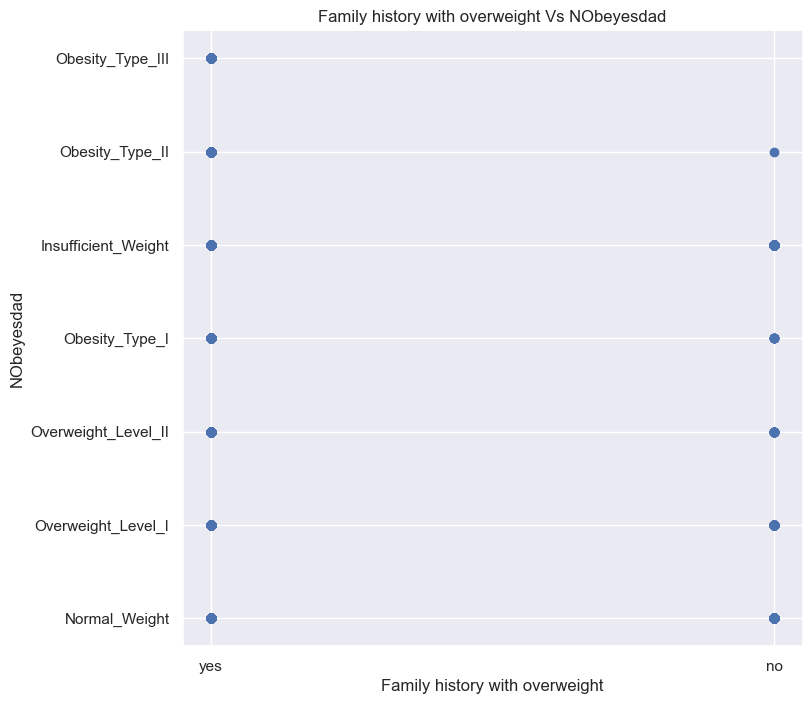
This shows the distribution of age over the obesity type. From this we can also see the age group of the people lying in each category of obesity. We can also see some outliers for age of people who fall under Normal weight category

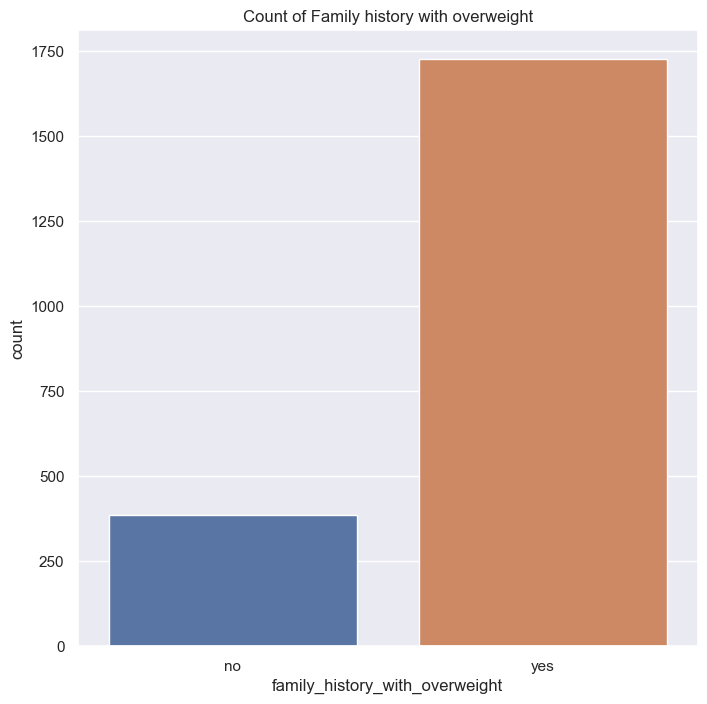


This graphic shows that the average age of persons classified as Overweight\_Level\_II and Obsity\_Type\_I is older than that of those classified as other types of obesity. As a result, we may conclude that Overweight\_Level\_II and Obesity\_Type\_I are not more likely to occur in persons under the age of 20.

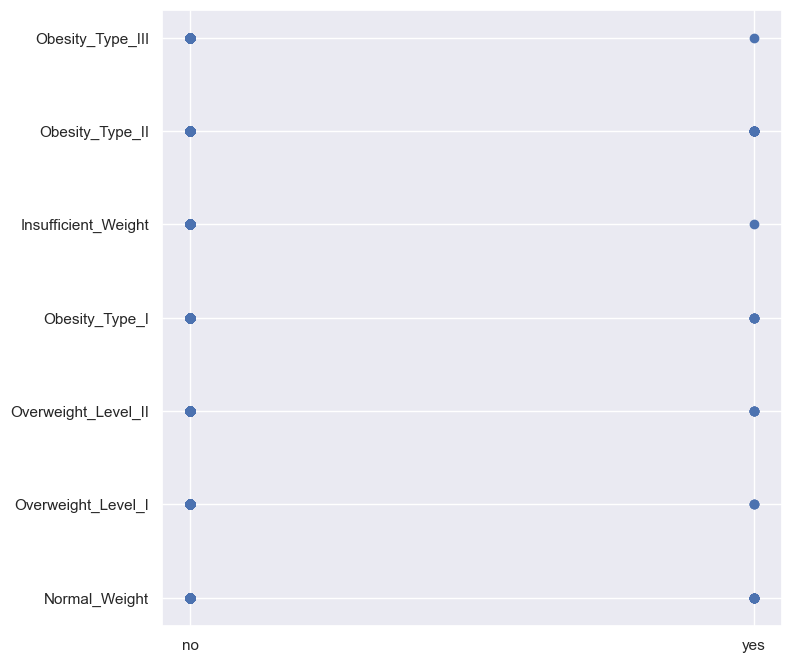


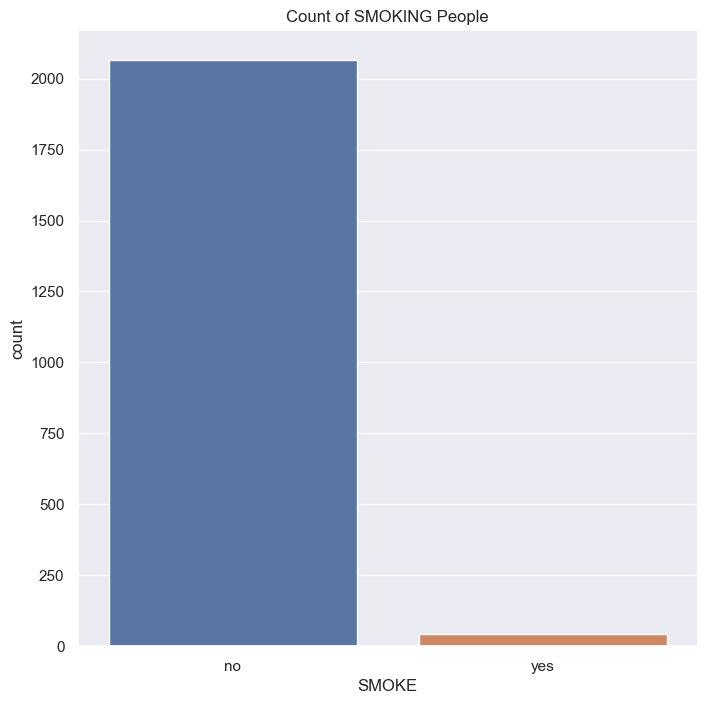
This figure shows that the density is highest when the age is between 15 and 25 and the weight is between 40 and 80.

This figure shows that consumption of water is almost same in each type except the normal weight

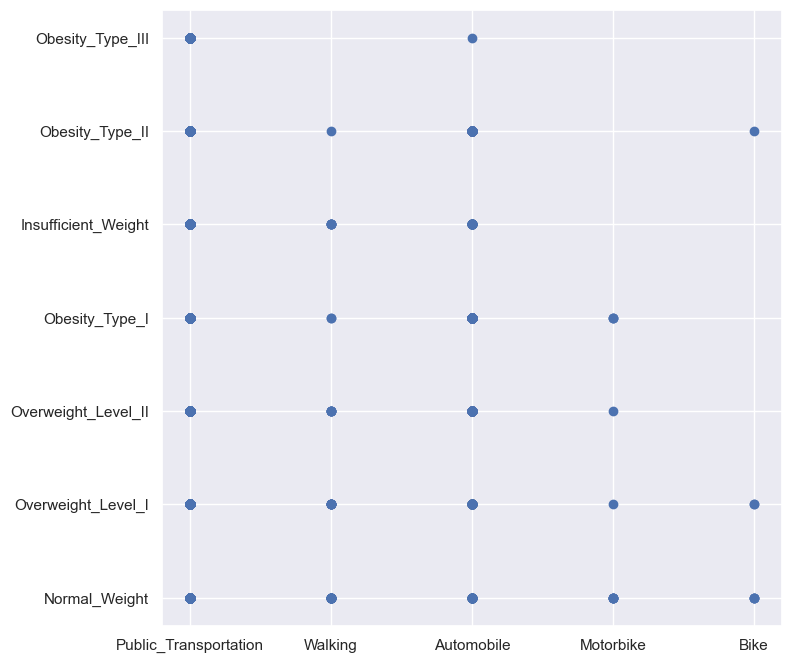


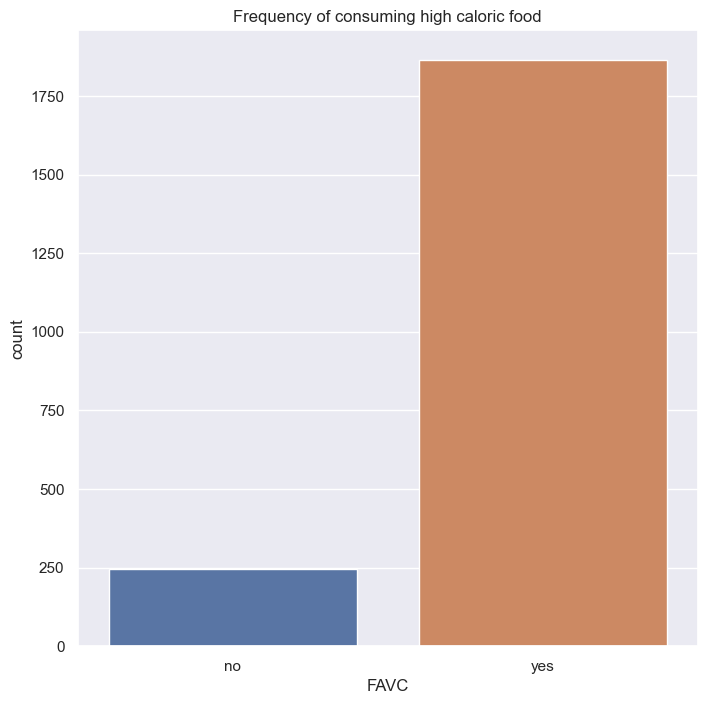
From this graph we can see that there is a bias in sampling of the data in terms of family history because we can see only a few samples of data for “no” category and we can see more number of data for yes category





We can also observe a bias in the data gathering from this graph since we don't have enough data for SMOKE = "yes."





The graph demonstrates that those who consume high calorie foods outnumber those who do not.

A graph with a green bar and blue bar

Description automatically generated with medium confidence

This graph shows that people who consume food between meals sometimes are more compared to other categories.

A graph of a box plot

Description automatically generated with medium confidence

The plot shows that people who consume food between meals sometimes are most likely to be in the age of 20-40

A graph of different colored squares

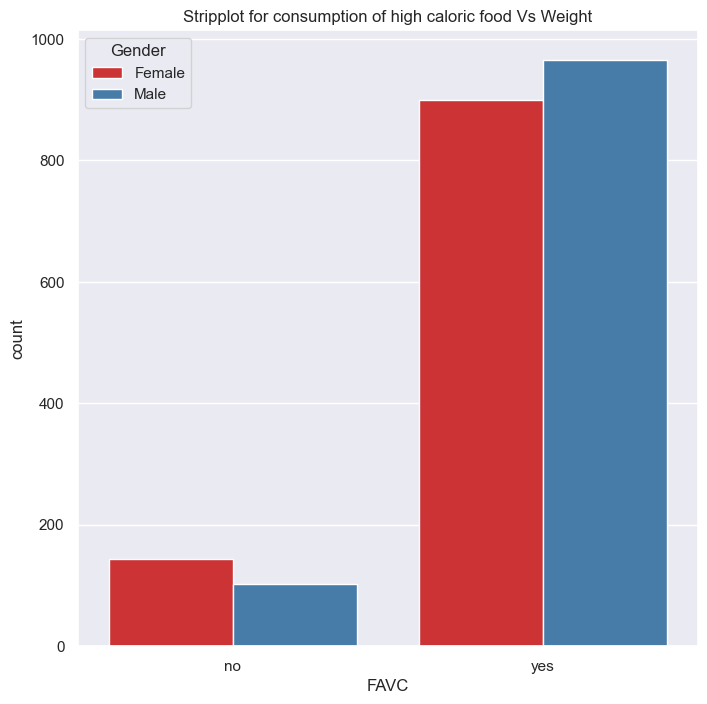
Description automatically generated

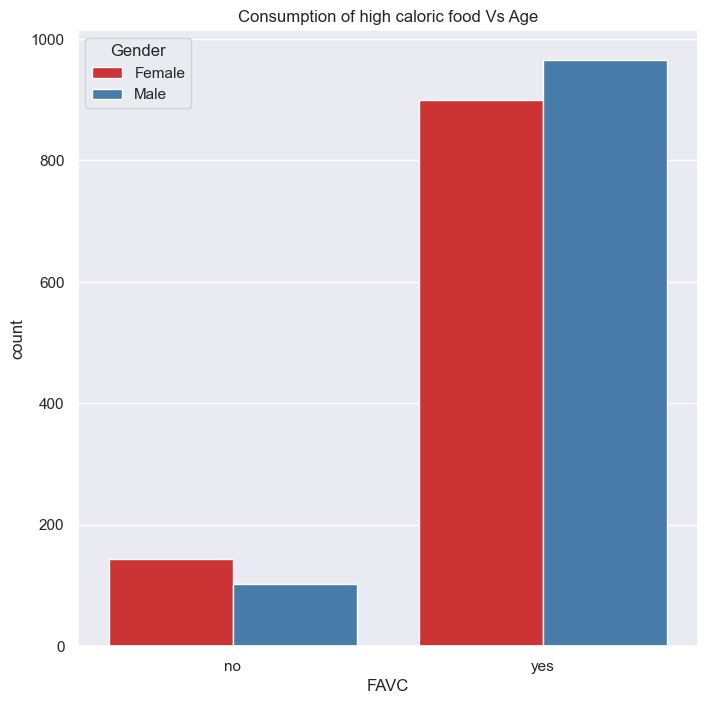
This graph shows that consumption of alcohol sometimes is more while compared to other categories.

A graph of different colored squares

Description automatically generated

This box plot demonstrates that most alcoholic drinkers are between the ages of 20 and 30.





The graph shows that both Gender consumes equal amount of caloric food as they age

Out[131]:

|  | **Gender** | **Age** | **Height** | **Weight** | **family\_history\_with\_overweight** | **FAVC** | **FCVC** | **NCP** | **CAEC** | **SMOKE** | **CH2O** | **SCC** | **FAF** | **TUE** | **CALC** | **MTRANS** | **NObeyesdad** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 21.0 | 1.62 | 64.0 | 1 | 0 | 2.0 | 3.0 | 1 | 0 | 2.0 | 0 | 0.0 | 1.0 | 0 | 0 | Normal\_Weight |
| **1** | 0 | 21.0 | 1.52 | 56.0 | 1 | 0 | 3.0 | 3.0 | 1 | 1 | 3.0 | 1 | 3.0 | 0.0 | 1 | 0 | Normal\_Weight |
| **2** | 1 | 23.0 | 1.80 | 77.0 | 1 | 0 | 2.0 | 3.0 | 1 | 0 | 2.0 | 0 | 2.0 | 1.0 | 2 | 0 | Normal\_Weight |
| **3** | 1 | 27.0 | 1.80 | 87.0 | 0 | 0 | 3.0 | 3.0 | 1 | 0 | 2.0 | 0 | 2.0 | 0.0 | 2 | 1 | Overweight\_Level\_I |
| **4** | 1 | 22.0 | 1.78 | 89.8 | 0 | 0 | 2.0 | 1.0 | 1 | 0 | 2.0 | 0 | 0.0 | 0.0 | 1 | 0 | Overweight\_Level\_II |

As we can see, each string values have been convert into numeric value according to their

We can observe that each variable is globally linearly independent of the others. As a result, ACP will be ineffective, hence we shall not conduct it. We will choose the variable that is associated with less than 30%.

### [¶](#step-3-Feature-selection)

### Key Findings from EVA:

## The dataset has a balanced representation of weight categories, which provides enough data for analysis. There are no significant correlations between the numerical variables, indicating that there are no difficulties with multicollinearity.

## Feature Choice

## Building an effective machine learning model requires careful feature selection. Two approaches of feature selection were used:

## Variance Threshold: To minimise dimensionality and enhance model performance, features with low variance (below the set threshold) were deleted.

## Select-K-Best: Based on the F-regression score, the top K features with the best predictive power were chosen. The selected characteristics were utilised for modelling after being picked using feature selection techniques.

## Model Implementation

## Several machine learning models were considered for predicting the weight category based on the selected features. The models used are as follows:

# [¶](#Using-Variance-Threshold)

#### Logistic Regression:

A simple yet effective classification algorithm suitable for binary and multiclass problems.

Model summary:

precision recall f1-score support

Insufficient\_Weight 0.81 0.86 0.84 107

Normal\_Weight 0.63 0.51 0.56 94

Obesity\_Type\_I 0.81 0.83 0.82 109

Obesity\_Type\_II 0.92 0.96 0.94 108

Obesity\_Type\_III 1.00 0.97 0.99 102

Overweight\_Level\_I 0.61 0.65 0.63 84

Overweight\_Level\_II 0.65 0.65 0.65 92

accuracy 0.79 696

macro avg 0.78 0.78 0.78 696

weighted avg 0.79 0.79 0.79 696

Logistic regression model is trained for the following c values and the optimum model is obtained C : Regularization parameter. 1,5,10,20

Accuracy: 0.7887931034482759

Gradient Boosting Classifier:

An ensemble technique combining weak learners (decision trees) to create a strong classifier.

Append the model's performance metrics (accuracy and model name "Gradient Boosting Classifier") to the 'perf' list:

perf.append([score, "Gradient Boosting Classifier"])

Confusion matrix:

[[ 98 6 0 1 0 1 1]

[ 1 77 4 1 1 9 1]

[ 0 2 105 1 0 0 1]

[ 0 3 2 103 0 0 0]

[ 1 1 0 2 97 0 1]

[ 0 8 1 0 0 72 3]

[ 0 6 5 0 0 2 79]]

Model summary:

precision recall f1-score support

Insufficient\_Weight 0.98 0.92 0.95 107

Normal\_Weight 0.75 0.82 0.78 94

Obesity\_Type\_I 0.90 0.96 0.93 109

Obesity\_Type\_II 0.95 0.95 0.95 108

Obesity\_Type\_III 0.99 0.95 0.97 102

Overweight\_Level\_I 0.86 0.86 0.86 84

Overweight\_Level\_II 0.92 0.86 0.89 92

accuracy 0.91 696

macro avg 0.91 0.90 0.90 696

weighted avg 0.91 0.91 0.91 696

Accuracy: 0.9066091954022989

## **Support Vector Machine (SVM):** A powerful algorithm for classification tasks, effective in handling high-dimensional data.

Append the model's performance metrics (accuracy and model name "SVM2") to the 'perf' list

perf.append([score, "SVM2"])

Confusion matrix:

[[ 97 7 0 1 0 1 1]

[ 11 68 3 1 0 10 1]

[ 0 1 101 0 0 1 6]

[ 0 3 2 103 0 0 0]

[ 2 0 0 0 99 0 1]

[ 0 8 1 0 0 66 9]

[ 0 4 3 1 0 5 79]]

Model summary:

precision recall f1-score support

Insufficient\_Weight 0.88 0.91 0.89 107

Normal\_Weight 0.75 0.72 0.74 94

Obesity\_Type\_I 0.92 0.93 0.92 109

Obesity\_Type\_II 0.97 0.95 0.96 108

Obesity\_Type\_III 1.00 0.97 0.99 102

Overweight\_Level\_I 0.80 0.79 0.79 84

Overweight\_Level\_II 0.81 0.86 0.84 92

accuracy 0.88 696

macro avg 0.88 0.88 0.88 696

weighted avg 0.88 0.88 0.88 696

Accuracy: 0.8807471264367817

## **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, suitable for handling discrete data.

Append the model's performance metrics (accuracy and model name "Naive Bayes") to the 'perf' list

perf.append([score, "Naive Bayes"])

Confusion matrix:

[[87 15 0 1 0 3 1]

[43 24 8 0 0 11 8]

[ 0 2 60 41 3 2 1]

[ 0 1 3 99 0 2 3]

[ 2 2 0 0 98 0 0]

[ 2 10 31 1 1 33 6]

[ 0 6 43 11 1 2 29]]

Model summary:

precision recall f1-score support

Insufficient\_Weight 0.65 0.81 0.72 107

Normal\_Weight 0.40 0.26 0.31 94

Obesity\_Type\_I 0.41 0.55 0.47 109

Obesity\_Type\_II 0.65 0.92 0.76 108

Obesity\_Type\_III 0.95 0.96 0.96 102

Overweight\_Level\_I 0.62 0.39 0.48 84

Overweight\_Level\_II 0.60 0.32 0.41 92

accuracy 0.62 696

macro avg 0.61 0.60 0.59 696

weighted avg 0.61 0.62 0.60 696

Accuracy: 0.617816091954023

## **Random Forest:** A popular ensemble method using multiple decision trees for classification tasks.

confusion matrix

[[ 94 10 0 1 0 1 1]

[ 4 79 3 0 1 6 1]

[ 0 1 106 0 0 0 2]

[ 0 3 1 104 0 0 0]

[ 1 1 0 0 99 1 0]

[ 0 7 1 0 0 73 3]

[ 0 4 3 0 0 1 84]]

model summary :

precision recall f1-score support

Insufficient\_Weight 0.95 0.88 0.91 107

Normal\_Weight 0.75 0.84 0.79 94

Obesity\_Type\_I 0.93 0.97 0.95 109

Obesity\_Type\_II 0.99 0.96 0.98 108

Obesity\_Type\_III 0.99 0.97 0.98 102

Overweight\_Level\_I 0.89 0.87 0.88 84

Overweight\_Level\_II 0.92 0.91 0.92 92

accuracy 0.92 696

macro avg 0.92 0.92 0.92 696

weighted avg 0.92 0.92 0.92 696

accuracy : 0.9181034482758621

## The models were trained on the training data and evaluated on the test data using accuracy, precision, recall, and F1-score metrics.

# Results Interpretation and Implications

## The machine learning models were evaluated based on their performance metrics. The results are summarized below:

## Random Forest Classifier:

## Accuracy: 91.81%

## Precision, Recall, and F1-score: The model achieved high precision, recall, and F1-score for most classes, indicating good performance in predicting different weight categories. The model is especially effective for Obesity Type II and Obesity Type III.

## Implications: Random Forest is a powerful ensemble method that performs well in both binary and multiclass classification tasks. Its ability to handle complex relationships in data and reduce overfitting makes it a reliable choice for this problem.

## Gradient Boosting Classifier:

## Accuracy: 90.66%

## Precision, Recall, and F1-score: The model shows strong performance in predicting most weight categories, especially for Obesity Type I, Obesity Type II, and Obesity Type III.

## Implications: Gradient Boosting is another ensemble technique that can achieve high accuracy by combining multiple weak learners. It is an effective choice for multiclass classification and can handle complex data distributions.

## Support Vector Machine (SVM):

## Accuracy: 88.07%

## Precision, Recall, and F1-score: SVM performed well for most classes, particularly for Insufficient Weight, Obesity Type I, and Obesity Type II.

## Implications: SVM is a powerful algorithm for classification tasks, especially when dealing with high-dimensional data. It is a versatile choice, but hyperparameter tuning and data scaling can significantly impact its performance.

## Naive Bayes:

## Accuracy: 61.78%

## Precision, Recall, and F1-score: Naive Bayes achieved moderate performance for some classes, but struggled with others. It performed best for Obesity Type II and Obesity Type III.

## Implications: Naive Bayes is a simple and fast classifier, suitable for discrete data. While it may not perform as well as other algorithms on this specific dataset, it can be useful for certain types of problems.

## Overall, the Random Forest Classifier stands out with the highest accuracy and balanced performance across all weight categories. It is the recommended model for predicting weight categories based on the provided dataset. However, further exploration and experimentation with other algorithms, hyperparameter tuning, and feature engineering may lead to even better results in the future. It's essential to continue refining the model to improve its performance and generalization capabilities on new data.

### Out-of-Sample Predictions

## Out-of-sample predictions were performed to simulate the model's performance in a real-world production environment using new data. This data was obtained separately from the test dataset used during model evaluation. The model's predictions were analyzed and compared to the actual weight categories to assess its performance in real-world scenarios.

### Concluding Remarks

## In conclusion, the analysis of the Obesity dataset has provided valuable insights into the factors influencing weight categories. The machine learning models developed in this project can effectively predict an individual's weight category based on their dietary and physical habits.

## It is essential to note that the choice of the most suitable model depends on the specific objectives and requirements of the project. Further analysis and fine-tuning of the models may be necessary to optimize their performance for specific use cases.

Overall, this project contributes to the understanding of obesity-related factors and offers predictive models that can assist in healthcare decision-making and interventions.[¶](#Linear-Regression-Model-with-Feature-Se)