Info 212: Data Science Programming I

Data Science Final Project Report:

Predictive Modeling and Key Factor Analysis of Obesity Levels

Group Number: 16
Group Members:

冯冠博	(Guanbo Feng)	- 320220940211,
成钢	(Gang Cheng)	- 320220940141,
叶禹昕	(Yuxin Ye)	- 320220940981,
高浩文	(Haowen Gao)	- 320220940221

Emails:
320220940211@lzu.edu.cn, 320220940141@lzu.edu.cn,
320220940981@lzu.edu.cn, 320220940221@lzu.edu.cn

Predictive Modeling and Key Factor Analysis of Obesity Levels

Contents

1 Executive Summary

Obesity is an important global health issue, leading to many chronic diseases and health complications. The project aims to use a comprehensive dataset from Mexico, Peru, and Colombia to analyze population and lifestyle factors that affect obesity levels.

The processing of data is indispensable. At this step, we processed the missing values and encoded the labels (for future machine learning). Afterwards, we also conducted EDA. Further research was conducted on the data characteristics. In this step, we also added a feature BMI to the dataset, which can more effectively reflect the two data of height and weight.

Next, we trained the model. The ***RandomForestClassifier*** model is used. We also evaluated the accuracy and recall of the model. The trained model performs well.

After developing the prediction model, we visualized the data. This helps us to visually analyze the features. Understand the causes of obesity, understand the distribution of data, and understand the correlation between features.

In order to gain a deeper understanding of the causes of obesity, we conducted in-depth analysis of various characteristics. Detailed description of the correlation between various features and obesity. We are not limited to this, but have also explored the relationships between different features and discovered many interesting conclusions.

2 Introduction

Obesity is a pervasive global health issue that poses significant physical and mental health challenges. The *World Health Organization (WHO)* has identified obesity as a major risk factor for a range of chronic diseases, including diabetes, cardiovascular diseases, and certain cancers. The increasing prevalence of obesity worldwide necessitates ongoing research to understand its determinants and develop predictive models to identify individuals at risk. In Latin America, countries such as **Mexico**, **Peru**, and **Colombia** have experienced alarming rates of obesity, making it imperative to investigate the factors contributing to this epidemic in these regions.

3 Dataset Description

3.1 Data Collection

This project utilizes a comprehensive dataset on Kaggle designed to estimate obesity levels among individuals from **Mexico**, **Peru**, and **Colombia**.

By calling the API Command provided by Kaggle, we downloaded the CSV data file of the dataset:

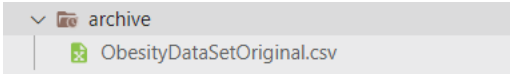

Figure 3.1 - File Directory

The dataset comprises **2111** records with **17** attributes, including **demographic**, **dietary**, and **lifestyle** factors. The target variable, ***NObeyesdad (Obesity Level)***, categorizes individuals into seven distinct obesity levels. The features selected for our analysis include:

Table 3.2 - Detailed Explanation of Variables

Gender:	Feature, Categorical, "Gender"
Age:	Feature, Continuous, "Age"
Height:	Feature, Continuous
Weight:	Feature, Continuous
FHWO:	Feature, Binary, "Has a family member suffered from overweight?"
FAVC:	Feature, Binary, "Do you eat high caloric food frequently?"
FCVC:	Feature, Integer, "Do you usually eat vegetables in your meals?"
NCP:	Feature, Continuous, "How many main meals do you have daily?"
CAEC:	Feature, Categorical, "Do you eat any food between meals?"
SMOKE:	Feature, Binary, "Do you smoke?"
CH2O:	Feature, Continuous, "How much water do you drink daily?"
SCC:	Feature, Binary, "Do you monitor the calories you eat daily?"
FAF:	Feature, Continuous, "How often do you have physical activity?"

Table 3.2 (Continued) - Detailed Explanation of Variables

TUE: Feature, Integer, "How much time do you use technological devices such as cell phone, videogames, television, computer and others?"
CALC: Feature, Categorical, "How often do you drink alcohol?"
MTRANS: Feature, Categorical, "Which transportation do you usually use?"
NObeyesdad: Target, Categorical, "Obesity level"

Further examination of the dataset revealed that the majority of the attributes are complete, with no missing values in **16** out of the **17** columns. However, the target variable ***NObeyesdad*** has **77** missing values out of **2111** entries. These missing values account for approximately **3.65%** of the total records, which is a relatively small portion. Given the small proportion of missing values, their presence is unlikely to significantly impact the overall analysis or the performance of predictive models.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             2111 non-null   float64
 1   Gender                          2111 non-null   object
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   CALC                            2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   SCC                             2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  family_history_with_overweight  2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CAEC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2034 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

Figure - 3.3 Data Inspection: Information

Consequently, the dataset remains robust for the intended analyses, allowing for accurate predictions and meaningful insights into the factors influencing obesity levels.

3.2 Handling Missing Values

In our data preprocessing phase, we addressed the issue of missing values within the dataset. Given that the missing values represented a very small portion of the dataset (approximately **3.65%**), we first isolated these instances into a separate subset named ***data_missing***.

Next, we excluded the records with missing values from the original dataset. This step ensured that our subsequent analysis and predictive modeling would be based on a complete and robust dataset, free from the potential biases and inaccuracies that incomplete data might introduce.

```
display(le_data.head())
display(le_data_missing.head())
```

	Age	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRANS	NObeyesdad
0	21.0	0	1.62	64.0	3	0	2.0	3.0	0	0	2.0	1	0.0	1.0	2	3	1
1	21.0	0	1.52	56.0	2	0	3.0	3.0	1	1	3.0	1	3.0	0.0	2	3	1
2	23.0	1	1.80	77.0	1	0	2.0	3.0	0	0	2.0	1	2.0	1.0	2	3	1
3	27.0	1	1.80	87.0	1	0	3.0	3.0	0	0	2.0	0	2.0	0.0	2	4	2
4	22.0	1	1.78	89.8	2	0	2.0	1.0	0	0	2.0	0	0.0	0.0	2	3	3

	Age	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRANS	NObeyesdad
0	29.0	1	1.62	53.0	2	1	2.0	3.0	0	0	2.0	0	0.0	0.0	2	0	NaN
1	31.0	0	1.58	68.0	2	0	2.0	1.0	0	0	1.0	1	1.0	0.0	2	3	NaN
2	21.0	0	1.62	69.0	3	1	1.0	3.0	0	0	2.0	1	0.0	1.0	1	3	NaN
3	23.0	1	1.72	70.0	1	0	2.0	3.0	0	0	2.0	0	3.0	1.0	2	3	NaN
4	22.0	0	1.50	49.0	3	0	2.0	1.0	0	0	2.0	1	3.0	0.0	2	4	NaN

Figure 3.4 - Data Inspection: Spilting Missing Values

To facilitate training our predictive models, we used the cleaned dataset, which now contained only complete records. After training the models, we planned to leverage the trained models to predict and fill in the missing values in the ***data_missing*** subset.

3.3 Label Encoding

To prepare the dataset for analysis and predictive modeling, we conducted a series of steps to convert **categorical data** into a numerical format. We applied **label encoding** to

these categorical columns. **Label encoding** is a method that converts categorical values into numerical values, which are more suitable for machine learning.

```
# Assuming encoders["NObeyesdad"] is already fitted
label_encoder = encoders["NObeyesdad"]

display(label_encoder)
# Print the classes and their corresponding integer labels
for index, class_label in enumerate(label_encoder.classes_):
    print(f"Class: {class_label.ljust(20)}, Encoded Label: {index}")
```

```
LabelEncoder
LabelEncoder()

Class: Insufficient_Weight , Encoded Label: 0
Class: Normal_Weight       , Encoded Label: 1
Class: Overweight_Level_I  , Encoded Label: 2
Class: Overweight_Level_II , Encoded Label: 3
Class: Obesity_Type_I      , Encoded Label: 4
Class: Obesity_Type_II     , Encoded Label: 5
Class: Obesity_Type_III    , Encoded Label: 6
```

Figure 3.5 - Label Encoder of *NObeyesdad*

We utilized a ***LabelEncoder*** for each categorical column, fitting the encoder to the unique values of the column and then transforming the column into numerical values. We also stored each ***LabelEncoder*** in a dictionary, ensuring that we could later reverse the transformation or apply the same encoding to other datasets:

	Age	Gender	Height	Weight	CALC
0	21.000000	Female	1.620000	64.000000	no
1	21.000000	Female	1.520000	56.000000	Sometimes
2	23.000000	Male	1.800000	77.000000	Frequently
3	27.000000	Male	1.800000	87.000000	Frequently
4	22.000000	Male	1.780000	89.800000	Sometimes
...	...	...	...	...	...
2029	20.976842	Female	1.710730	131.408528	Sometimes
2030	21.982942	Female	1.748584	133.742943	Sometimes
2031	22.524036	Female	1.752206	133.689352	Sometimes
2032	24.361936	Female	1.739450	133.346641	Sometimes
2033	23.664709	Female	1.738836	133.472641	Sometimes

	Age	Gender	Height	Weight	CALC	FAVC
0	21.000000	0	1.620000	64.000000	3	0
1	21.000000	0	1.520000	56.000000	2	0
2	23.000000	1	1.800000	77.000000	1	0
3	27.000000	1	1.800000	87.000000	1	0
4	22.000000	1	1.780000	89.800000	2	0
...	...	...	...	...	...	...
2029	20.976842	0	1.710730	131.408528	2	1
2030	21.982942	0	1.748584	133.742943	2	1
2031	22.524036	0	1.752206	133.689352	2	1
2032	24.361936	0	1.739450	133.346641	2	1
2033	23.664709	0	1.738836	133.472641	2	1

Figure 3.6 - Data Before (Left) and After (Right) Applying Label Encoding

4 Exploratory Data Analysis (EDA)

Notes: In this section, we present some initial calculations and previews of the data as part of our preliminary **Exploratory Data Analysis**. More comprehensive visualizations and in-depth analysis will be showcased in **6: Visualization and Insights**.

4.1 Dataset Overview

Display the first few rows of the dataset to get an initial sense of the data structure:

```
data.head()
```

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

Figure 4.1 - First Few Rows of the Dataset

Summarize the dataset to understand the number of records, data types, and the presence of missing values:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2034 entries, 0 to 2033
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Age         2034 non-null   float64
 1   Gender      2034 non-null   object
 2   Height      2034 non-null   float64
 3   Weight      2034 non-null   float64
 4   CALC        2034 non-null   object
 5   FAVC        2034 non-null   object
 6   FCVC        2034 non-null   float64
 7   NCP         2034 non-null   float64
 8   SCC         2034 non-null   object
 9   SMOKE       2034 non-null   object
 10  CH2O        2034 non-null   float64
 11  FHWO        2034 non-null   object
 12  FAF         2034 non-null   float64
 13  TUE         2034 non-null   float64
 14  CAEC        2034 non-null   object
 15  MTRANS      2034 non-null   object
 16  NObeyesdad  2034 non-null   object
dtypes: float64(8), object(9)
memory usage: 270.3+ KB
```

Figure 4.2 – Basic Information of the Dataset

Since we have completed the loading of the dataset and handling missing values in the previous **Step 2: Data Collection**, we will not repeat these steps here.

4.2 Summary Statistics

Calculating summary statistics such as ***mean***, ***median***, ***standard deviation***, ***minimum***, and ***maximum*** for **continuous variables** provides a concise overview of the dataset. These statistics help in understanding the central tendency, dispersion, and range of the data:

```
data.describe()
```

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2034.000000	2034.000000	2034.000000	2034.000000	2034.000000	2034.000000	2034.000000	2034.000000
mean	24.286542	1.701898	86.604216	2.418067	2.684921	2.010149	1.011160	0.657801
std	6.354498	0.093147	26.298687	0.533195	0.779085	0.610696	0.849520	0.606660
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.899719	1.630000	65.130595	2.000000	2.657909	1.598178	0.121980	0.000000
50%	22.737801	1.701383	83.000000	2.378672	3.000000	2.000000	1.000000	0.629431
75%	26.000000	1.768234	107.999704	3.000000	3.000000	2.475631	1.668462	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

Figure 4.3 – Summary Statistics of the Dataset

Examining the distribution of **categorical variables** provides insights into the frequency and proportion of different categories within each variable. This analysis is crucial for understanding the composition of the dataset and identifying any imbalances or dominant categories:

```
Distribution of Categorical Variables:
```

Gender	Count
Male	1028
Female	1006

FAVC	Count
yes	1799
no	235

SCC	Count
no	1945
yes	89

SMOKE	Count
no	1990
yes	44

FHWO	Count
yes	1661
no	373

CALC	Count
Sometimes	1347
no	618
Frequently	68
Always	1

CAEC	Count
Sometimes	1702
Frequently	230
Always	51
no	51

MTRANS	Count
Public_Transportation	1526
Automobile	438
Walking	53
Motorbike	11
Bike	6

NObeyesdad	Count
Obesity_Type_I	338
Obesity_Type_III	312
Obesity_Type_II	287
Overweight_Level_I	279
Overweight_Level_II	279
Normal_Weight	271
Insufficient_Weight	268

Figure 4.4 – Distribution of Categorical Variables

4.3 Feature Engineering

Feature engineering involves creating new features or modifying existing ones to improve the performance of machine learning models. For this project, we engineered a new feature known as **Body Mass Index (BMI)**, which is a widely recognized measure used to classify individuals based on their weight relative to their height. The calculation is as follows:

$$BMI = \frac{Weight\ (\boldsymbol{kg})}{Height^{\boldsymbol{2}}(\boldsymbol{m^2})} \qquad \textbf{(Formula 4.5)}$$

	Age	Gender	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRANS	BMI	NObeyesdad
0	21.000000	0	3	0	2.0	3.0	0	0	2.000000	1	0.000000	1.000000	2	3	24.386526	1
1	21.000000	0	2	0	3.0	3.0	1	1	3.000000	1	3.000000	0.000000	2	3	24.238227	1
2	23.000000	1	1	0	2.0	3.0	0	0	2.000000	1	2.000000	1.000000	2	3	23.765432	1
3	27.000000	1	1	0	3.0	3.0	0	0	2.000000	0	2.000000	0.000000	2	4	26.851852	2
4	22.000000	1	2	0	2.0	1.0	0	0	2.000000	0	0.000000	0.000000	2	3	28.342381	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2029	20.976842	0	2	1	3.0	3.0	0	0	1.728139	1	1.676269	0.906247	2	3	44.901475	6
2030	21.982942	0	2	1	3.0	3.0	0	0	2.005130	1	1.341390	0.599270	2	3	43.741923	6
2031	22.524036	0	2	1	3.0	3.0	0	0	2.054193	1	1.414209	0.646288	2	3	43.543817	6
2032	24.361936	0	2	1	3.0	3.0	0	0	2.852339	1	1.139107	0.586035	2	3	44.071535	6
2033	23.664709	0	2	1	3.0	3.0	0	0	2.863513	1	1.026452	0.714137	2	3	44.144338	6

Figure 4.6 - Feature Engineering: Calculate BMI

4.4 Pearson Correlation Matrix

In this step, we calculate and analyze the Pearson correlation matrix for the dataset. The Pearson correlation coefficient measures the linear relationship between two variables, with values ranging from -1 to 1.

	Age	Gender	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRAI
Age	1.000000	0.045994	-0.049882	0.062370	0.013493	-0.043225	-0.116777	0.094235	-0.044363	0.209452	-0.143834	-0.297343	0.088552	-0.6045
Gender	0.045994	1.000000	0.003878	0.066970	-0.277700	0.069374	-0.105671	0.045707	0.105073	0.108039	0.188989	0.017002	0.100819	-0.1297
CALC	-0.049882	0.003878	1.000000	-0.094375	-0.066488	-0.070936	-0.009208	-0.084127	-0.090530	0.035659	0.084518	0.054691	-0.041252	-0.0068
FAVC	0.062370	0.066970	-0.094375	1.000000	-0.025888	-0.012824	-0.193362	-0.051975	0.014378	0.214236	-0.096502	0.060735	0.149524	-0.0627
FCVC	0.013493	-0.277700	-0.066488	-0.025888	1.000000	0.036275	0.062903	0.014886	0.073930	0.045574	0.017572	-0.092804	-0.049592	0.0662
NCP	-0.043225	0.069374	-0.070936	-0.012824	0.036275	1.000000	-0.014885	0.008085	0.055862	0.078002	0.130907	0.035825	-0.092441	-0.0540
SCC	-0.116777	-0.105671	-0.009208	-0.193362	0.062903	-0.014885	1.000000	0.050798	0.011416	-0.178123	0.070513	-0.004350	-0.101119	0.0519
SMOKE	0.094235	0.045707	-0.084127	-0.051975	0.014886	0.008085	0.050798	1.000000	-0.033249	0.018067	0.011295	0.018034	-0.057157	-0.0113
CH2O	-0.044363	0.105073	-0.090530	0.014378	0.073930	0.055862	0.011416	-0.033249	1.000000	0.144962	0.166188	0.005526	0.156835	0.0487
FHWO	0.209452	0.108039	0.035659	0.214236	0.045574	0.078002	-0.178123	0.018067	0.144962	1.000000	-0.053125	0.017758	0.169453	-0.1054
FAF	-0.143834	0.188989	0.084518	-0.096502	0.017572	0.130907	0.070513	0.011295	0.166188	-0.053125	1.000000	0.062432	-0.017880	0.0031
TUE	-0.297343	0.017002	0.054691	0.060735	-0.092804	0.035825	-0.004350	0.018034	0.005526	0.017758	0.062432	1.000000	-0.061761	0.1774
CAEC	0.088552	0.100819	-0.041252	0.149524	-0.049592	-0.092441	-0.101119	-0.057157	0.156835	0.169453	-0.017880	-0.061761	1.000000	-0.0562
MTRANS	-0.604549	-0.129791	-0.006860	-0.062771	0.066298	-0.054061	0.051931	-0.011347	0.048756	-0.105489	0.003141	0.177487	-0.056225	1.0000
BMI	0.244793	-0.049838	-0.170224	0.246821	0.262139	0.037284	-0.180453	-0.000753	0.148356	0.485519	-0.174309	-0.103449	0.311445	0.0194
NObeyesdad	0.284058	-0.027544	-0.151801	0.247289	0.226015	0.024836	-0.190660	0.003698	0.137605	0.507164	-0.195341	-0.113177	0.326098	0.0082

Figure 4.7 - Pearson Correlation Matrix

5 Predictive Modeling

In this section, we detail the steps taken to develop and evaluate a predictive model for estimating obesity levels. The model employs a **RandomForestClassifier**, a robust and widely used machine learning algorithm known for its high performance in classification tasks.

5.1 Data Preparation

To begin, we loaded the preprocessed and label-encoded dataset (***le_data***). The target variable for our predictive modeling is ***NObeyesdad***, which represents the obesity levels. We excluded the ***Height*** and ***Weight*** column as they have already been used to calculate ***BMI*** and could introduce multicollinearity into the model. The features and target variable were then separated as follows:

```
X = le_data.drop(["NObeyesdad", "Height", "Weight"], axis=1)
y = le_data["NObeyesdad"]
```

5.2 Data Splitting

The dataset was split into training and testing sets to evaluate the model's performance on unseen data. We allocated **20%** of the data for testing, ensuring that the split was stratified to maintain the distribution of the target variable across both sets:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

5.3 Data Standardization

Standardization of the features was performed to ensure that all features contribute equally to the model, especially since **RandomForestClassifier** can benefit from standardized data. We used the ***StandardScaler*** to transform the training and testing sets.

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

5.4 Model Definition and Cross-Validation

We defined the ***RandomForestClassifier*** with a fixed random state to ensure reproducibility. To assess the model's performance, we employed **5-fold stratified cross-validation**, which helps in evaluating the model's generalization capability across different subsets of the training data.

```
# Define the RandomForestClassifier model
rf_clf = RandomForestClassifier(random_state=42)

# Set up K-fold cross-validation
kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Perform K-fold cross-validation
cv_scores = cross_val_score(rf_clf, X_train_scaled, y_train, cv=kfold,
scoring="accuracy")

print("Cross-validation scores:", cv_scores)
print("Mean cross-validation score:", cv_scores.mean())
```

5.5 Model Training and Prediction

Following **cross-validation**, the model was trained on the entire training set. Predictions were then made on the testing set to evaluate the model's performance on unseen data.

```
# Train the RandomForestClassifier model on the full training set
rf_clf.fit(X_train_scaled, y_train)

# Make predictions
y_pred = rf_clf.predict(X_test_scaled)
```

5.6 Model Evaluation

To evaluate the model, we generated a confusion matrix, classification report, and accuracy score. These metrics provide insights into the model's performance in terms of precision, recall, F1-score, and overall accuracy.

We also plotted the ROC curve of the machine learning model (Random Forest) and calculated the AUC. By binarizing each label and calculating the ROC multiple times(for each tag), we can evaluate the model performance by each criterion.

Cross-validation scores: [0.96932515 0.9601227 0.97230769 0.96307692 0.98461538]
Mean cross-validation score: 0.9698895705521473

Confusion Matrix:

```
[[45  2  0  0  0  0  0]
 [ 0 54  0  0  0  0  0]
 [ 0  6 40  1  0  0  0]
 [ 0  2  2 60  0  0  0]
 [ 0  0  0  0 70  2  0]
 [ 0  0  0  0  0 59  0]
 [ 0  0  0  0  0  0 64]]
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	47
1	0.84	1.00	0.92	54
2	0.95	0.85	0.90	47
3	0.98	0.94	0.96	64
4	1.00	0.97	0.99	72
5	0.97	1.00	0.98	59
6	1.00	1.00	1.00	64
accuracy			0.96	407
macro avg	0.96	0.96	0.96	407
weighted avg	0.97	0.96	0.96	407

Accuracy Score: 0.9631449631449631

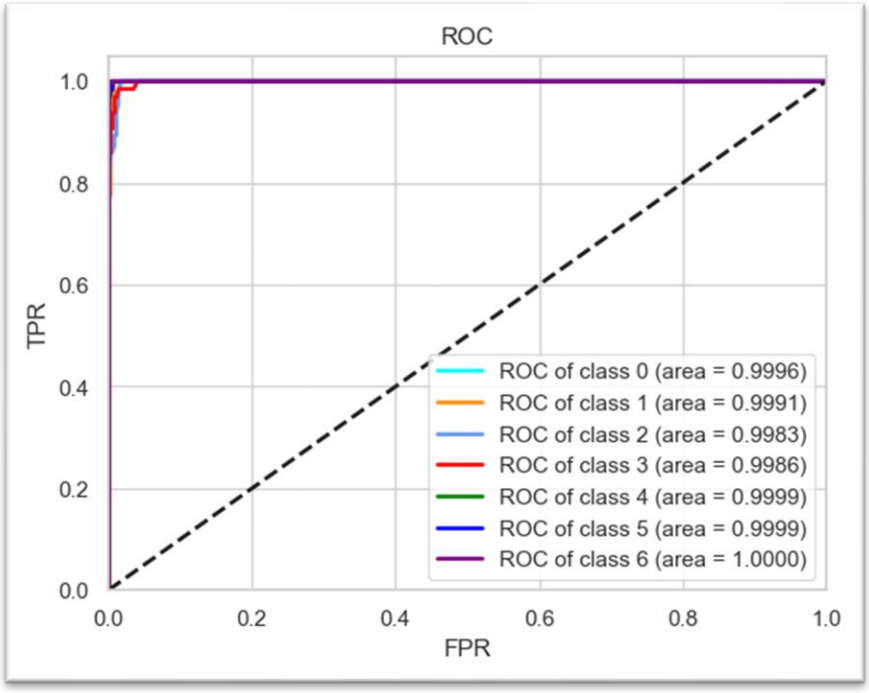

Figure 5.1 - ROC curve

Cross-Validation:

The cross-validation scores indicate the model's consistency and reliability across different subsets of the data. The mean cross-validation score of approximately **0.9698** suggests that the model performs consistently well across multiple folds, reflecting its generalizability and robustness.

Confusion Matrix:

The confusion matrix provides a granular view of the model's performance by showing the number of correct and incorrect predictions for each class. The high number of correct predictions across all classes indicates that the model has effectively learned to distinguish between different obesity levels. The few misclassifications observed are minimal and do not significantly impact the overall performance.

Classification Report:

The classification report offers a detailed breakdown of **precision**, **recall**, and **F1-score** for each class. The high **precision** and **recall** values across most classes indicate that the model is both accurate and reliable in its predictions. The **F1-scores**, which balance precision and recall, further affirm the model's strong performance.

Accuracy Score:

The overall accuracy score of approximately **0.9631** demonstrates that the model correctly predicts the obesity level for the vast majority of instances in the dataset. This high accuracy underscores the effectiveness of the selected features and the model's ability to capture the underlying patterns in the data.

5.7 Filling Missing Values

The primary objective of this section is to determine the most effective methods for handling missing data in the dataset to ensure the integrity and accuracy of the subsequent analysis. To address the issue of missing values in the target dataset, ***data_missing***, we

developed a function to predict and fill these missing values using a trained ***RandomForestClassifier***.

Age	Gender	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRANS	BMI	NObeyesdad
29.000000	Male	Sometimes	yes	2.0	3.0	no	no	2.000000	no	0.000000	0.000000	Sometimes	Automobile	20.195092	NaN
31.000000	Female	Sometimes	no	2.0	1.0	no	no	1.000000	yes	1.000000	0.000000	Sometimes	Public_Transportation	27.239224	NaN
21.000000	Female	no	yes	1.0	3.0	no	no	2.000000	yes	0.000000	1.000000	Frequently	Public_Transportation	26.291724	NaN
23.000000	Male	Frequently	no	2.0	3.0	no	no	2.000000	no	3.000000	1.000000	Sometimes	Public_Transportation	23.661439	NaN
22.000000	Female	no	no	2.0	1.0	no	no	2.000000	yes	3.000000	0.000000	Sometimes	Walking	21.777778	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
25.524336	Female	Sometimes	yes	3.0	3.0	no	no	1.436616	yes	0.167086	0.764717	Sometimes	Public_Transportation	37.614319	NaN
25.921678	Female	Sometimes	yes	3.0	3.0	no	no	1.031701	yes	0.034650	0.912345	Sometimes	Public_Transportation	39.419298	NaN
25.940153	Female	Sometimes	yes	3.0	3.0	no	no	1.000536	yes	0.005939	0.566353	Sometimes	Public_Transportation	40.128624	NaN
25.289428	Female	Sometimes	yes	3.0	3.0	no	no	1.299194	yes	0.234303	0.946888	Sometimes	Public_Transportation	36.856473	NaN
25.722004	Female	Sometimes	yes	3.0	3.0	no	no	2.487070	yes	0.067329	0.455823	Sometimes	Public_Transportation	40.430771	NaN

Figure 5.2 - Datasets with Missing Target Variables

Gender	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	FHWO	FAF	TUE	CAEC	MTRANS	BMI	NObeyesdad
Male	Sometimes	yes	2.0	3.0	no	no	2.000000	no	0.000000	0.000000	Sometimes	Automobile	20.195092	Normal_Weight
Female	Sometimes	no	2.0	1.0	no	no	1.000000	yes	1.000000	0.000000	Sometimes	Public_Transportation	27.239224	Overweight_Level_II
Female	no	yes	1.0	3.0	no	no	2.000000	yes	0.000000	1.000000	Frequently	Public_Transportation	26.291724	Normal_Weight
Male	Frequently	no	2.0	3.0	no	no	2.000000	no	3.000000	1.000000	Sometimes	Public_Transportation	23.661439	Normal_Weight
Female	no	no	2.0	1.0	no	no	2.000000	yes	3.000000	0.000000	Sometimes	Walking	21.777778	Normal_Weight
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Female	Sometimes	yes	3.0	3.0	no	no	1.436616	yes	0.167086	0.764717	Sometimes	Public_Transportation	37.614319	Obesity_Type_III
Female	Sometimes	yes	3.0	3.0	no	no	1.031701	yes	0.034650	0.912345	Sometimes	Public_Transportation	39.419298	Obesity_Type_III
Female	Sometimes	yes	3.0	3.0	no	no	1.000536	yes	0.005939	0.566353	Sometimes	Public_Transportation	40.128624	Obesity_Type_III
Female	Sometimes	yes	3.0	3.0	no	no	1.299194	yes	0.234303	0.946888	Sometimes	Public_Transportation	36.856473	Obesity_Type_III
Female	Sometimes	yes	3.0	3.0	no	no	2.487070	yes	0.067329	0.455823	Sometimes	Public_Transportation	40.430771	Obesity_Type_III

Figure 5.3 - Datasets Filled with Classifier Model Predictions

5.7 User Input Prediction

The objective of this section is to develop a user-oriented interface to allow users to input their details and predict their obesity level using the trained model. To achieve this, we developed a user-friendly interface that leverages a trained ***RandomForestClassifier*** to predict an individual's obesity level based on their input data.

The interface guides users through the process of entering their information and provides clear instructions on the required input format. This user-centric design ensures that individuals can easily use the tool without requiring extensive technical knowledge.

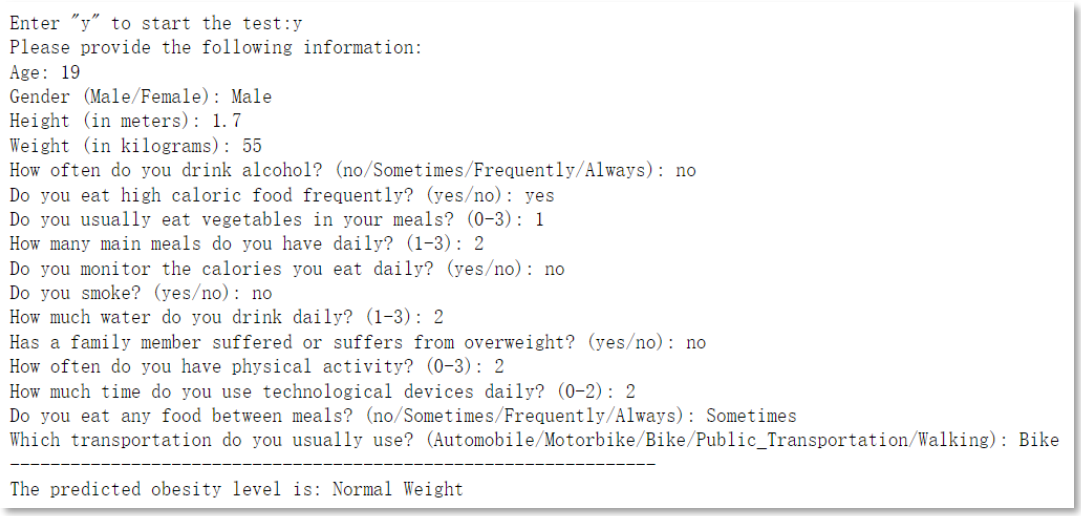

Figure 5.4 - User-Friendly Interface for Obesity Prediction

6 Visualization and Insights

6.1 Proportion of People with Different Types of Obesity

To understand the distribution of different obesity levels within the dataset, we performed an analysis to count the occurrences of each obesity type. The results are visualized in a pie chart to provide a clear representation of the proportion of people falling into each category.

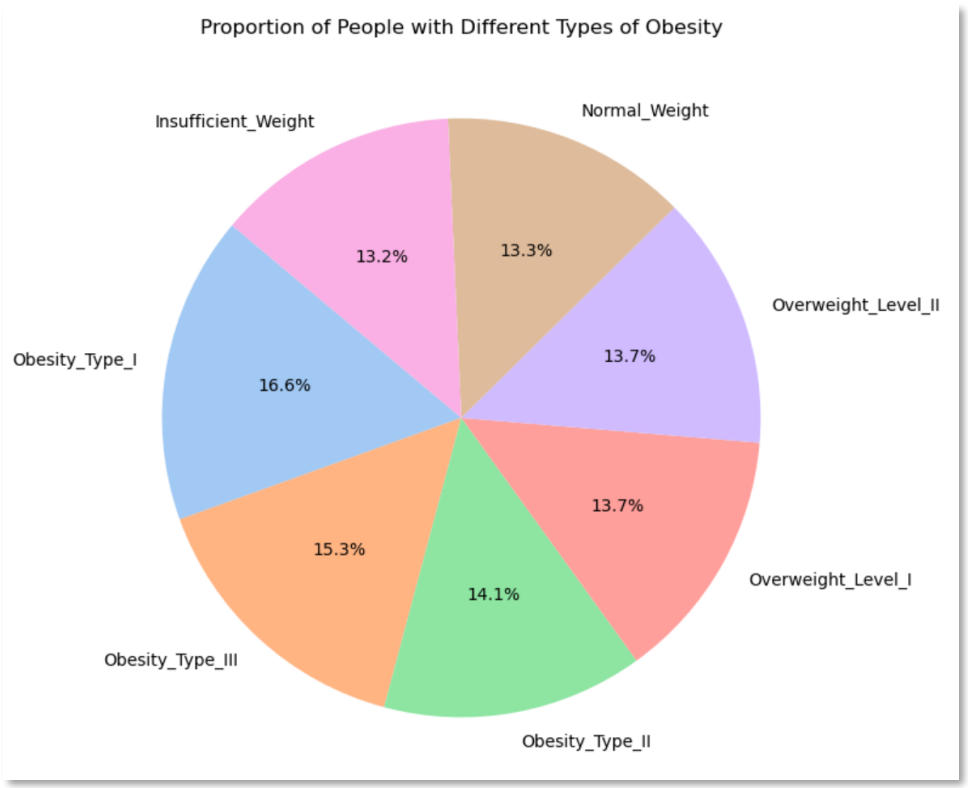

Figure 6.1 - Proportion of People with Different Types of Obesity

The pie chart reveals a relatively even distribution across different obesity levels. The balanced distribution across overweight and obesity levels underscores the complexity and varied nature of obesity within the dataset. This distribution can inform targeted interventions and strategies, ensuring that efforts to address obesity are inclusive of all weight categories, from insufficient to severe obesity.

Because of this feature, we can do machine learning modeling without significant bias in the data, resulting in a better fit of the model.

6.2 Handling Categorical Variables

In order to more intuitively reflect the distribution of categorical variables, we have drawn corresponding histograms

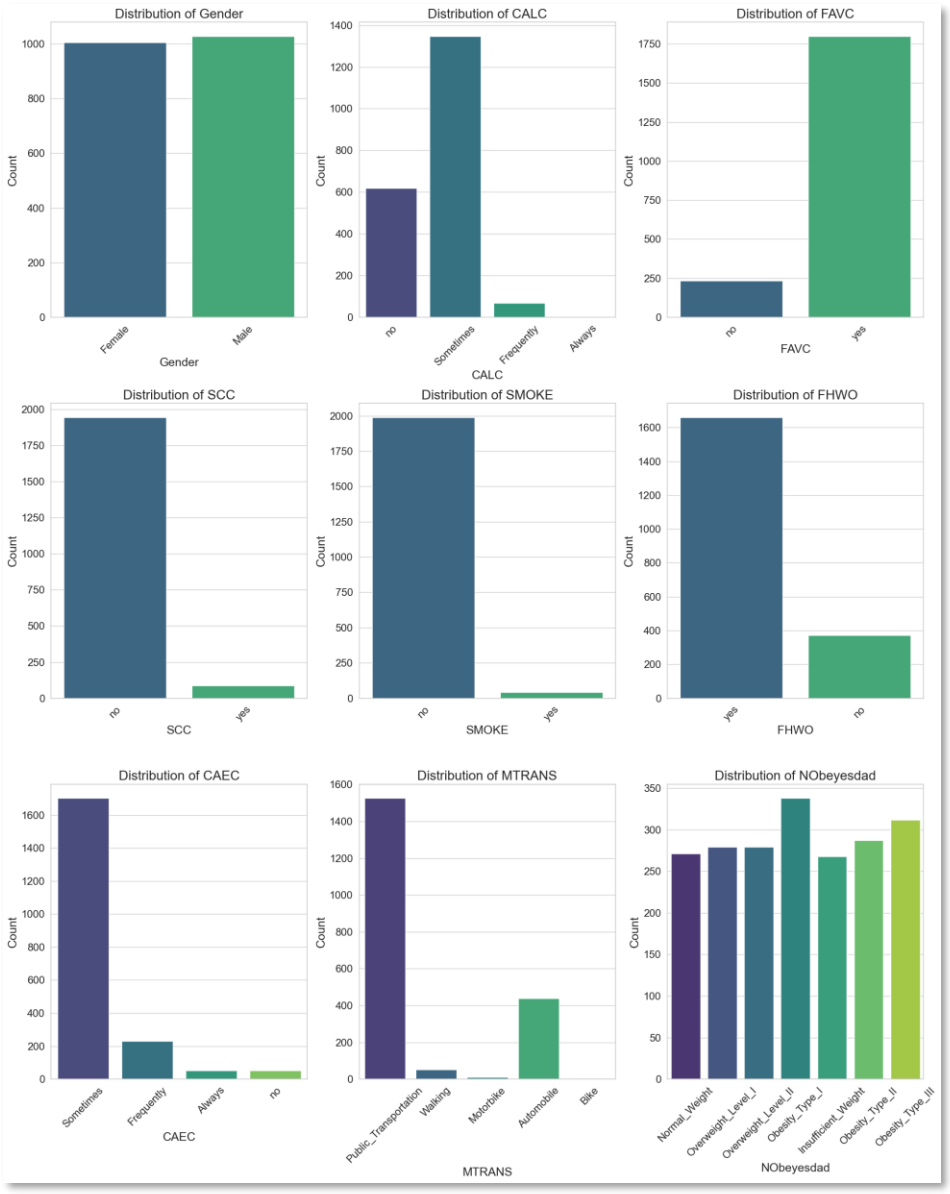

Figure 6.2 - The distribution of all categorical variables

After drawing the histogram, we can observe the distribution of different variables very clearly. At the same time, we can also conduct a preliminary analysis of the data. The following is our observation and analysis.

Observations:

- In the surveyed population, the distribution of gender and obesity types is relatively uniform, and the number of people in each category is similar.
- The drinking habits of the majority of the surveyed population are generally non alcoholic or occasional.
- Most of the surveyed population frequently consume high calorie foods, with only about one in seven people not frequently consuming high calorie foods.
- Almost no one monitors their calorie intake.
- Most of the surveyed population are non-smokers
- Many surveyed individuals have a family history of being overweight or obese, but about a quarter of them do not have a family history of being overweight or obese.
- The vast majority of the surveyed population will engage in inter meal eating, but the frequency is not high for most people.
- In terms of transportation, public transportation is the most common way, followed by walking and cars.

Analysis:

- The objectivity of the population surveyed in this dataset can be demonstrated by the gender distribution and obesity type distribution of the surveyed population.
- Because the number of people with normal weight and the number of people with underweight only account for two-thirds of the total population, if there are characteristics that only these two categories have, it may lead to a small amount of data, but we cannot ignore these data. Because these data features are likely the true cause of obesity.
- It is found from the data that about one in seven people do not frequently consume high calorie foods, which is likely to be mostly in the normal weight and underweight population. This factor is likely to be an important factor leading to weight gain and obesity. Of course, we need further analysis to determine.

- About a quarter of people do not have a family history of overweight or obesity, which may also indicate that some people's obesity is potentially hereditary.
- Whether the diet between meals can have little effect on obesity, what is important is likely the frequency of the diet between meals.

Afterwards, we drew corresponding histograms for different populations (overweight, normal, lean) and variables of interest, and plotted them in the same graph for easy comparison. This helps us to make preliminary conclusions on whether some variables affect obesity.

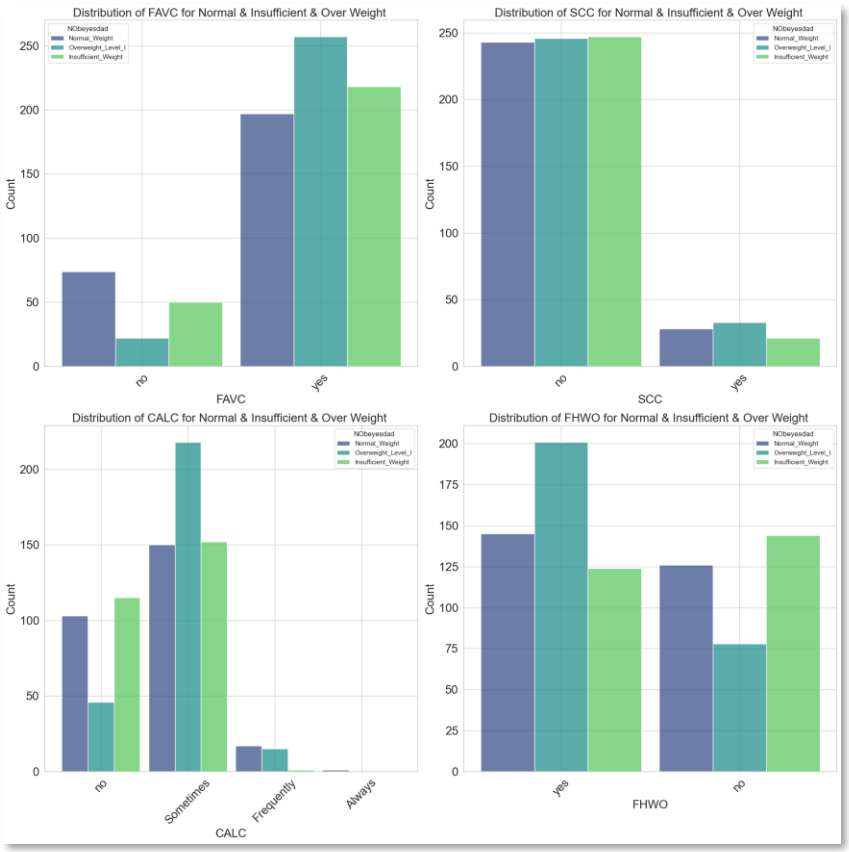

Figure 6.3 - Reflect of different obesity in FAVC, SCC, CALC, and FHWO (rough)

Conclusion:

- About 150 people in the normal weight and underweight population do not frequently consume high calorie foods, accounting for half of the population who

do not consume high calorie foods. This is a good indication that consuming high calories is related to obesity.

- Whether to monitor calorie intake is not related to obesity. From the graph, it can be observed that the distribution of overweight and normal weight individuals is similar.
- Sometimes drinking alcohol may be related to obesity, and there are more non drinking individuals in the normal weight and underweight population.
- Obesity may be related to family genetics, but there is no necessary relationship.

6.3 Handling Continuous Variables

For continuous variables, we not only care about their distribution but also about the distribution trend of the data. So we have plotted histograms of Age, BMI FCVC, NCP, CH2O, FAF, TUE and other variables one by one, and there are curves showing their trends. This helps us to have a preliminary understanding of the meaning of these continuous variables.

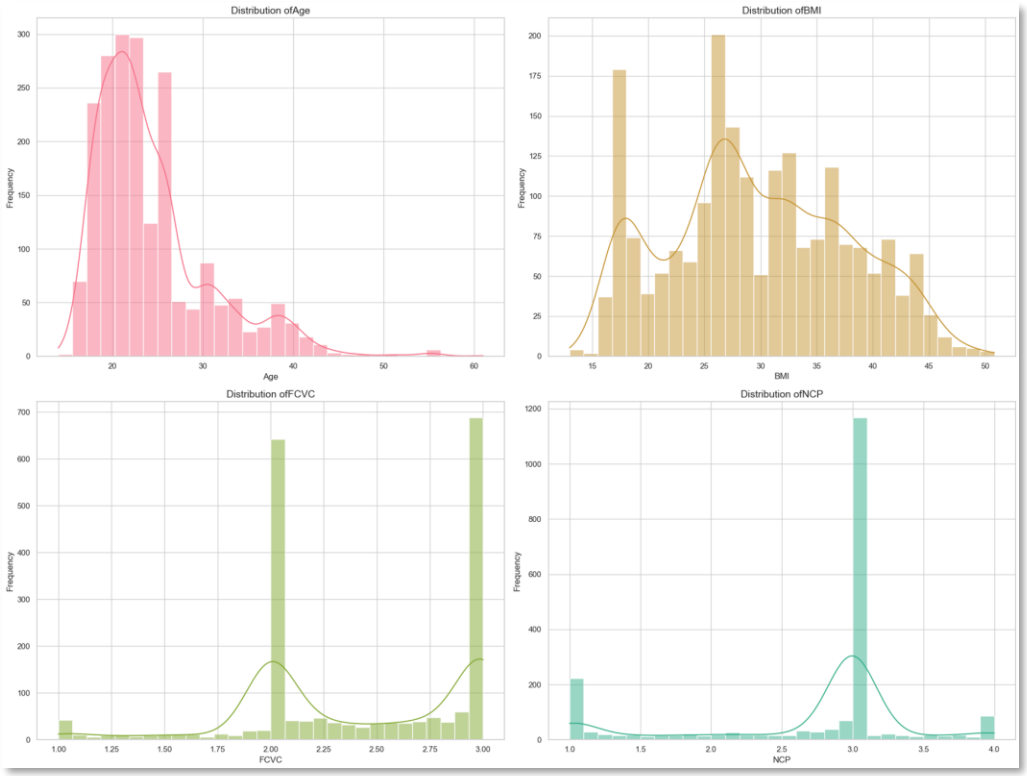

Figure 6.4 - The distribution of all continuous variables(1)

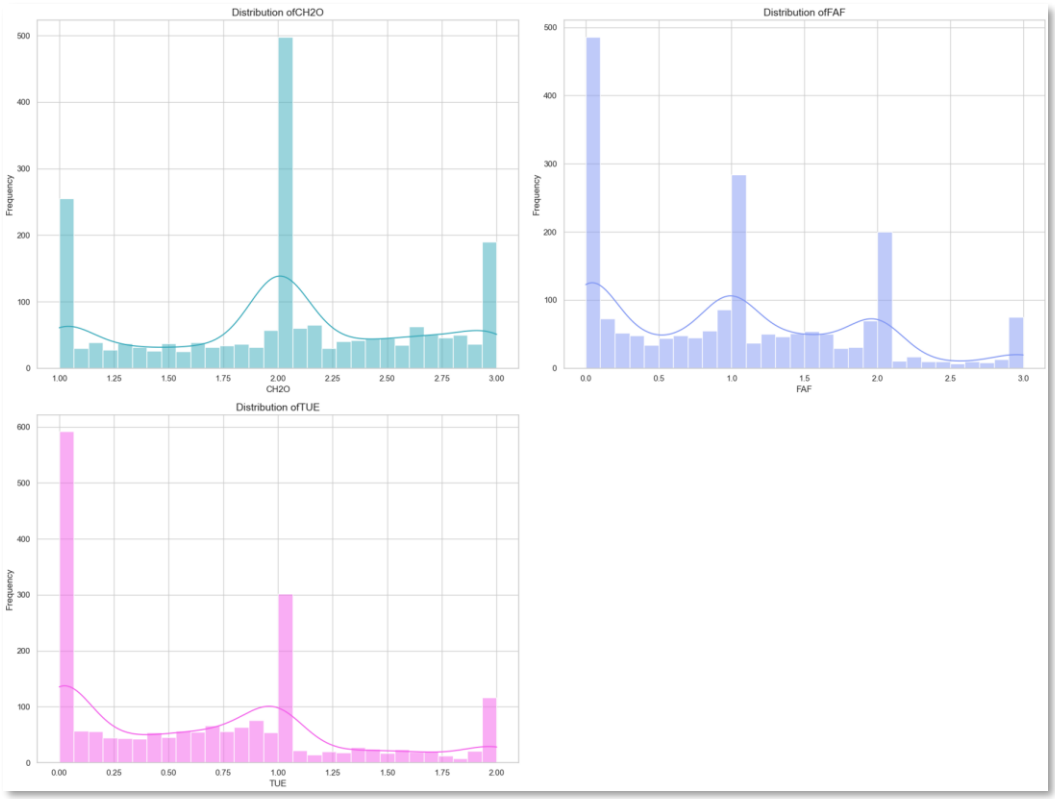

Figure 6.5 - The distribution of all continuous variables(2)

The following is our observations and analysis about continuous variable.

Observations:

- The majority of respondents are aged between 20 and 30, especially in the age range of 20 to 25. A small number of respondents are over 50 years old.
- The BMI value shows a bimodal distribution, with one peak around 20 and the other peak around 30. This indicates that the data is concentrated in two main BMI ranges, one is the normal range and the other is the obesity range.
- The frequency of vegetable intake is mainly concentrated between 2 and 3, indicating that the majority of respondents frequently consume vegetables.
- Most respondents have 3 main meals per day
- The daily water intake is very uniform, and the data is only concentrated between 1, 2, and 3.

- Most respondents exercise 1-2 times a week.
- The time spent using technical equipment is concentrated between 0 and 1, indicating less usage time

Analysis:

- From the age distribution of the surveyed population, it can be seen that this obesity study mainly targets young people.
- The majority of BMI is greater than 25, indicating that the surveyed population is mainly obese, but there are also many normal and lean individuals. This is consistent with the data from NObeyesdad mentioned earlier.
- The frequency of vegetable intake may not be related to obesity, perhaps we should ask more about the proportion of vegetables consumed in the main diet
- The amount of exercise should be closely related to obesity. Fewer people with more physical activity

In order to investigate whether there are obvious outliers in these continuous variables and whether these outliers actually exist, we have drawn corresponding box plots. Box plots can provide us with a very intuitive representation of outliers in data, and we can also analyze these outliers to see if they are reasonable and if they need to be removed. At the same time, we can also visually observe the key values of each data, such as what the median is.

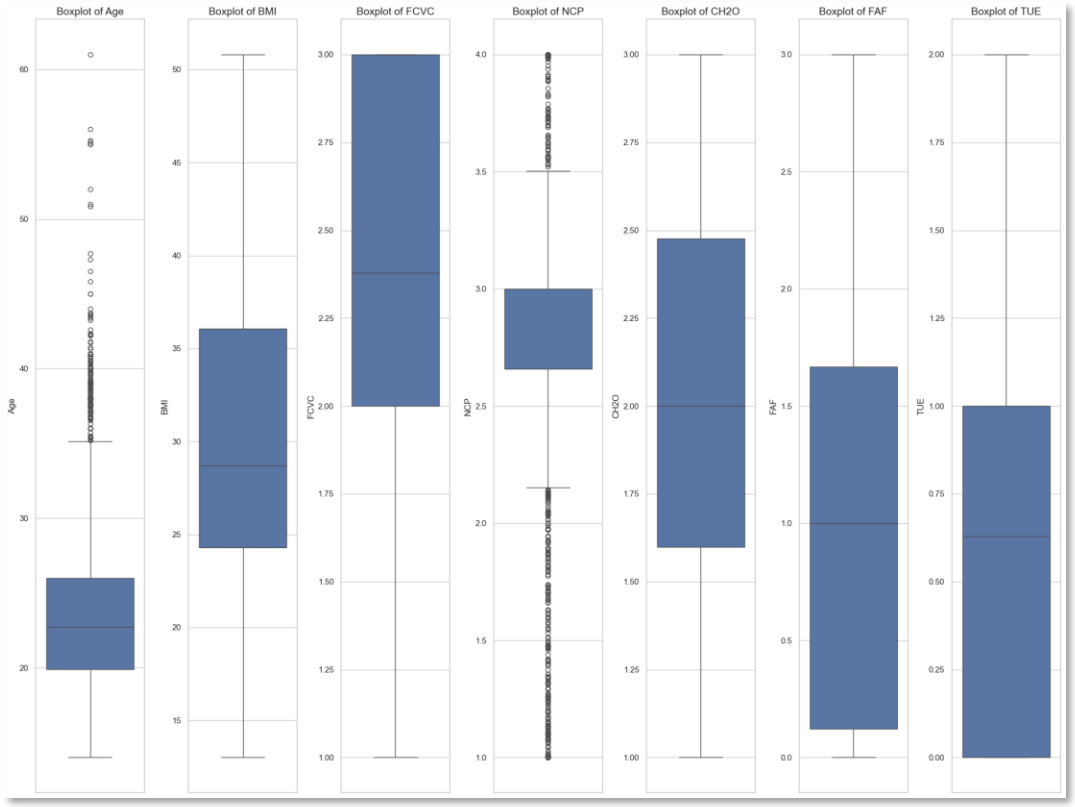

Figure 6.6 - Box plots of continuous variables

The following is our observations and analysis about the box plots of continuous variables.

Observations:

- The majority of respondents are aged between 20 and 30 years old. People over 40 years old are outliers, especially those close to 60 years old.
- The median BMI value is approximately 28, with a distribution concentrated between 25 and 35, and no significant outliers.
- The median frequency of vegetable intake is 2.5, with most data distributed between 2 and 3 and no significant outliers.
- The median number of main meals is 3, with a relatively concentrated distribution, but there are outliers at times 1 and 4.
- The median water consumption is 2.0 liters, and the data distribution is relatively concentrated with no significant outliers.

- The median frequency of exercise per week is 1.0, with a wide distribution but no significant outliers.
- The median time spent using technical equipment is 1.0 hours, with a relatively concentrated distribution and no significant outliers.

Analysis:

- The vast majority of the data are excellent with no significant outliers. Very advantageous for analysis.
- To ensure age diversity, it is not necessary to remove these outliers.
- For the amount of drinking water, considering that different people have different needs for drinking water, and as can be seen from the previous data distribution uniformity, these abnormal values do not need to be removed.

In order to analyze and visualize the variability of selected variables in a dataset, we calculate the standard deviation, sort the variables, and create an interactive bar chart.

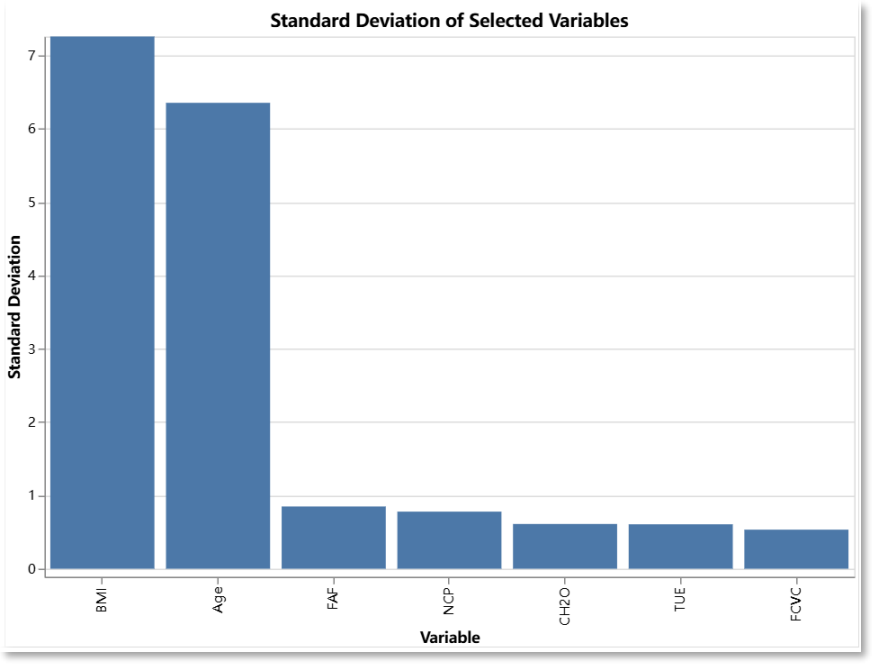

Figure 6.7 - Histogram of continuous variables' standard deviation

From the above figure, we can draw the following observations and conclusions.

Observations:

- This graph analyzes the standard deviation of continuous variables. The variance of a ll variables except BMI and age is not significant.

Analysis:

- The standard deviation of BMI and Age indicates that the dataset has a wide range of interviewees and covers a wide range of topics
- The standard deviation of the remaining data is very small because there is not much difference in the numerical values of the answers, which is limited by the human bod y. For example, most people can only exercise for less than 3 hours a day, which is n ot very realistic. This demonstrates the accuracy of our dataset.

6.4 Correlation Heatmap of Obesity-Related Factors

To gain a deeper understanding of the relationships between the various features in our dataset, we created a heatmap of the Pearson correlation coefficients. This visualization helps us identify which features are strongly correlated with each other, providing insights into potential multicollinearity issues and highlighting important predictors for our target variable, `NObeyesdad`.

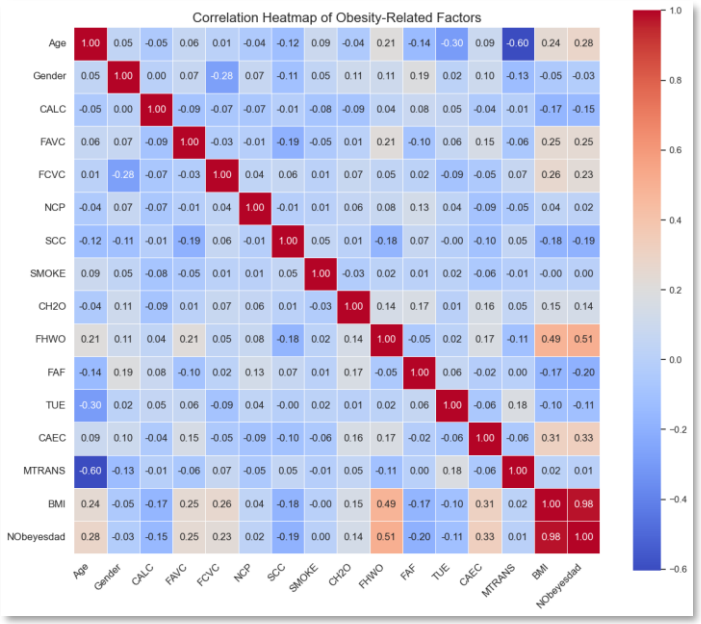

Figure 6.7 - Histogram of continuous variables' standard deviation

The heatmap provides a visual representation of the correlations between the features in our dataset. Strong positive correlations are represented by dark red colors, while strong negative correlations are represented by dark blue colors. This helps us visually display the correlation between various features.

7 Feature Importance and Factor Analysis

7.1 Demographic and Lifestyle Factors Linked to Obesity Levels

To analyze the relationship between demographic factors (such as **age**, **gender**, and **family history**) and lifestyle choices (such as **smoking** and **alcohol** consumption) with obesity levels, we performed a correlation analysis using the dataset. The correlation coefficients between individual variables and obesity levels (***NObeyesdad***) were calculated, and the key findings from the correlation matrix are summarized below.

	NObeyesdad
FHWO	0.507164
Age	0.284058
CALC	-0.151801
Gender	-0.027544
SMOKE	0.003698

Figure 7.1 - Demographic and Lifestyle Factors Linked to Obesity Levels

- **Family History with Overweight (FHWO)**:

  This factor has the highest positive correlation with obesity levels (**0.507164**). This suggests that individuals with a family history of overweight are more likely to have higher obesity levels. This finding highlights the potential genetic or familial lifestyle influences on obesity.

- **Age:**

  The correlation coefficient for age is **0.284058**, indicating a moderate positive relationship with obesity levels. As age increases, there is a tendency for obesity levels to increase. This could be due to various factors such as changes in metabolism, physical activity levels, and lifestyle habits over time.

- **Gender:**

  Gender has a very weak negative correlation with obesity levels (**-0.027544**), suggesting that gender differences alone do not significantly impact obesity levels in this dataset.

- **Consumption of Alcohol (CALC):**

  The correlation coefficient for alcohol consumption is **-0.151801**, indicating a slight negative relationship with obesity levels. This means that individuals who consume alcohol more frequently tend to have slightly lower obesity levels. This counterintuitive result could be influenced by other confounding factors such as the type of alcohol consumed, associated dietary habits, and lifestyle choices.

- **Smoking (SMOKE):**

  The correlation coefficient for smoking is almost negligible (**0.003698**), indicating no significant relationship between smoking habits and obesity levels in this dataset.

These factors were grouped into two major categories for further analysis:

Demographic Factors: Age, Gender, Family History with Overweight (FHWO)

Lifestyle Factors: Consumption of Alcohol (CALC), Smoking (SMOKE)

When considered as a whole, The aggregated correlation of demographic factors with obesity levels is significantly higher. This suggests that inherent characteristics such as family history of overweight and age play a crucial role in determining obesity risk. The

strong correlation with family history underscores the importance of genetic predispositions and potentially inherited lifestyle patterns in the development of obesity. In contrast, lifestyle factors show a weaker overall correlation with obesity levels. Although lifestyle choices like alcohol consumption and smoking are important for overall health, their direct impact on obesity appears to be less significant in this dataset.

7.2 Eating Habits, Physical Activity and Sedentary Behavior about Obesity Levels

To identify specific dietary patterns that contribute to higher or lower obesity levels and assess the impact of physical activity and sedentary behavior, we performed a correlation analysis using the dataset. The correlation coefficients between individual variables and obesity levels (NObeyesdad) were calculated, and the key findings from the correlation matrix are summarized below.

	NObeyesdad
CAEC	0.326098
FAVC	0.247289
FCVC	0.226015
FAF	-0.195341
SCC	-0.190660
CH2O	0.137605
TUE	-0.113177
NCP	0.024836
MTRANS	0.008208

Figure 7.2 - Eating Habits, Physical Activity and Sedentary Behavior Linked to Obesity Levels

- **Consumption of Food Between Meals (CAEC):** This factor has the highest positive correlation with obesity levels (0.326098). This suggests that individuals who frequently consume food between meals are more likely to have higher obesity levels. This finding highlights the potential impact of snacking habits on obesity.

- **Frequent Consumption of High-Caloric Food (FAVC):** The correlation coefficient for frequent consumption of high-caloric food is 0.247289, indicating a moderate positive relationship with obesity levels. Individuals who often consume high-caloric foods tend to have higher obesity levels, suggesting the influence of diet quality on obesity.
- **Frequency of Consumption of Vegetables (FCVC):** The correlation coefficient for vegetable consumption is 0.226015, suggesting a moderate positive relationship with obesity levels. This result may seem counterintuitive, possibly due to confounding dietary habits or portion sizes not captured in the data.
- **Physical Activity Frequency (FAF):** The correlation coefficient for physical activity frequency is -0.195341, indicating a negative relationship with obesity levels. This suggests that individuals who engage in physical activity more frequently tend to have lower obesity levels, underscoring the importance of regular exercise.
- **Calories Consumption Monitoring (SCC):** The correlation coefficient for monitoring calorie intake is -0.190660, showing a negative relationship with obesity levels. Individuals who monitor their calorie consumption tend to have lower obesity levels, highlighting the role of dietary awareness and self-regulation in weight management.
- **Daily Water Consumption (CH2O):** The correlation coefficient for daily water consumption is 0.137605, indicating a weak positive relationship with obesity levels. This suggests that higher water intake alone may not be a strong determinant of obesity.
- **Time Using Technology Devices (TUE):** The correlation coefficient for time spent using technology is -0.113177, suggesting a weak negative relationship with obesity levels. This indicates that more time spent on technological devices may be associated with slightly lower obesity levels, possibly reflecting sedentary behavior not being fully captured by technology usage alone.
- **Number of Main Meals (NCP):** The correlation coefficient for the number of main meals is 0.024836, indicating a negligible relationship with obesity levels. This suggests that the number of main meals per day does not significantly impact obesity in this dataset.

- **Mode of Transportation (MTRANS):** The correlation coefficient for transportation mode is 0.008208, showing an almost negligible relationship with obesity levels. This suggests that the type of transportation used does not have a significant impact on obesity in this dataset.
- When aggregating the findings, we can categorize the factors into three major groups: Eating Habits, Physical Activity, and Sedentary Behavior to better understand their overall impact on obesity levels.
- **Eating Habits:** This category includes CAEC, FAVC, FCVC, SCC, CH2O, and NCP. The combined impact of these factors shows a significant relationship with obesity levels. Frequent consumption of food between meals (CAEC) and high-caloric food (FAVC) are strongly associated with higher obesity levels, highlighting the importance of diet quality and snacking habits in obesity. Monitoring calorie intake (SCC) shows a negative relationship, emphasizing the role of dietary awareness.
- **Physical Activity:** This category includes FAF. The combined analysis indicates that physical activity frequency (FAF) has a notable negative correlation with obesity levels, underscoring the importance of regular exercise in managing weight.
- **Sedentary Behavior:** This category includes TUE. Time spent using technology devices (TUE) shows a weak negative correlation, suggesting that while sedentary behavior is important, it may not be fully captured by technology usage alone.

7.3 Correlated Independent Variables Pairs

When carrying out **Factor Analysis**, we focus on each variable that affects obesity, but instead of studying their influence on obesity, we focus on the correlation between Independent Variables.

	Variable 1	Variable 2	Correlation Coefficient
0	Age	MTRANS	-0.604549
1	FHWO	BMI	0.485519
2	CAEC	BMI	0.311445
3	Age	TUE	-0.297343
4	Gender	FCVC	-0.277700
5	FCVC	BMI	0.262139
6	FAVC	BMI	0.246821
7	Age	BMI	0.244793
8	FAVC	FHWO	0.214236
9	Age	FHWO	0.209452

Figure 7.3 - Correlated Independent Variables

In order to better understand the scatter distribution and potential linear relationships between each pair of variables. We selected the top 5 correlated variables, namely Age, FCVC, FHWO, BMI, and TUE. Then we generate a pair plot to visualize the relationships between these variables.

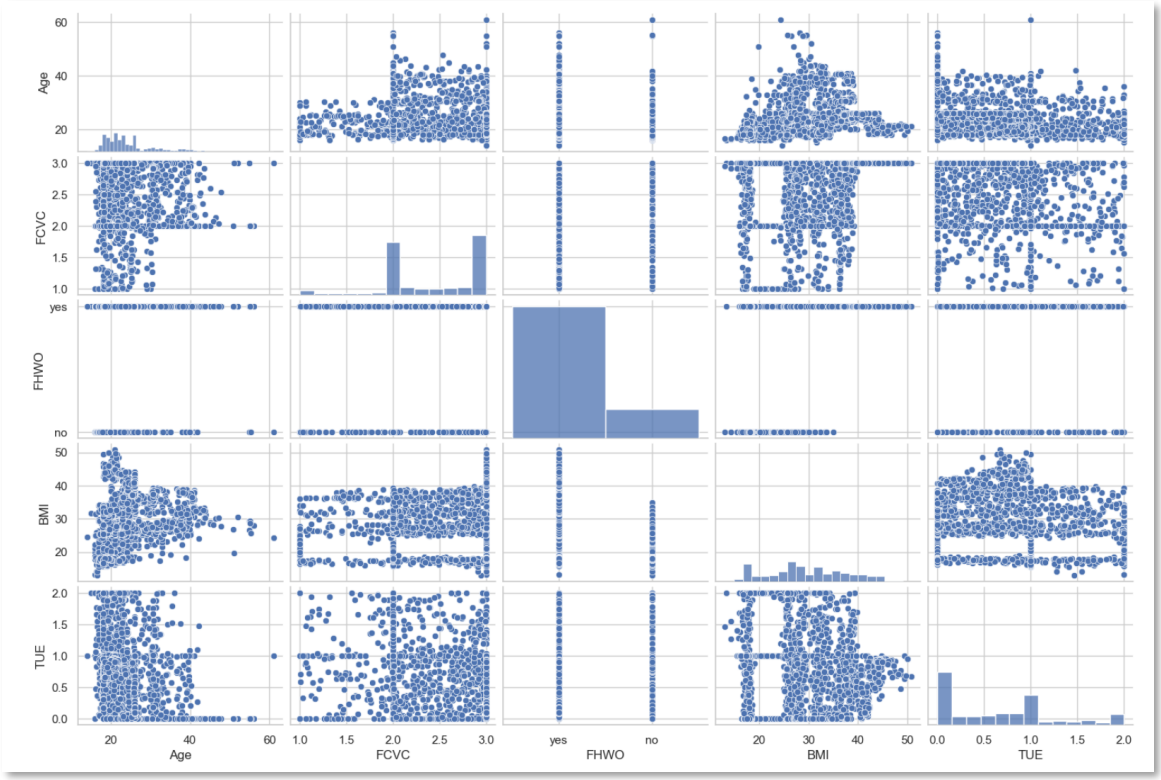

Figure 7.4 - Pair Plot of Top 5 Correlated Variables

Based on the pair plot generated, we can derive several insights from the most correlated pairs of variables, and we get some **interesting finding**:

Age and MTRANS:

Correlation Coefficient: -0.604549

Interpretation: As age increases, the use of transportation modes such as walking and public transportation decreases. This suggests that older individuals prefer using automobiles or motorbikes over walking or using public transport.

FHWO and BMI:

Correlation Coefficient: 0.485519

Interpretation: There is a moderate positive correlation between having a family history of overweight and the individual's BMI. Individuals with a family history of overweight tend to have higher BMI values, indicating a possible genetic or lifestyle influence.

CAEC and BMI:

Correlation Coefficient: 0.311445

Interpretation: The frequency of eating food between meals is positively correlated with BMI. People who often eat between meals are likely to have a higher BMI.

Age and TUE:

Correlation Coefficient: -0.297343

Interpretation: There is a negative correlation between age and time spent using technology. Younger individuals spend more time on technological devices compared to older individuals.

Gender and FCVC:

Correlation Coefficient: -0.277700

Interpretation: Males and females differ in their consumption of vegetables, with females generally consuming vegetables more frequently.

7.4 Correlation Between Various Within Groups With Different Levels of Obesity

The primary objective of this analysis is to explore the correlation between various behaviors within groups characterized by different levels of obesity. For example, non-obese individuals might display different patterns of behavior interaction compared to obese individuals. Identifying these differences can guide the creation of specific interventions that address the unique needs of each group.

We focuses on comparing the correlation patterns between two distinct obesity groups: **Insufficient Weight (NObeyesdad == 0)** and **Obesity Type II (NObeyesdad == 5)**, highlighting how the correlation between features differs between the two groups.

	Feature 1	Feature 2	Correlation Difference
0	CH2O	CALC	1.018997
1	MTRANS	Age	0.817028
2	MTRANS	FAF	-0.785771
3	TUE	FCVC	-0.740091
4	NCP	CALC	0.705706
5	FHWO	CALC	0.643776
6	FHWO	NCP	0.594968
7	CAEC	Age	-0.575116
8	CAEC	FHWO	-0.525381
9	MTRANS	Gender	-0.490420

Figure 7.5 - Correlation Between Various in Groups With Vary Levels of Obesity

Hydration and Alcohol Consumption (CH2O and CALC):

In the Obesity Type II group, there is a stronger positive correlation between water intake and alcohol consumption. This suggests that individuals with Obesity Type II who consume more water are also likely to consume more alcohol, indicating that their hydration habits are closely tied to their drinking habits.

Mode of Transportation and Age (MTRANS and Age):

In the Obesity Type II group, age has a more significant positive influence on transportation choices. Older individuals in this group are more likely to choose specific modes of transportation such as automobiles or motorbikes over walking or public transportation, reflecting age-related preferences or limitations.

Transportation and Physical Activity Frequency (MTRANS and FAF):

In the Insufficient Weight group, there is a stronger negative correlation between the mode of transportation and physical activity frequency. This suggests that individuals in this group who use active transportation modes (e.g., walking, biking) are more likely to engage in frequent physical activity compared to those with Obesity Type II.

Technology Use and Vegetable Consumption (TUE and FCVC):

In the Insufficient Weight group, higher usage of technological devices is more negatively correlated with the frequency of vegetable consumption. This implies that individuals who spend more time using technology tend to consume fewer vegetables, might indicating a potential trade-off between sedentary behavior and healthy eating habits.

Meal Frequency and Alcohol Consumption (NCP and CALC):

In the Obesity Type II group, there is a stronger positive correlation between the number of main meals consumed daily and alcohol consumption. This suggests that individuals who have more meals per day are also more likely to consume alcohol frequently.

7.5 Conclusion

Among the topics in ***7.1 and 7.2*** we analyzed, **Family History with Overweight** (FHWO) has the highest correlation coefficient with obesity levels, at 0.507164. This suggests that having a family history of overweight is the single most significant predictor of obesity in this dataset. This finding underscores the importance of genetic or familial influences on obesity, indicating that individuals with a family history of overweight are more likely to have higher obesity levels.

When comparing the overall impact of the five major categories, it is evident that **Demographic Factors** have the most significant combined correlation with obesity levels. This suggests that inherent characteristics such as family history of overweight and age play a crucial role in determining obesity risk. In contrast, Lifestyle Factors show a weaker overall correlation, and while Eating Habits and Physical Activity are also important, their combined impact is less pronounced than that of Demographic Factors.

8 Conclusions and Recommendations

8.1 Conclusion of question.

Finally, we would like to answer the questions raised in the proposal.

1. **What demographic and lifestyle factors are most strongly associated with different levels of obesity?**

   Answer: From 7.1, we can know that the most significant demographic factor associated with obesity levels is a family history of overweight (FHWO), with a correlation coefficient of 0.507164. This highlights the strong genetic or familial influence on obesity. Age is another critical demographic factor, influencing obesity risk and transportation choices among different groups.

   Among lifestyle factors, alcohol consumption, meal frequency, and technology use show varying degrees of association with obesity levels. But it's not very obvious.

2. **Can we develop a predictive model to classify individuals into different obesity levels based on their eating habits and physical activities?**

   Answer: Of course! In our part 5, we developed a predictive model. We used the ***RandomForestClassifier*** model, which combines demographic factors with lifestyle choices, providing a reliable basis for predicting obesity levels. The developed model allows users to input their demographic factors and lifestyle choices, thereby obtaining their own obesity type.

3. **How do eating habits, physical activity, and sedentary behavior impact obesity levels?**

Answer: Dietary habits and physical activity significantly affect obesity levels. In this experiment, we found it difficult to determine the impact of sedentary behavior on obesity. Because the data related to sedentary behavior refers to the usage time of the device, and prolonged use of the device does not necessarily mean prolonged sitting.

From the conclusion (in 7.2), we can conclude that Consumption of Food Between Meals (CAEC) significantly affects obesity levels and is positively correlated. The Physical Activity Frequency (FAF) also shows a negative correlation with obesity levels. Time Using Technology Devices (TUE) has a negative correlation with obesity levels. This indicates that the shorter the time the device is used, the fatter it will be. Considering that using devices may not necessarily be sedentary behavior. So, we cannot accurately determine the impact of sedentary behavior on obesity levels.

4. **Is there a correlation between various behaviors within groups with different levels of obesity?**

   Answer: Yes, there are notable correlations between behaviors within groups with different levels of obesity. For instance, individuals with higher obesity levels who consume more water are also likely to consume more alcohol. Additionally, older individuals in the Obesity Type II group tend to prefer motorized transportation over active modes.

   In contrast, those with insufficient weight show a stronger negative correlation between active transportation and physical activity frequency.

5. **How can we impute missing values in the dataset?**

   Answer: In 3.2, We have mentioned that. The missing values in the dataset can be imputed by using predictive modeling. Given the small proportion of missing values (approximately 3.65%), their impact on the overall analysis is minimal. The approach involves training models on the complete data and then predicting the missing values based on the patterns learned from the data.

8.2 Recommendations

- **Focus on Family History and Demographic Factors:**

A strong correlation between family history and obesity levels. This may be closely related to the family environment, but the role of genetics cannot be ignored. In short, for this group of people, we suggest that they exercise more, eat more vegetables, have a healthy diet frequency, and try not to choose mobile oriented transportation. This will greatly help their physical health overcome obesity.

- **Calories Consumption Monitoring:**

For people who have already been found to be obese or even overweight, monitoring their calorie intake is a good solution. Although there is not much behavior of monitoring calorie intake in daily life, it shows a significant negative correlation with obesity levels. This is likely because people who monitor calorie intake behavior tend to exercise more and eat more vegetables.

- **Promote Physical Activity:**

Programs encouraging physical activity, especially those promoting active transportation modes, should be developed to combat obesity.

9 References

1. https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels/data
2. https://crunchingthedata.com/data-science-project-proposals/
3. https://www.slideteam.net/blog/top-10-data-science-proposal-templates-with-examples-and-samples
4. Sebastian Raschka and Vahid Mirjalili. 2017. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition (2nd. ed.). Packt Publishing.
5. Kendall, Kenneth E. and Julie E. Kendall. 2019. Systems Analysis & Design, Tenth Edition. Pearson.