# OBESITY CLASSIFICATION

machine learning
models

supervisor

Eng .Abdelrahman Eid

# Team Members

Mariam Osama Shaker

Yasmin kadry El-Sayed

Aya Attia Abd Elhamed

Mennatullah Tarek Arafat

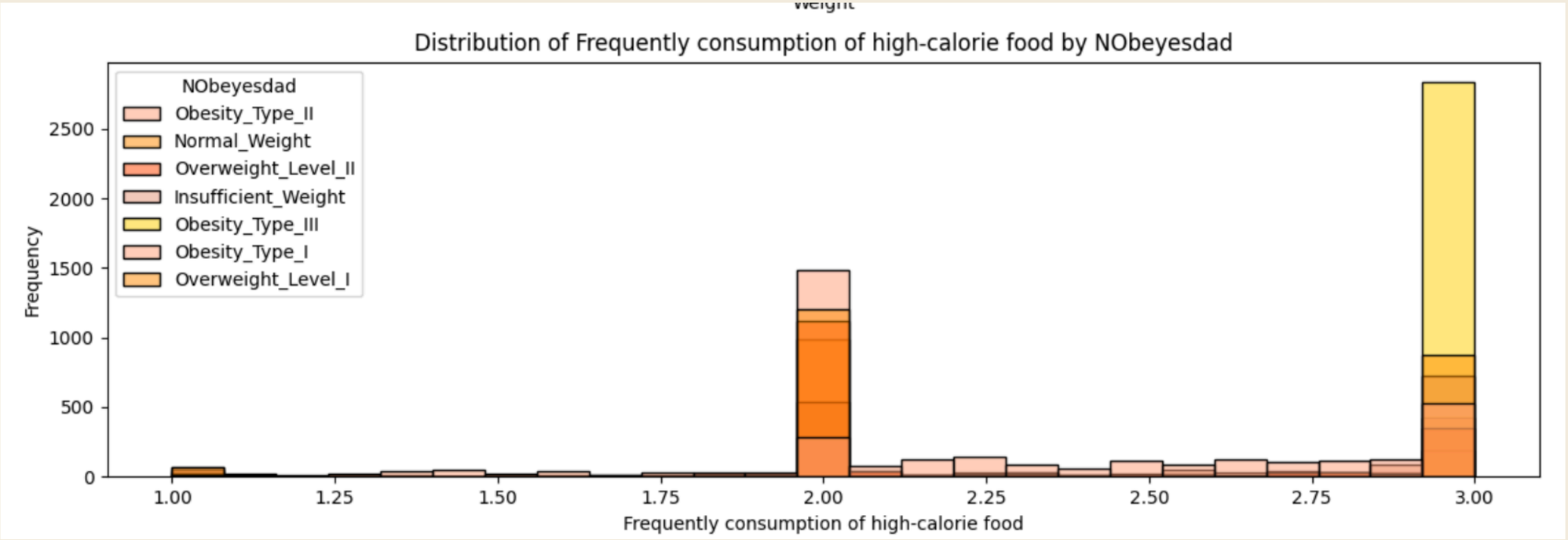
# INTRODUCTION

The problem :
is classification people obesity to avoid health issues

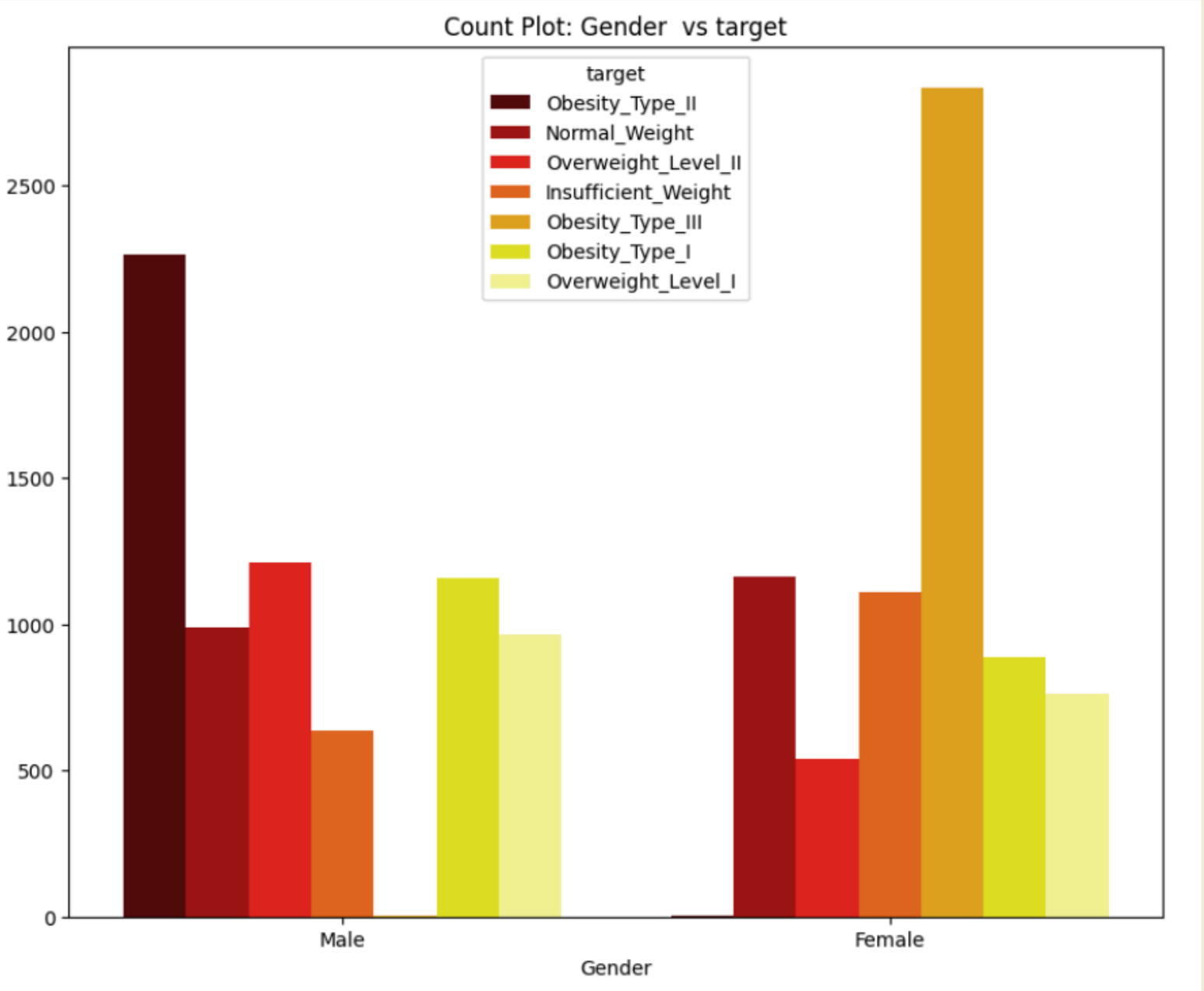
# DATA OVERVIEW

```
<class 'pandas.core.frame.DataFrame'>
Index: 14530 entries, 3969 to 10741
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Gender                          14530 non-null  object 
 1   Age                             14530 non-null  float64
 2   Height                          14530 non-null  float64
 3   Weight                          14530 non-null  float64
 4   family_history_with_overweight  14530 non-null  object 
 5   FAVC                            14530 non-null  object 
 6   FCVC                            14530 non-null  float64
 7   NCP                             14530 non-null  float64
 8   CAEC                            14530 non-null  object 
 9   SMOKE                           14530 non-null  object 
 10  CH2O                            14530 non-null  float64
 11  SCC                             14530 non-null  object 
 12  FAF                             14530 non-null  float64
 13  TUE                             14530 non-null  float64
 14  CALC                            14530 non-null  object 
 15  MTRANS                          14530 non-null  object 
dtypes: float64(8), object(8)
memory usage: 1.9+ MB
```

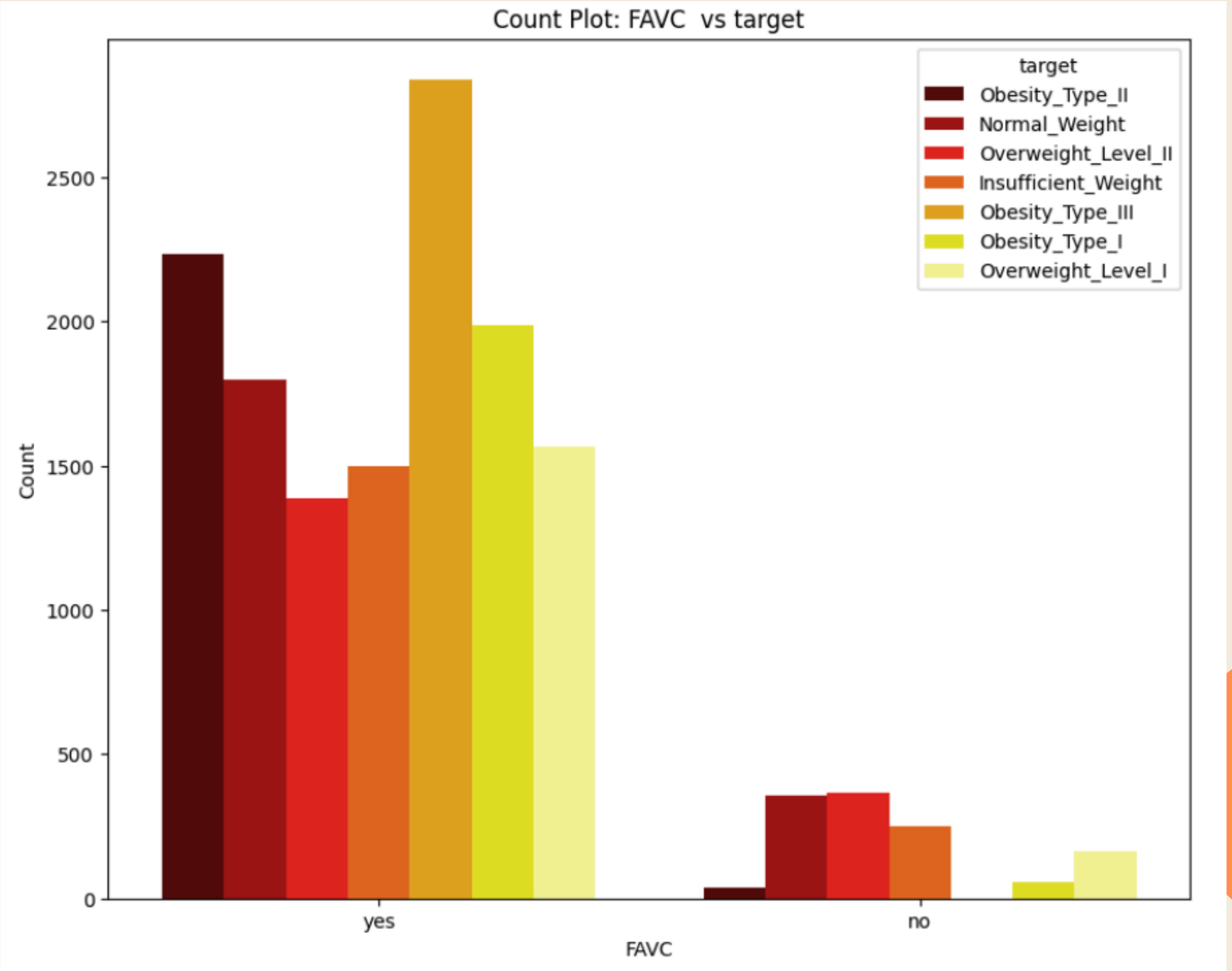
# EDA


Distribution of Frequently consumption of high-calorie food by NObeyesdad
NObeyesdad
Obesity_Type_II
Normal_Weight
Overweight_Level_II
Insufficient_Weight
Obesity_Type_III
Obesity_Type_I
Overweight_Level_I
Frequency
0
500
1000
1500
2000
2500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Frequently consumption of high-calorie food

# EDA


Count Plot: Gender  vs target
target
Obesity_Type_II
Normal_Weight
Overweight_Level_II
Insufficient_Weight
Obesity_Type_III
Obesity_Type_I
Overweight_Level_I
2500
2000
1500
1000
500
0
Male
Female
Gender

# EDA


Count Plot: FAVC  vs target
target
Obesity_Type_II
Normal_Weight
Overweight_Level_II
Insufficient_Weight
Obesity_Type_III
Obesity_Type_I
Overweight_Level_I
Count
2500
2000
1500
1000
500
0
yes
no
FAVC

# DATA PREPROCESSING :

## 1. Encoding Categorical Data

Goal: Convert text categories into numbers for machine learning.

### A. Ordinal Encoding (Order Matters)

Used for categories with a natural order:

CAEC (EATING BETWEEN MEALS):

NO → 0, SOMETIMES → 1, FREQUENTLY → 2, ALWAYS → 3

MTRANS (TRANSPORTATION):

WALKING → 0, BIKE → 1, MOTORBIKE → 2, PUBLIC_TRANSPORTATION → 3, AUTOMOBILE → 4

CALC (ALCOHOL INTAKE):

SAME AS CAEC (NO → 0, ALWAYS → 3)

# DATA PREPROCESSING :

## 1. Encoding Categorical Data

Goal: Convert text categories into numbers for machine learning.

### B. Label Encoding (No Order)

Used for binary/yes-no categories:

GENDER: FEMALE → 0, MALE → 1

FAVC (FAST FOOD FREQUENT): NO → 0, YES → 1

SMOKE: NO → 0, YES → 1

# DATA PREPROCESSING :

## 1. Encoding Data

Goal : to make all values in the same range

```
SCALER = ROBUSTSCALER()
```

due to the outliers

# MODELS(SOLUTION)

by using “XGBClassifier” with kfold

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.91 | 0.92 | 0.92 | 395 |
| 1.0 | 0.87 | 0.88 | 0.87 | 468 |
| 2.0 | 0.80 | 0.81 | 0.80 | 336 |
| 3.0 | 0.83 | 0.84 | 0.83 | 391 |
| 4.0 | 0.88 | 0.87 | 0.88 | 430 |
| 5.0 | 0.97 | 0.96 | 0.97 | 493 |
| 6.0 | 1.00 | 1.00 | 1.00 | 601 |
| accuracy | | | 0.91 | 3114 |
| macro avg | 0.90 | 0.90 | 0.90 | 3114 |
| weighted avg | 0.91 | 0.91 | 0.91 | 3114 |

# MODELS(SOLUTION)

by using “XGBClassifier” with kfold

```
Train Accuracy: 0.9514108740536821
Validation Accuracy: 0.9149004495825305
```

```
Test accuarcy:  0.9142581888246628
```

# MODELS(SOLUTION)

RandomForestClassifier

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.93 | 0.92 | 0.92 | 395 |
| 1.0 | 0.87 | 0.89 | 0.88 | 468 |
| 2.0 | 0.79 | 0.77 | 0.78 | 336 |
| 3.0 | 0.80 | 0.83 | 0.82 | 391 |
| 4.0 | 0.91 | 0.87 | 0.89 | 430 |
| 5.0 | 0.97 | 0.97 | 0.97 | 493 |
| 6.0 | 1.00 | 1.00 | 1.00 | 601 |
| accuracy | | | 0.90 | 3114 |
| macro avg | 0.89 | 0.89 | 0.89 | 3114 |
| weighted avg | 0.91 | 0.90 | 0.90 | 3114 |

# MODELS(SOLUTION)

## 1. Voting Classifier (Accuracy: 81%)

- **Approach: Combined multiple models (e.g., logistic regression, SVM, random forest) for majority voting.**
- **Insight: Ensemble diversity helped, but accuracy was limited.**

## 2. Bagging (Accuracy: 90%)

- **Model: BaggingClassifier with decision trees.**
- **Why? Reduces overfitting by averaging multiple tree predictions (bootstrap samples).**
- **Result: 10% boost over Voting—shows trees work well for this data!**

# MODELS(SOLUTION)

## 3. XGBoost + GridSearchCV (Accuracy: 90%)

- **Optimization: Used GridSearchCV to tune hyperparameters (e.g., max_depth, learning_rate).**
- **Why XGBoost? Handles imbalances and complex relationships better than bagging.**
- **Result: Matched bagging's accuracy but with less variance (more reliable).**

## 4. (Extra Experiment)

- **LogisticRegression**
- **Voting with LogisticRegression - SGDClassifier - DecisionTreeClassifier**

# CONCLUSION

To sum up, this project aimed to classify individuals based on their obesity risk to help prevent health problems. The model showed good performance, and with more data or refined techniques, it could become even more reliable. Ultimately, this work highlights how machine learning can contribute to tackling obesity and improving health outcomes.

# THANK YOU!