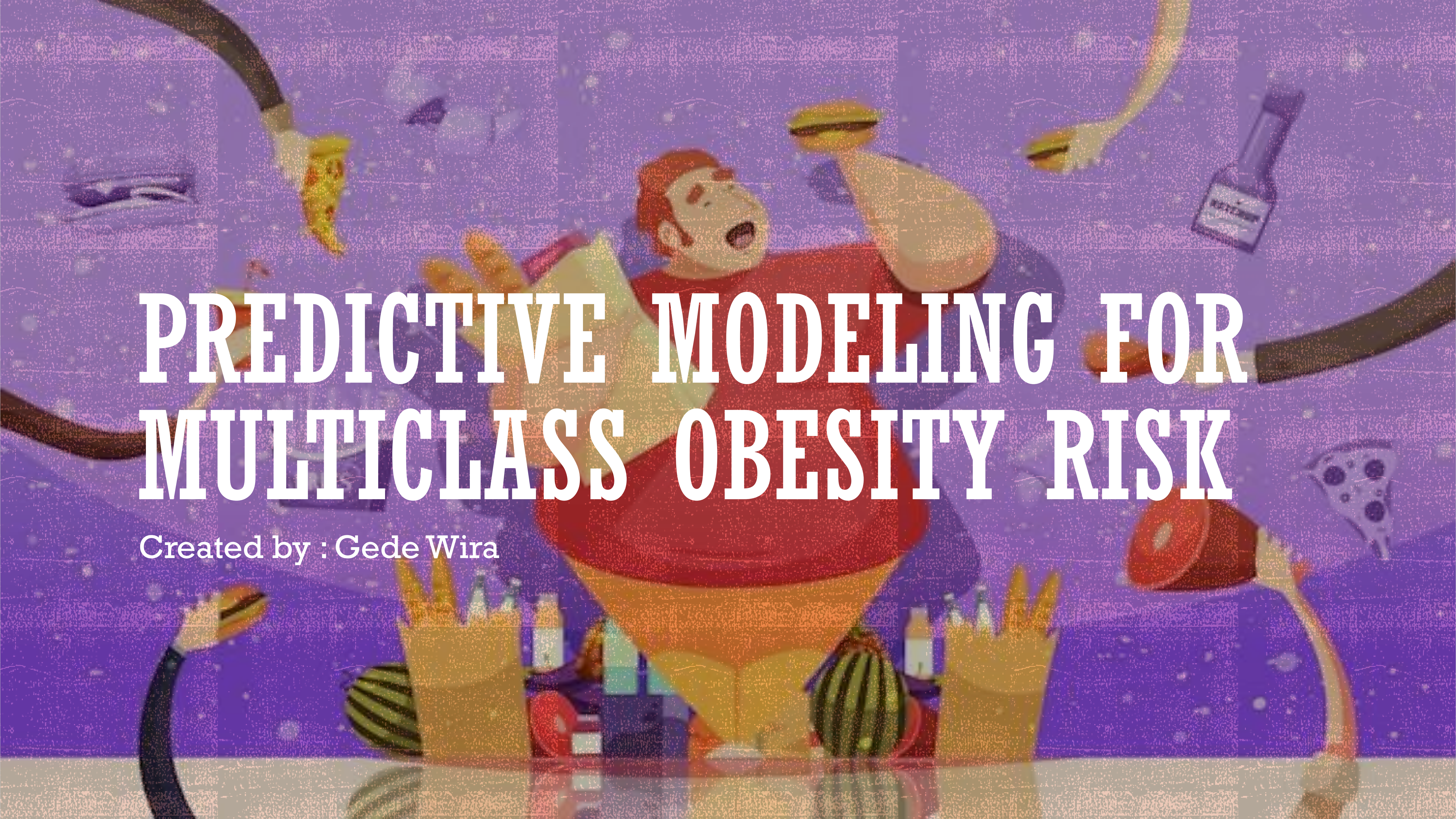


PREDICTIVE MODELING FOR MULTICLASS OBESITY RISK

Created by : Gede Wira

PROJECT GOALS

- Develop a predictive model aimed at identifying obesity risk across multiple classes with high accuracy.
- Ensure the model achieves a high level of accuracy in identifying obesity risk on a multi-class level.

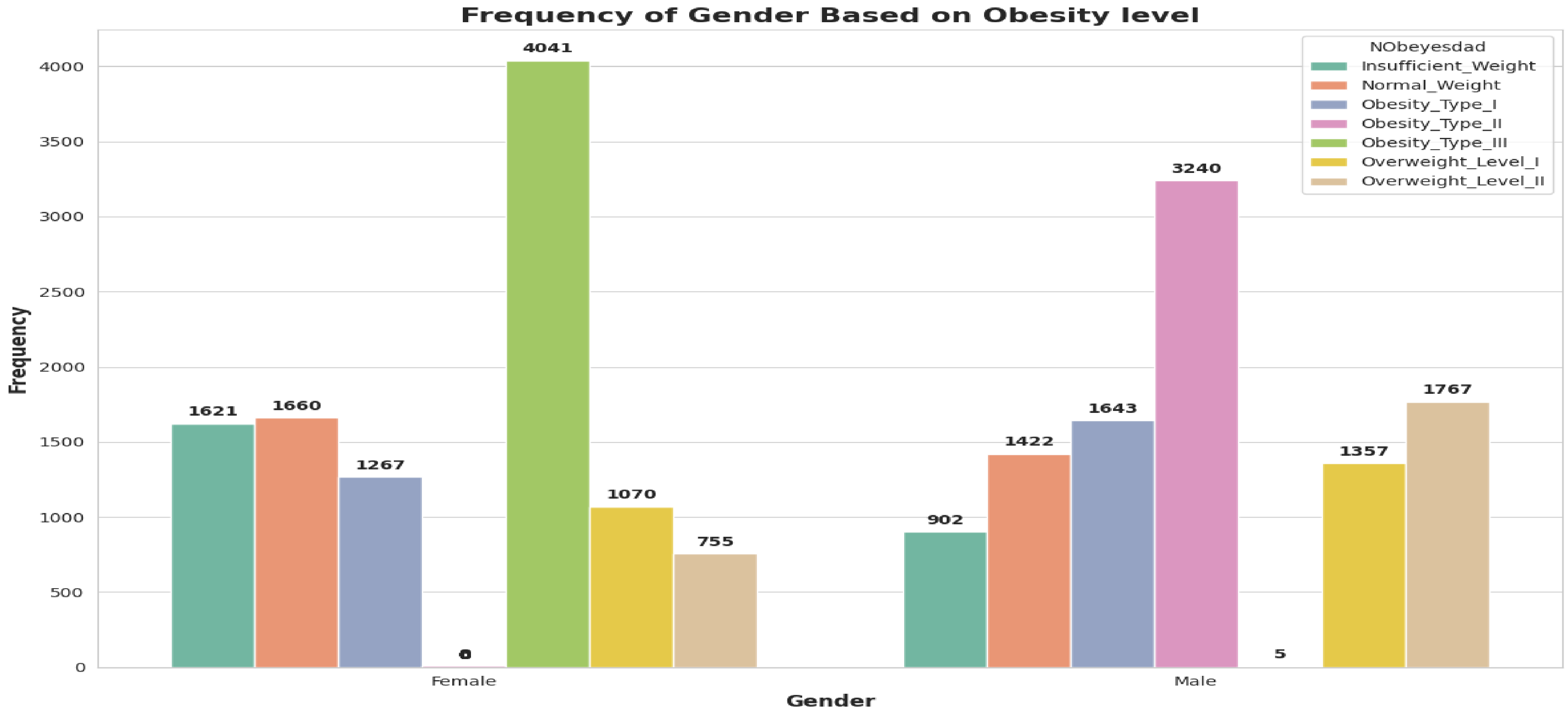
Data Information

- Dataset Source from Kaggle : Multi-Class Prediction of Obesity Risk
- The Training Dataset consists of 18 columns and 20,758 rows of data.
- There are no null values or duplicate data in the dataset.

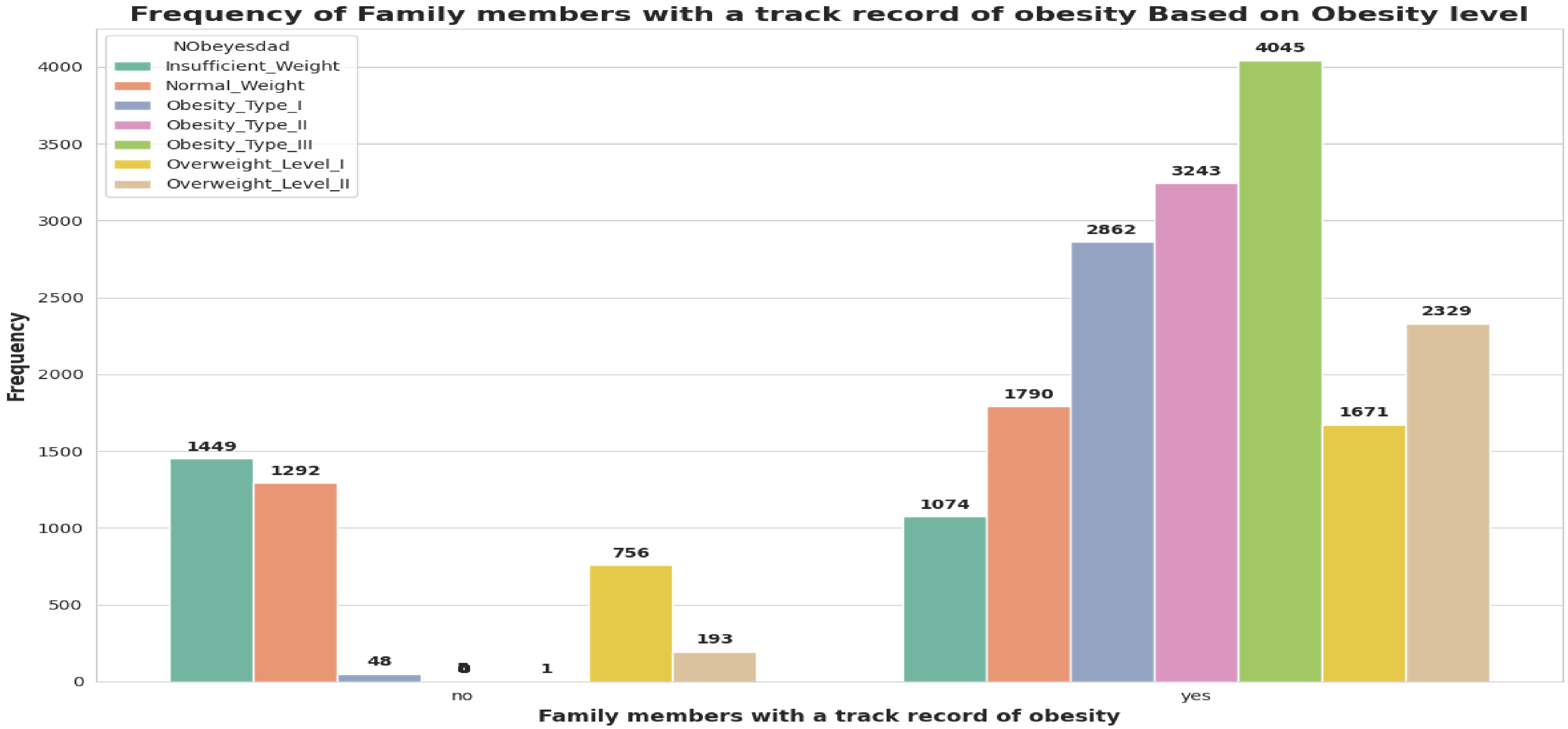
1. EXPLORATORY DATA ANALYSIST

Exploratory Data Analysis, is the preliminary process of analyzing data to understand its patterns, relationships, and characteristics before proceeding with further analysis.

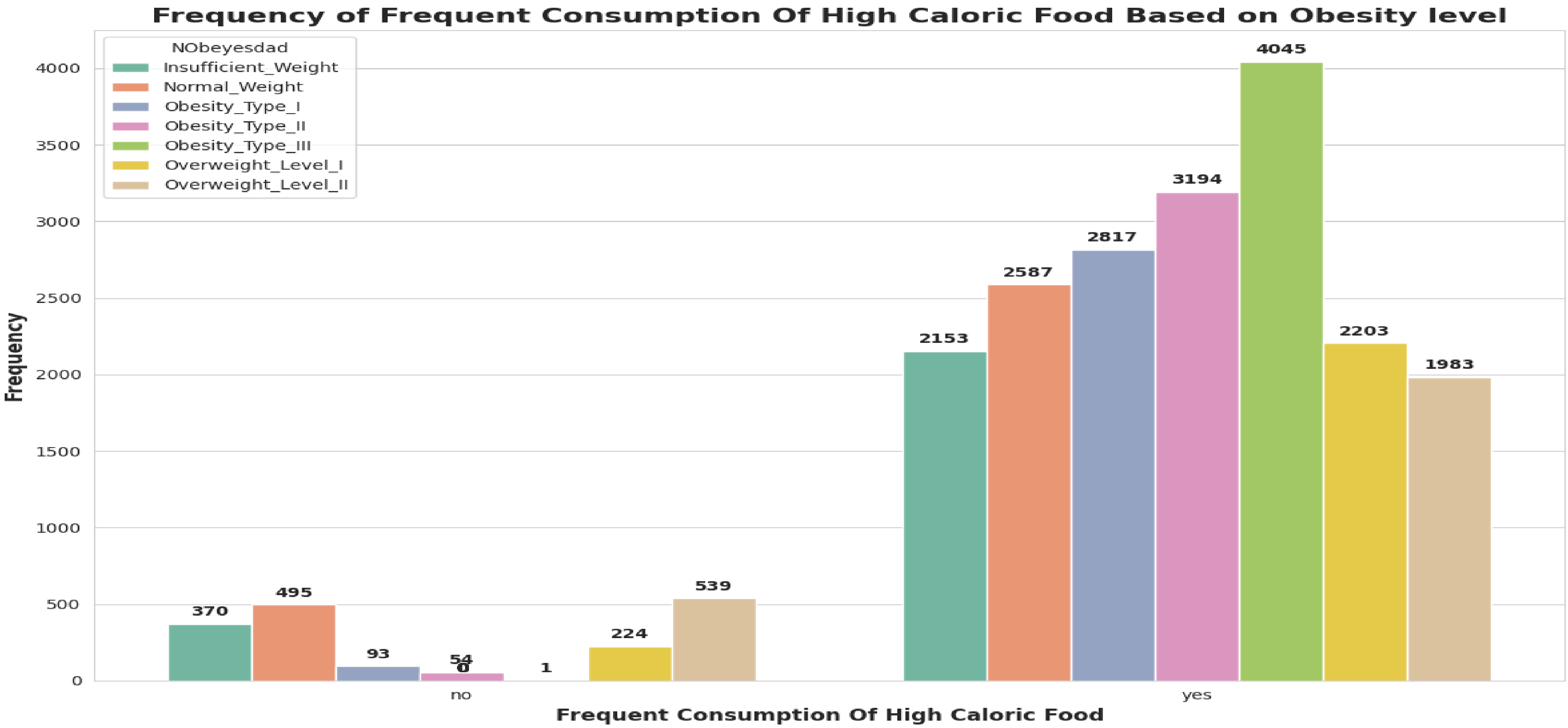
FEMALES HAVE A HIGHER RISK OF EXPERIENCING TYPE III OBESITY, WHILE MALES HAVE A HIGHER RISK OF TYPE II OBESITY.



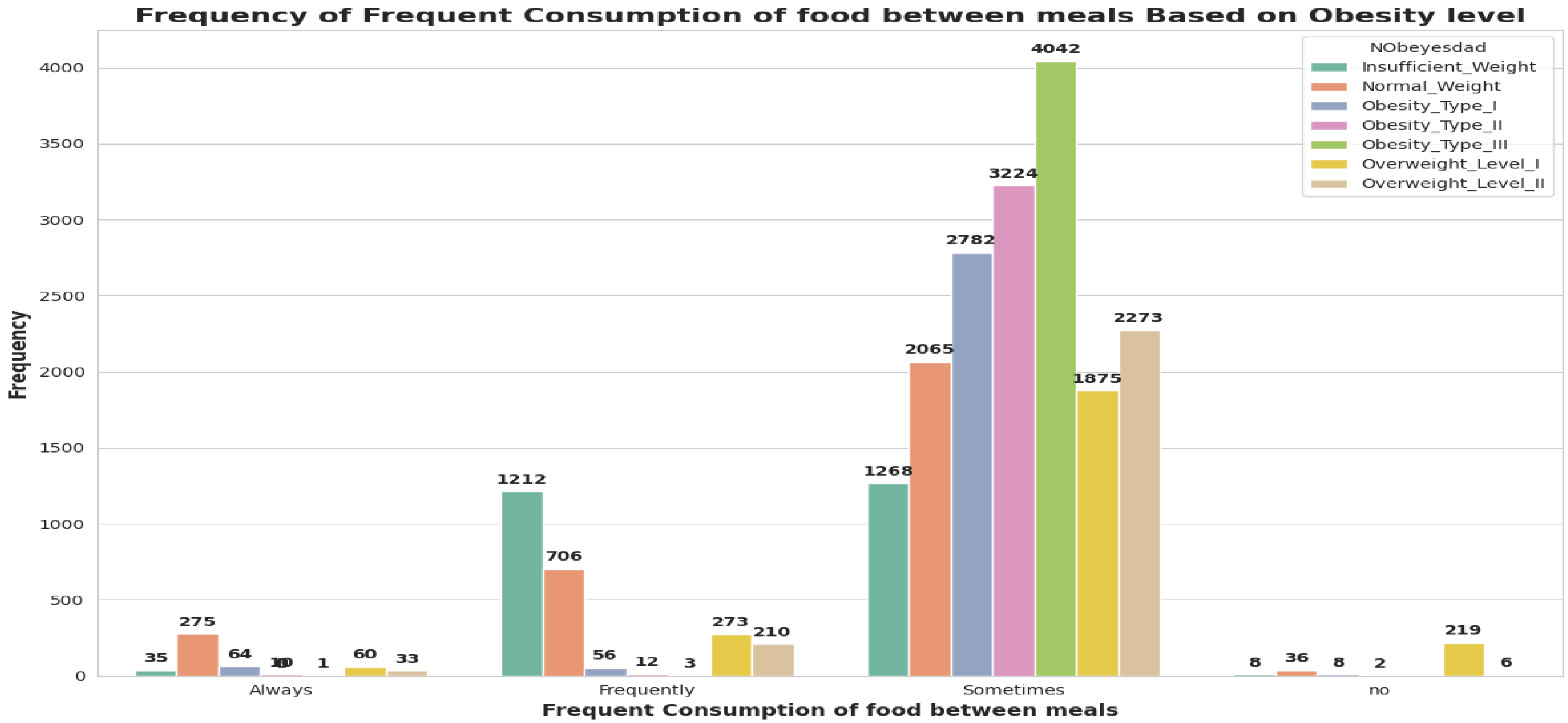
INDIVIDUALS WITH A FAMILY HISTORY OF OVERWEIGHT HAVE A GREATER POTENTIAL TO EXPERIENCE OBESITY.



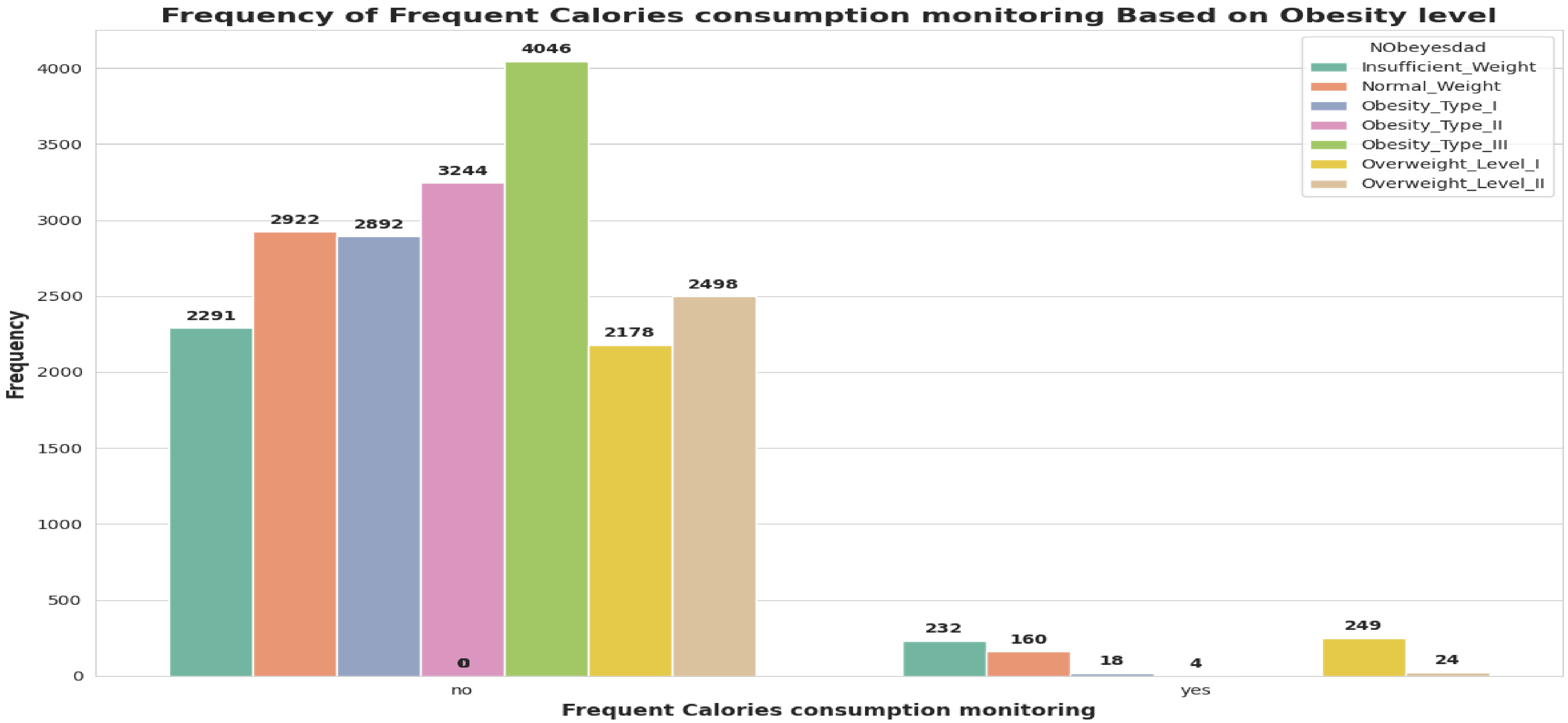
CONSUMING HIGH-CALORIE FOODS INCREASES THE POTENTIAL FOR OBESITY COMPARED TO CONSUMING LOW-CALORIE FOODS.



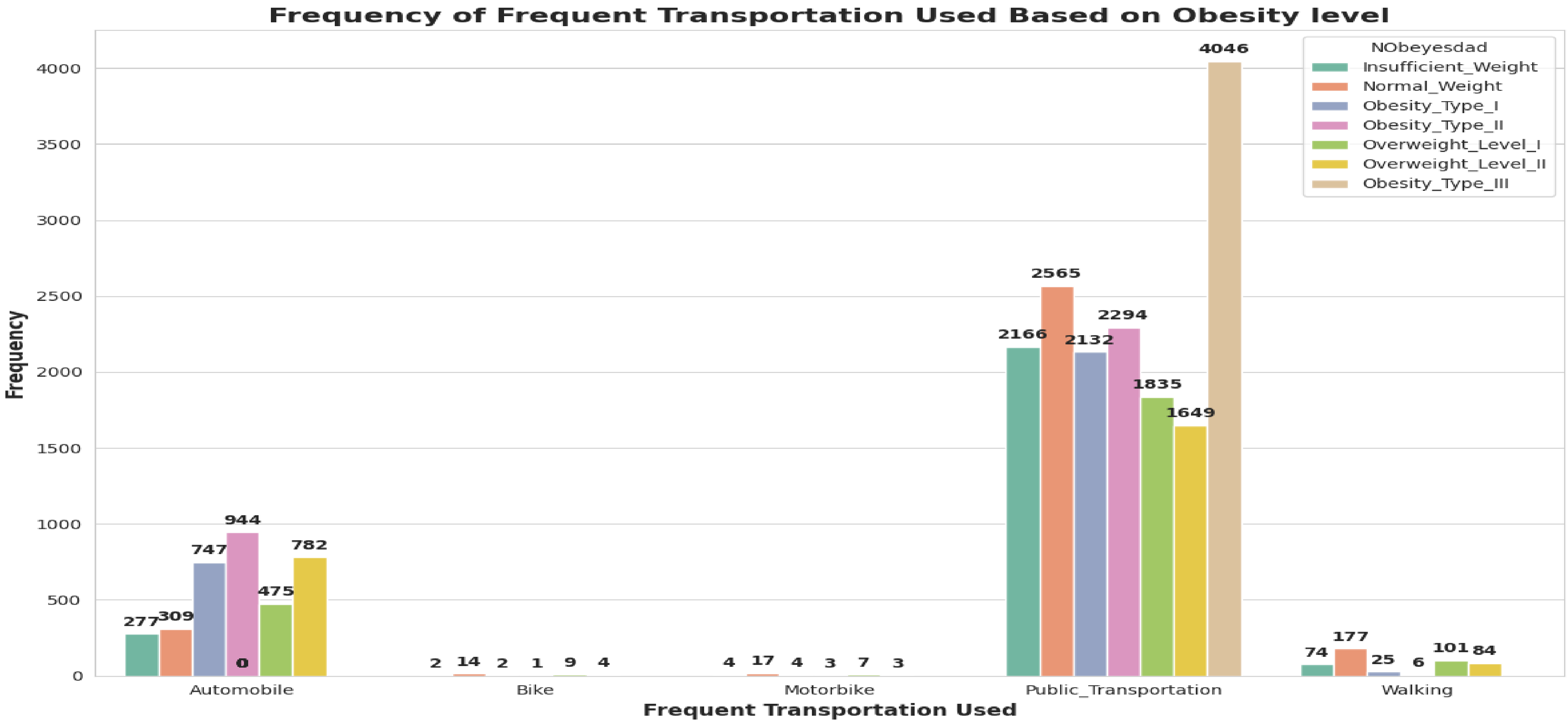
CONSUMING FOOD OUTSIDE MEAL TIMES INCREASES THE RISK OF OBESITY.



NOT MONITORING CALORIE INTAKE INCREASES THE RISK OF OBESITY.



INDIVIDUALS WHO RARELY MOVE (WALK OR BIKE) HAVE A HIGHER RISK OF OBESITY.



INDIVIDUALS WITH OBESITY AND OVERWEIGHT TEND TO ENGAGE LESS IN PHYSICAL ACTIVITY.

```
Average FAF for each 'NObeyesdad':
           NObeyesdad       FAF
0  Insufficient_Weight  1.201782
1        Normal_Weight  1.189580
2       Obesity_Type_I  0.922710
3      Obesity_Type_II  1.029579
4     Obesity_Type_III  0.549225
5   Overweight_Level_I  1.134657
6  Overweight_Level_II  1.060895
```

FAF refers to Physical Activity Frequency.

And NObeyesdad refers to Obesity Level.

THE AVERAGE WATER CONSUMPTION OF INDIVIDUALS WITH OVERWEIGHT AND OBESITY IS HIGHER COMPARED TO THOSE WITH NORMAL WEIGHT.

```
Average CH2O for each 'NObeyesdad':
             NObeyesdad      CH2O
0   Insufficient_Weight  1.744163
1         Normal_Weight  1.806204
2        Obesity_Type_I  2.129783
3       Obesity_Type_II  1.985064
4      Obesity_Type_III  2.332338
5    Overweight_Level_I  2.069366
6   Overweight_Level_II  2.004470
```

CH2O refers to the Consumption of Water Daily

And NObeyesdad refers to Obesity Level.

2. DATA PRE-PROCESING

Data preprocessing in machine learning is the initial process of cleaning and preparing data for modeling. It involves handling missing values, removing duplicates, scaling features, and encoding categorical variables for accurate results.

AS NULL VALUES AND DUPLICATES WERE NOT FOUND, AND THE DATASET HAS BEEN PRE-NORMALIZED, PREPROCESSING ENTAILS FEATURE ENCODING

Binary columns utilize label encoding

While others employ one-hot encoding

For the target column, mapping is conducted.

3. MODELING

Modeling in machine learning refers to the process of developing and adjusting mathematical models or algorithms to learn patterns from data and make predictions or decisions based on these patterns.

THE MODELS EXHIBITING THE HIGHEST ACCURACY AND ROC AUC CURVE ARE GRADIENT BOOSTING AND LIGHTGBM. CONSEQUENTLY, WE WILL PROCEED WITH HYPERPARAMETER TUNING FOR BOTH ALGORITHMS TO ENHANCE THE MODEL'S PERFORMANCE.

Algorithm	Accuracy Train	Accuracy Test	ROC AUC Train	ROC AUC Test
Decision Tree	1.00	0.85	1.00	0.90
Random Forest	1.00	0.89	1.00	0.98
AdaBoost	0.42	0.43	0.75	0.75
Gradient Boosting	0.92	0.90	0.99	0.99
XGBoost	0.99	0.90	1.00	0.99
CatBoost	0.96	0.90	1.00	0.99
LGBM	0.98	0.91	1.00	0.99

GRADIENT BOOSTING METRIC EVALUATION

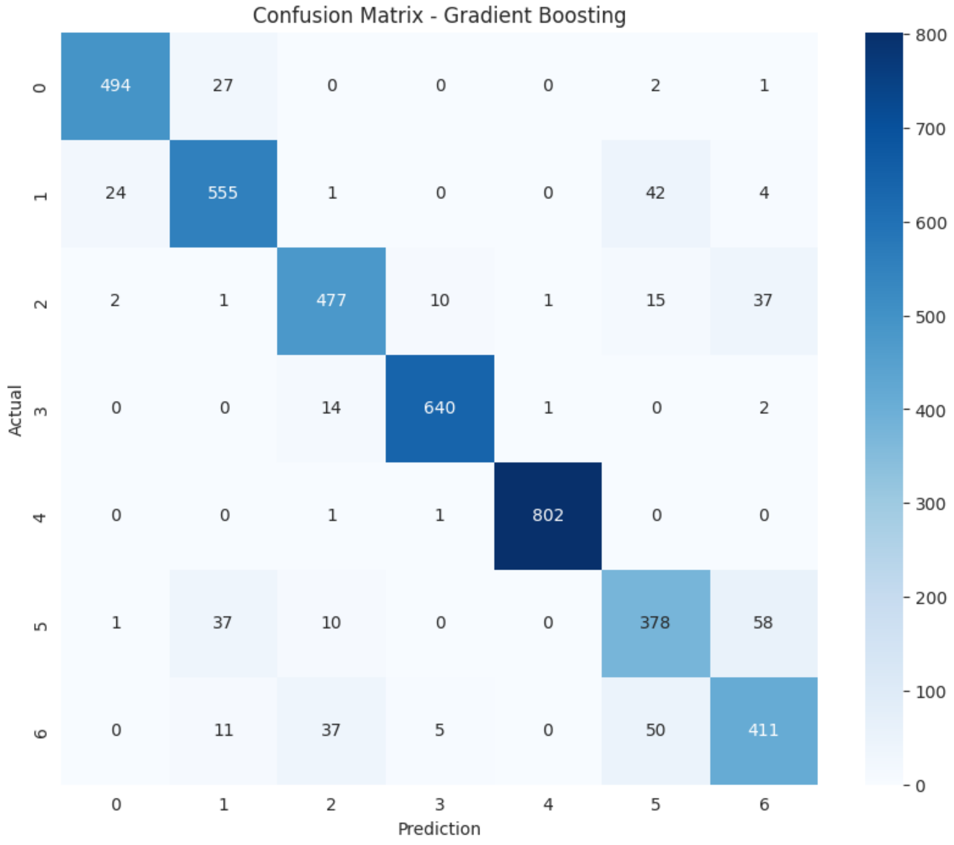
Train Classification Report:

	precision	recall	f1-score	support
0	0.93	0.95	0.94	1999
1	0.89	0.90	0.89	2456
2	0.92	0.92	0.92	2367
3	0.97	0.98	0.98	2591
4	1.00	1.00	1.00	3242
5	0.84	0.80	0.82	1943
6	0.84	0.85	0.84	2008
accuracy			0.92	16606
macro avg	0.91	0.91	0.91	16606
weighted avg	0.92	0.92	0.92	16606

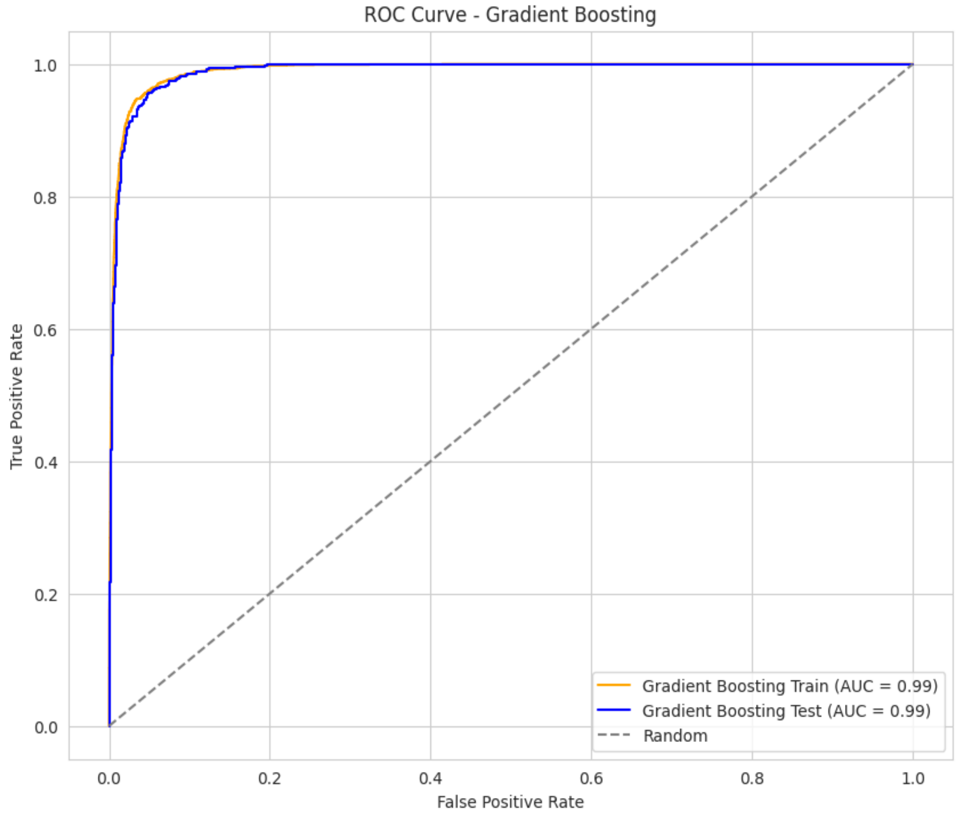
Test Classification Report:

	precision	recall	f1-score	support
0	0.95	0.94	0.95	524
1	0.88	0.89	0.88	626
2	0.88	0.88	0.88	543
3	0.98	0.97	0.97	657
4	1.00	1.00	1.00	804
5	0.78	0.78	0.78	484
6	0.80	0.80	0.80	514
accuracy			0.90	4152
macro avg	0.89	0.89	0.89	4152
weighted avg	0.90	0.90	0.90	4152

Gradient Boosting Confusion Matrix



Gradient Boosting ROC AUC Curve



LGBM METRIC EVALUATION

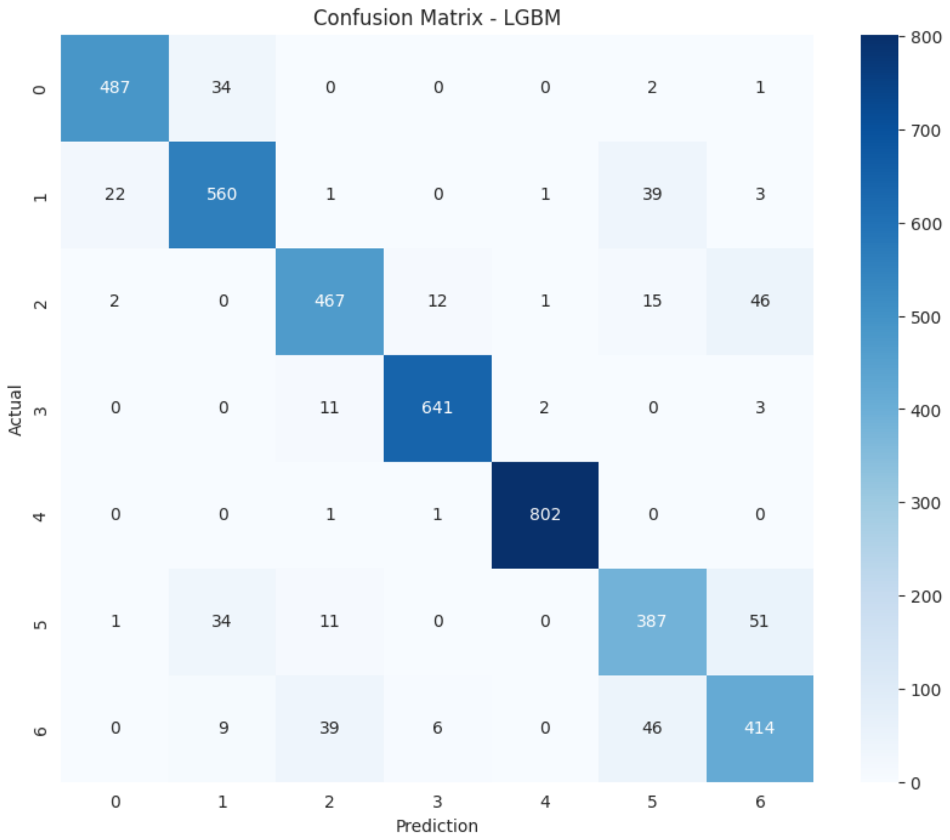
Train Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1999
1	0.97	0.98	0.97	2456
2	0.99	0.99	0.99	2367
3	1.00	1.00	1.00	2591
4	1.00	1.00	1.00	3242
5	0.96	0.93	0.95	1943
6	0.96	0.96	0.96	2008
accuracy			0.98	16606
macro avg	0.98	0.98	0.98	16606
weighted avg	0.98	0.98	0.98	16606

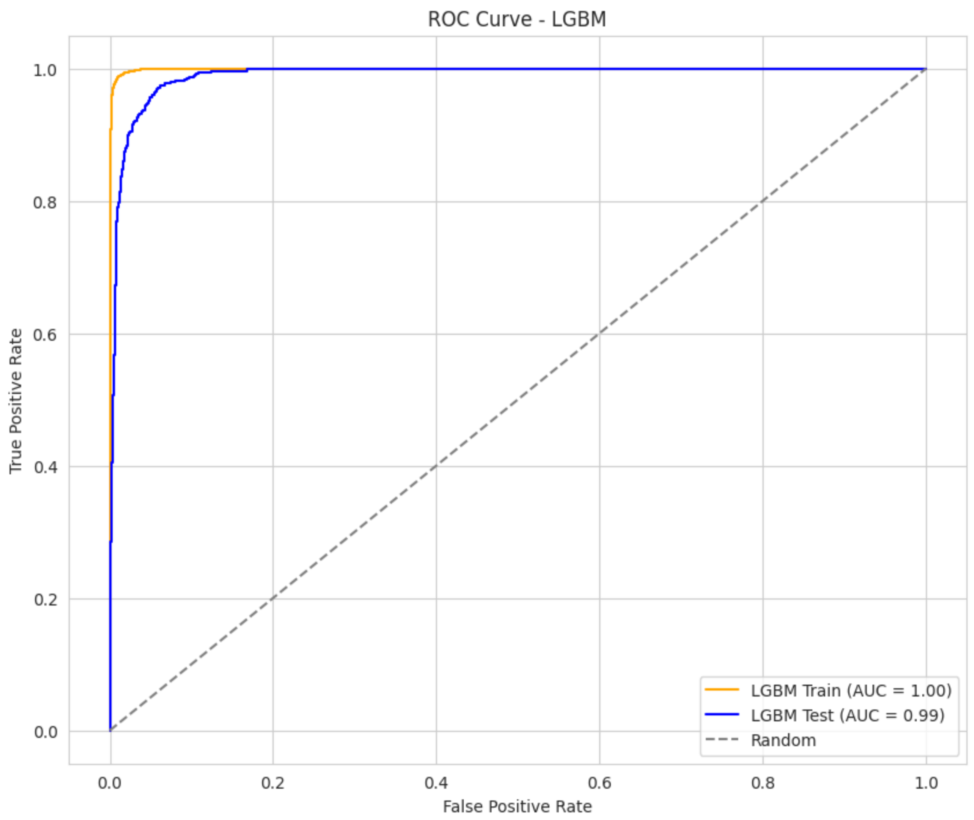
Test Classification Report:

	precision	recall	f1-score	support
0	0.95	0.93	0.94	524
1	0.88	0.89	0.89	626
2	0.88	0.86	0.87	543
3	0.97	0.98	0.97	657
4	1.00	1.00	1.00	804
5	0.79	0.80	0.80	484
6	0.80	0.81	0.80	514
accuracy			0.91	4152
macro avg	0.90	0.89	0.89	4152
weighted avg	0.91	0.91	0.91	4152

LGBM Confusion Matrix



LGBM ROC AUC Curve



AFTER HYPERPARAMETER TUNING, BOTH ALGORITHMS ACHIEVED THE SAME SCORE. HOWEVER, CONSIDERING THAT THE **LGBM ALGORITHM IS MORE TIME-EFFICIENT**, WE HAVE DECIDED TO PROCEED WITH USING **LGBM**.

After Tuning Hyperparameters				
Algorithm	**Accuracy Train**	**Accuracy Test**	**ROC AUC Train**	**ROC AUC Test**
Gradient Boosting	0.92	0.90	0.99	0.99
LGBM	0.92	0.90	0.99	0.99

LGBM METRIC EVALUATION AFTER TUNING

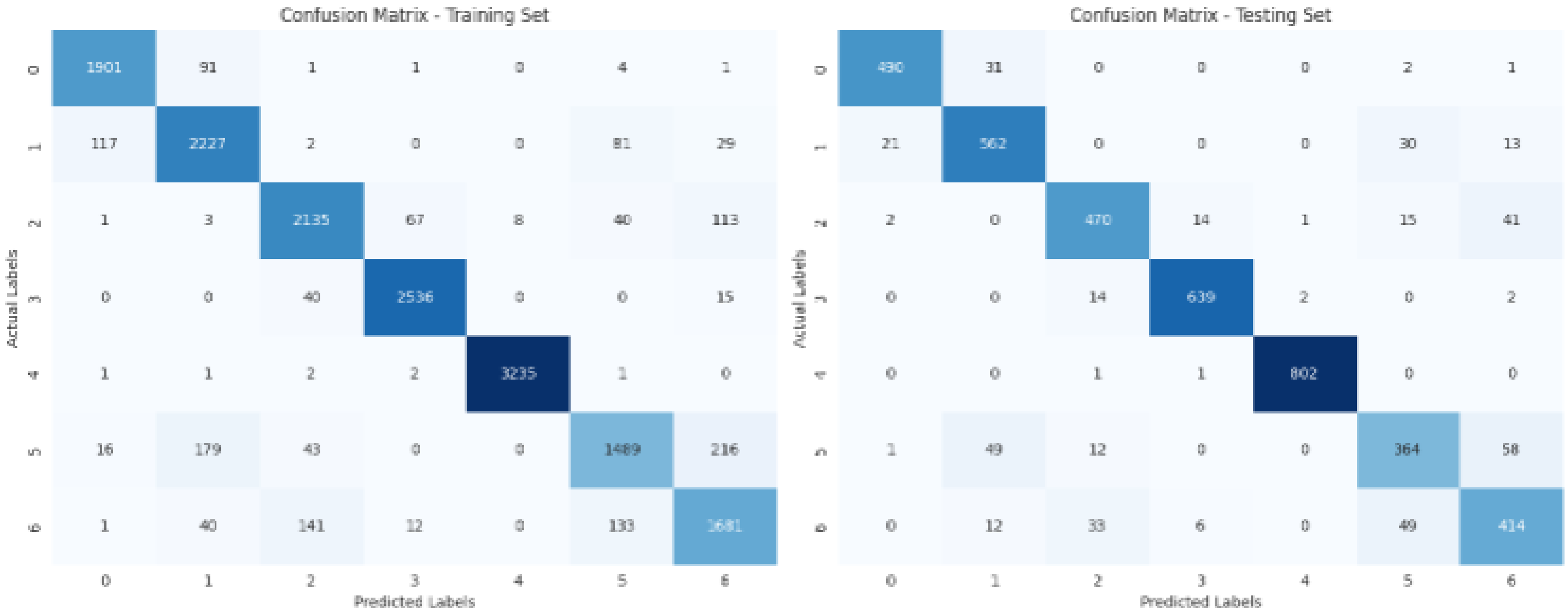
Classification Report - Training Set

	precision	recall	f1-score	support
0	0.93	0.95	0.94	1999
1	0.88	0.91	0.89	2456
2	0.90	0.90	0.90	2367
3	0.97	0.98	0.97	2591
4	1.00	1.00	1.00	3242
5	0.85	0.77	0.81	1943
6	0.82	0.84	0.83	2008
accuracy			0.92	16606
macro avg	0.91	0.91	0.91	16606
weighted avg	0.92	0.92	0.92	16606

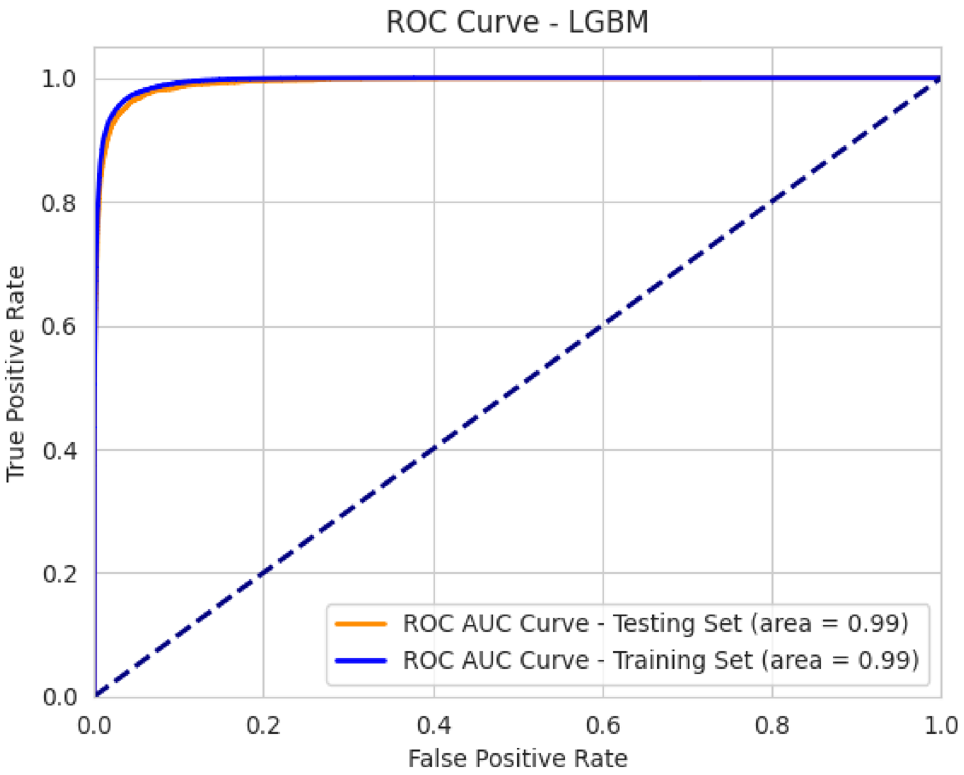
Classification Report - Testing Set

	precision	recall	f1-score	support
0	0.95	0.94	0.94	524
1	0.86	0.90	0.88	626
2	0.89	0.87	0.88	543
3	0.97	0.97	0.97	657
4	1.00	1.00	1.00	804
5	0.79	0.75	0.77	484
6	0.78	0.81	0.79	514
accuracy			0.90	4152
macro avg	0.89	0.89	0.89	4152
weighted avg	0.90	0.90	0.90	4152

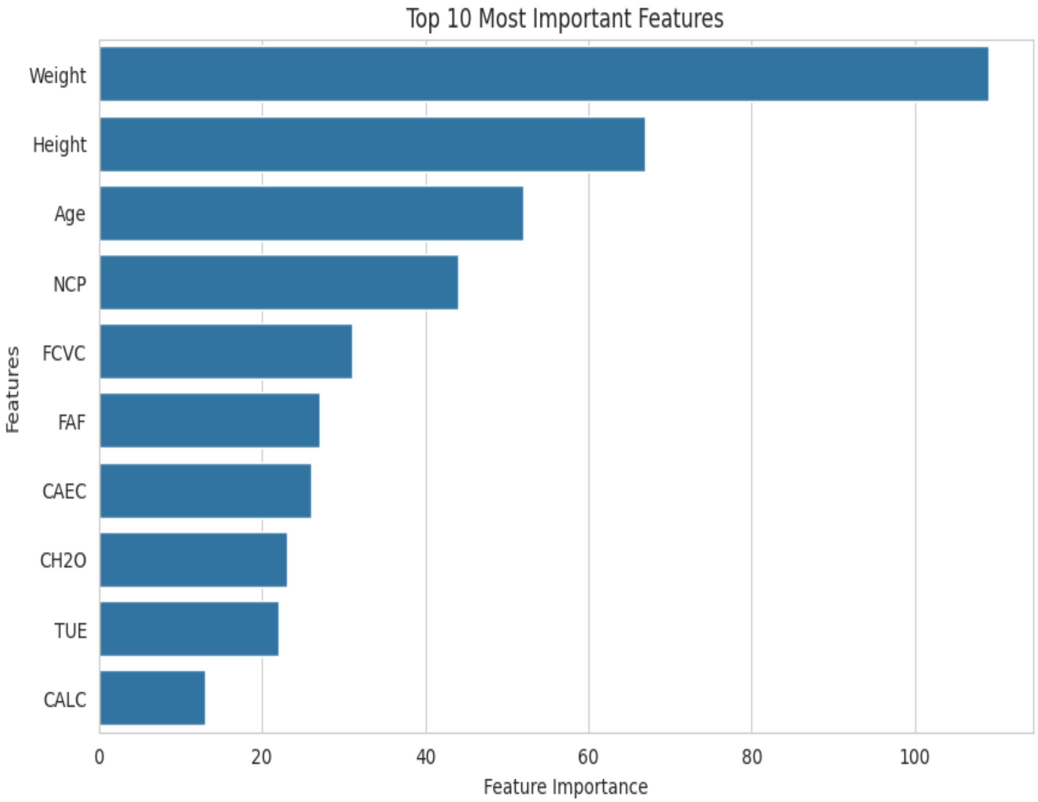
LGBM CONFUSION MATRIX AFTER TUNING



LGBM ROC AUC Curve After Tuning



LGBM Feature Importance



Next, I proceeded to perform predictions on the 'test.csv' data using the pre-tuned LGBM model. The structure of the columns in the 'test.csv' data is identical to that of the 'train.csv' data, except that it lacks the target variable ("Nobeyesdad").

The pre-processing steps used for the 'test.csv' data are the same as those used for the 'train.csv' data.

TO VIEW THE SOURCE CODE AND PREDICTION RESULTS, PLEASE VISIT:

Github: Link

4. CONCLUSION

With a high ROC AUC (0.99) on both testing and training data, as well as relatively stable accuracy (0.92 on training data and 0.90 on testing data), it can be concluded that the model exhibits excellent predictive capability for the target classes in the testing data. A ROC AUC approaching 1 indicates the model's outstanding ability to distinguish between positive and negative classes.

Despite a slight decrease in accuracy from the training to testing data, this difference remains small and can be considered a sign of the model's strong generalization ability on previously unseen data. Therefore, the conclusion drawn is that the model demonstrates exceptional performance in predicting the target classes in the multiclass dataset.


THANK YOU
gedewirawasistha39@gmail.com