

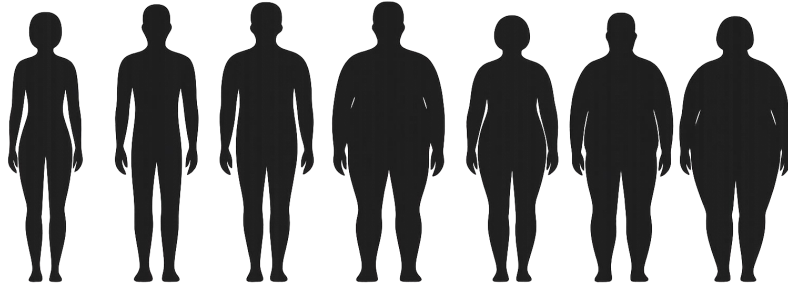
Machine Learning Final Project

Obesity Levels Based On Eating Habits and Physical Condition



Why I Chose This Project/Dataset?

- Obesity is a major global health issue, making prediction and prevention highly relevant.
- The dataset combines demographic and lifestyle factors, offering clear insights into obesity levels.
- The dataset is structured, accessible, and suitable for both statistical analysis and machine learning.



Dataset Upload and Overview

```
1 # =====
2 # Load Dataset
3 # =====
4 # Load the dataset from CSV file
5 df = pd.read_csv('ObesityDataset_raw_and_data_sinthetic.csv')
6
7 # Check dataset dimensions (rows, columns)
8 print("Dataset shape:", df.shape)
9
10 # Display summary information: column types, non-null counts
11 df.info()
12
13 # Generate summary statistics for numerical columns
14 df.describe()
15
16 # Check for missing values in each column
17 missing = df.isnull().sum()
18 print("Missing values per column:\n", missing[missing > 0])
```

```
Dataset shape: (2111, 17)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
```

#	Column	Non-Null Count	Dtype
0	Gender	2111 non-null	object
1	Age	2111 non-null	float64
2	Height	2111 non-null	float64
3	Weight	2111 non-null	float64
4	family_history_with_overweight	2111 non-null	object
5	FAVC	2111 non-null	object
6	FCVC	2111 non-null	float64
7	NCP	2111 non-null	float64
8	CAEC	2111 non-null	object
9	SMOKE	2111 non-null	object
10	CH2O	2111 non-null	float64
11	SCC	2111 non-null	object
12	FAF	2111 non-null	float64
13	TUE	2111 non-null	float64
14	CALC	2111 non-null	object
15	MTRANS	2111 non-null	object
16	NObeyesdad	2111 non-null	object

```
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
Missing values per column:
Series([], dtype: int64)
```

Variables Description



family_history_with_overweight: Has a family member suffered or suffers from overweight?

FAVC: Do you eat high caloric food frequently?

FCVC: Do you usually eat vegetables in your meals?

NCP: How many main meals do you have daily?

CAEC: Do you eat any food between meals?

SMOKE: Do you smoke?

CH2O: How much water do you drink daily?

SCC: Do you monitor the calories you eat daily?

FAF: How often do you have physical activity?

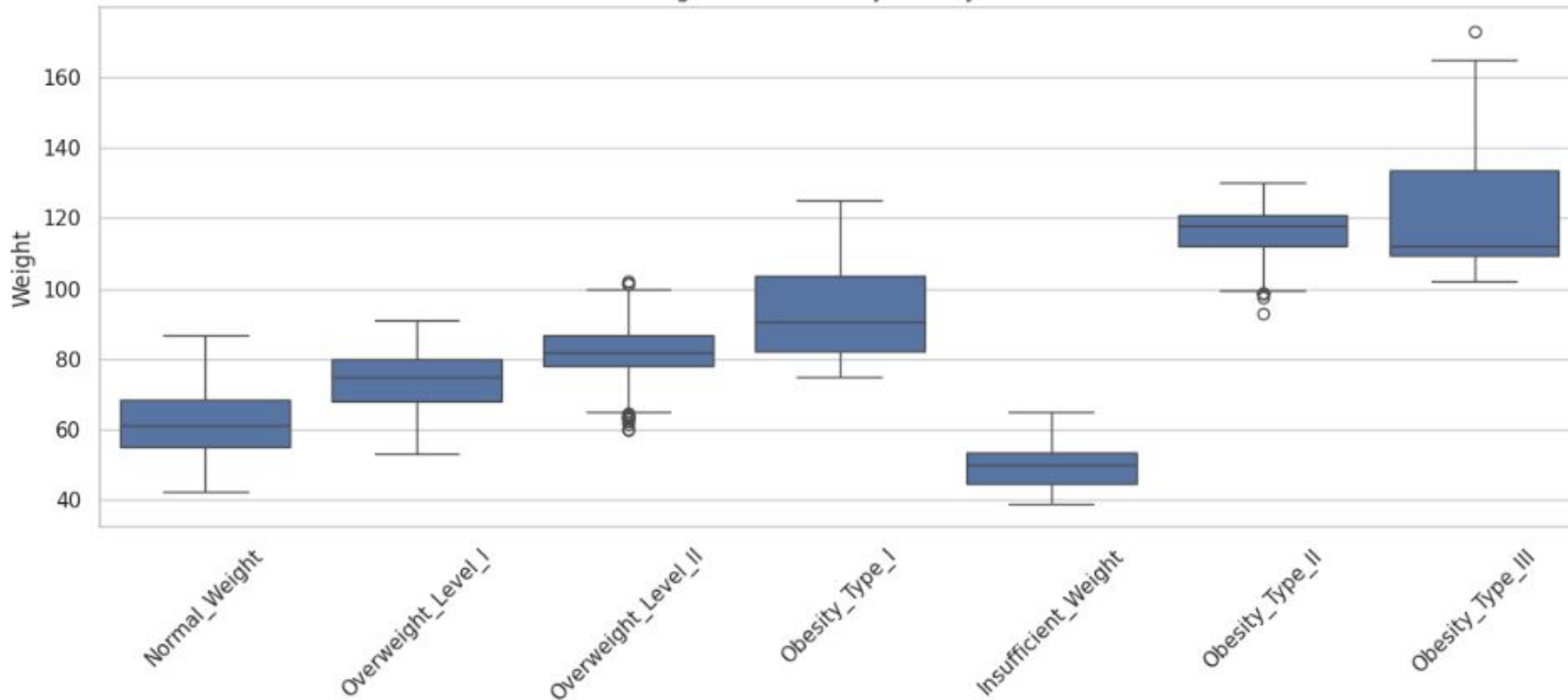
TUE: How much time do you use technological devices such as cell phone, videogames, television, computer and others?

CALC: How often do you drink alcohol?

MTRANS: Which transportation do you usually use?

Exploratory Data Analysis

Weight Distribution by Obesity Level

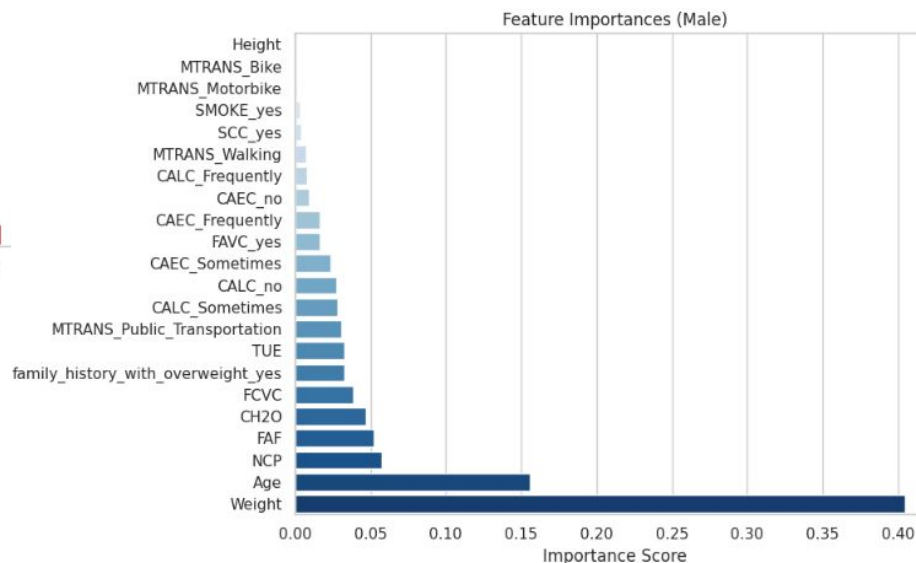
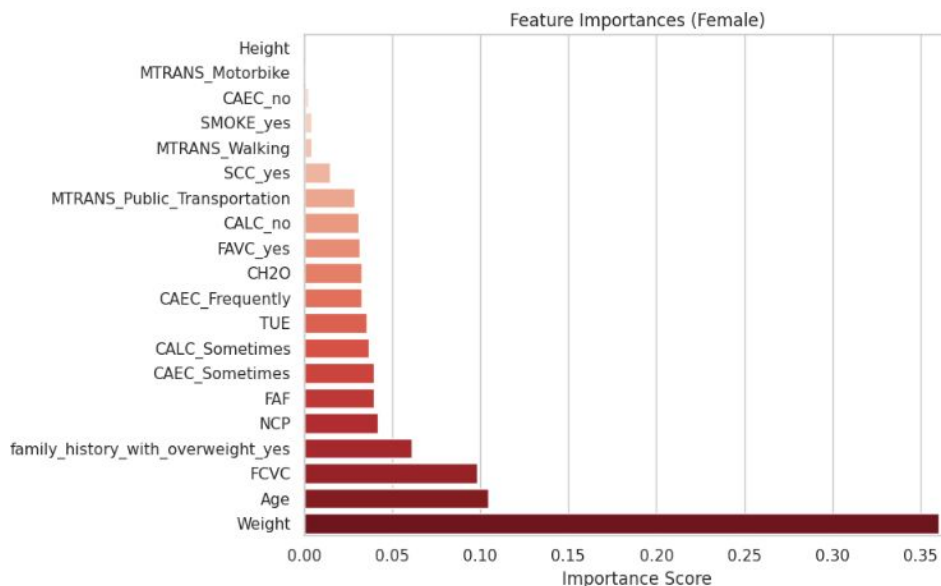


Statistical Test (ANOVA)

```
1 # =====
2 # Statistical Test (ANOVA)
3 # =====
4 # ANOVA checks if mean values of Weight differ significantly across obesity categories
5
6 groups = [df[df['NObeyesdad'] == label]['Weight'] for label in df['NObeyesdad'].unique()]
7 f_stat, p_val = f_oneway(*groups)
8
9 print(f"ANOVA for Weight across Obesity Levels: F={f_stat:.2f}, p={p_val:.4f}")
10
11 # Interpretation:
12 # - F-statistic: higher values indicate stronger differences between groups
13 # - p-value: probability that differences are random
14 #   If  $p < 0.05$  → statistically significant differences
```

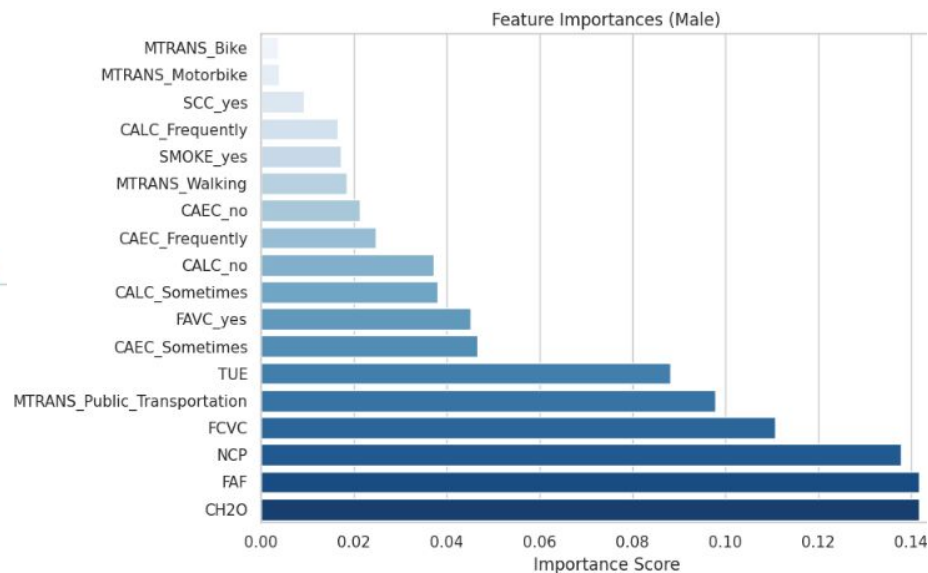
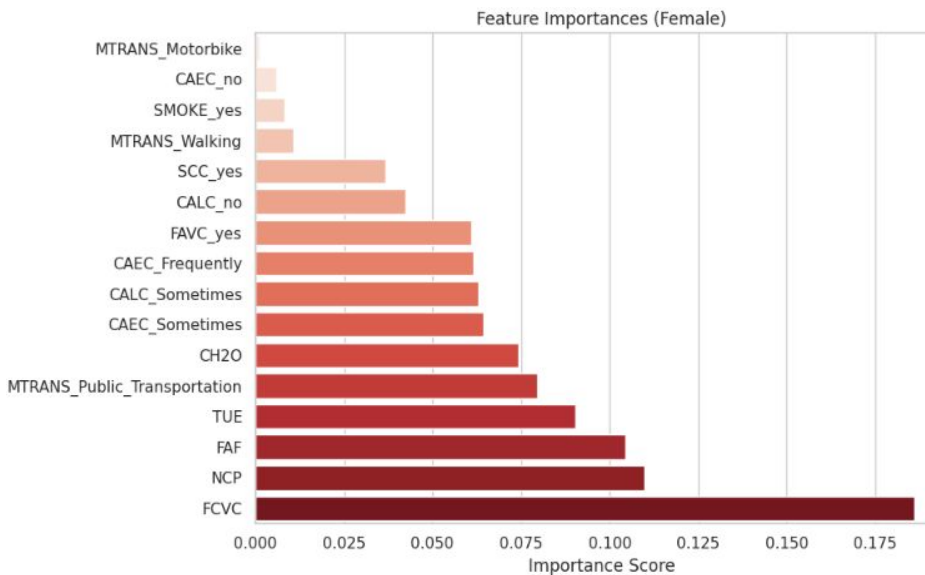
ANOVA for Weight across Obesity Levels: F=1966.52, p=0.0000

Feature Importance (Random Forest)



Feature Importance (Random Forest)

Without Weight, Height, Age, and family_history_with_overweight



Number of Records in Each Category

Women

	count
NObeyesdad	
Obesity_Type_III	323
Insufficient_Weight	173
Obesity_Type_I	156
Overweight_Level_I	145
Normal_Weight	141
Overweight_Level_II	103
Obesity_Type_II	2

Gender

Male	1068
Female	1043

Men

	count
NObeyesdad	
Obesity_Type_II	295
Obesity_Type_I	195
Overweight_Level_II	187
Normal_Weight	146
Overweight_Level_I	145
Insufficient_Weight	99
Obesity_Type_III	1

Obesity Categories Grouped

Women

Female distribution:

Obesity_Grouped

Obesity	481
---------	-----

Overweight	248
------------	-----

Insufficient_Weight	173
---------------------	-----

Normal_Weight	141
---------------	-----

Name: count, dtype: int64

Men

Male distribution:

Obesity_Grouped

Obesity	491
---------	-----

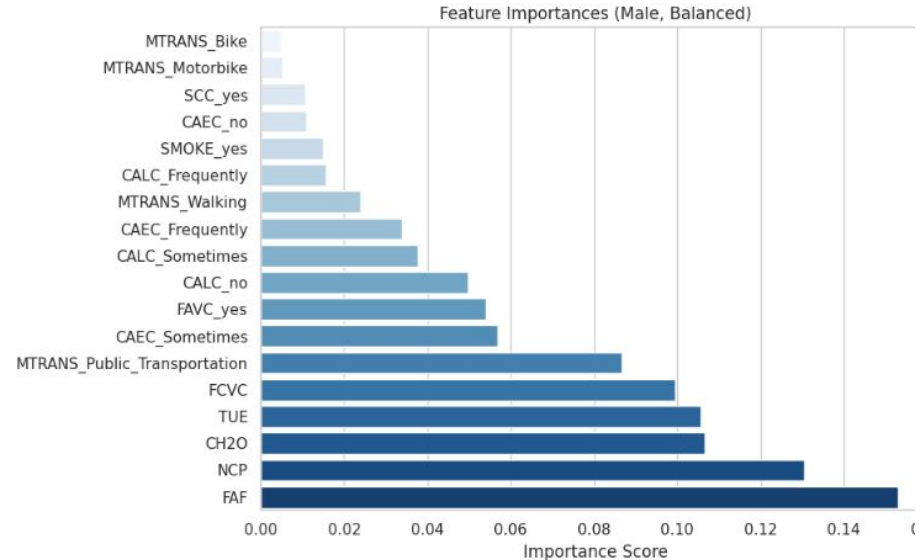
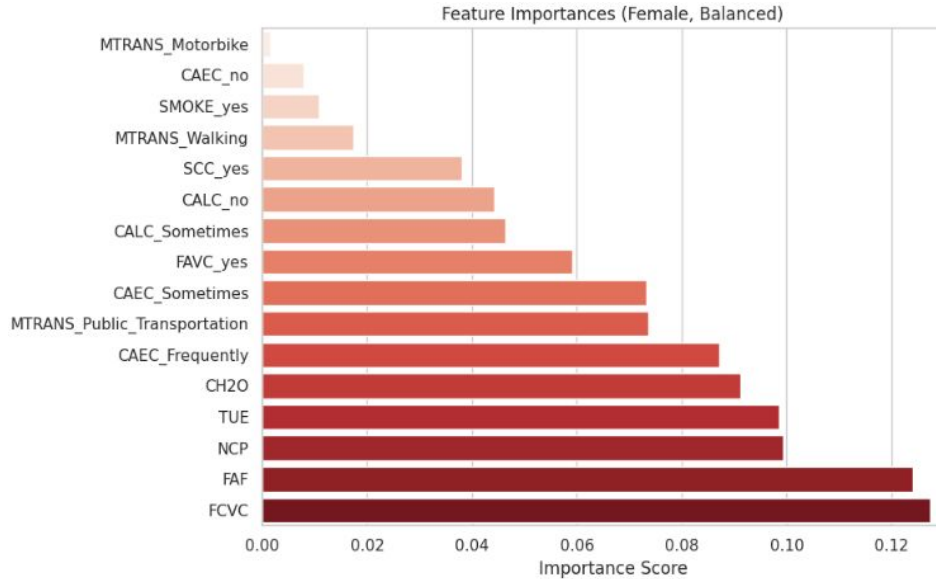
Overweight	332
------------	-----

Normal_Weight	146
---------------	-----

Insufficient_Weight	99
---------------------	----

Name: count, dtype: int64

Random Forest With Class Balancing



Multiclass Logistic Regression



Classification Report (Female):

	precision	recall	f1-score	support
Insufficient_Weight	0.76	0.69	0.73	173
Normal_Weight	0.48	0.49	0.48	141
Obesity	0.85	0.77	0.81	481
Overweight	0.55	0.69	0.61	248
accuracy			0.70	1043
macro avg	0.66	0.66	0.66	1043
weighted avg	0.72	0.70	0.70	1043

Classification Report (Male):

	precision	recall	f1-score	support
Insufficient_Weight	0.32	0.65	0.43	99
Normal_Weight	0.58	0.63	0.60	146
Obesity	0.65	0.61	0.63	491
Overweight	0.54	0.40	0.46	332
accuracy			0.55	1068
macro avg	0.52	0.57	0.53	1068
weighted avg	0.57	0.55	0.55	1068

F1-Score Visualization



Confusion Matrix

Confusion Matrix (Female)



Confusion Matrix (Male)



Model Performance Overview

The logistic regression model performed better for women than for men.

Women

- **Average f1 score:** 0.66.
- **Best prediction:** Obesity -> 0.81.
- **Worst prediction:**
Normal_Weight -> 0.48.
- **Often confused:** Obesity with Overweight -> 92 cases out of 481.

Men

- **Average f1 score:** 0.53.
- **Best prediction:** Obesity -> 0.63.
- **Worst prediction:**
Insufficient_Weight -> 0.43.
- **Often confused:** Overweight with Obesity -> 133 cases out of 332.

Why Other Models Are Less Suitable

- **Ordinary Linear Regression**: continuous values, not suitable for categories.
- **Multinomial Linear Regression**: continuous outcomes, not probabilities, not suitable for classification.
- **Decision Trees**: prone to overfitting with categorical obesity levels.
- **Random Forests**: hard to explain.
- **Support Vector Machines (SVM)**: less transparent, hard to present.
- **Neural Networks**: require large datasets and resources.

Factor Importance Changes

Feature importance depends on included features.

Women

1. **Weight.**
2. **Age.**
3. **Do you usually eat vegetables in your meals?**

Men

1. **Weight.**
2. **Age.**
3. **How many main meals do you have daily?**

Factor Importance Changes

After the ***Weight, Height, Age***, and ***family_history_with_overweight factors*** were removed, the 3 top features became:

Women

1. Do you usually eat vegetables in your meals?
2. How many main meals do you have daily?
3. How often do you have physical activity?

Men

1. How much water do you drink daily?
2. How often do you have physical activity?
3. How many main meals do you have daily?

Stable Lifestyle Factors

After obesity categories grouping and balancing,
the 3 top features became:

Women

1. **Do you usually eat vegetables in your meals?**
2. **How often do you have physical activity?**
3. **How many main meals do you have daily?**

Men

1. **How often do you have physical activity?**
2. **How many main meals do you have daily?**
3. **How much water do you drink daily?**

Lifestyle Factors Summary



This research shows that among the most important lifestyle factors always remain:

Women

Do you usually eat vegetables in your meals?

Men

How much water do you drink daily?

Everyone

- 1. How often do you have physical activity?**
- 2. How many main meals do you have daily?**

Difficulties

- **Dataset selection**

Many free datasets contained unrealistic or synthetic-looking data, making it difficult to ensure validity.

- **Gender-specific modeling**

Treating gender as a simple encoded variable did not reflect biological and lifestyle differences.

Solution: conducted separate analysis for male and female datasets, since physiological and lifestyle distinctions (hormonal profile, body composition, habits) affect obesity classification.

- **Imbalance in obesity categories**

Some classes had very few samples, which affected model performance.

Solution: grouped obesity types into broader categories and applied class balancing techniques.

Future Improvements

- **Expand dataset size**

Collect more samples (currently ~2000) to improve statistical reliability and generalization.

- **Enhance model performance**

Improve precision, recall, and F1-score by testing other algorithms and applying cross-validation for more robust evaluation

- **External validation**

Test models on independent datasets to check generalizability beyond the current sample.

How Could the Project be Used?



- **Identify key factors influencing obesity**

Determine the most important predictors separately for women and men.

- **Support gender-specific health interventions**

Show that lifestyle factors (diet, physical activity, habits) have different impact levels depending on gender.

- **Support targeted health interventions**

Provide insights for designing gender-specific prevention programs and public health strategies.

- **Improve awareness and education**

Help individuals understand which behaviors are most critical for maintaining healthy weight.

- **Provide framework for future research**

Offer a framework for analyzing obesity with balanced datasets and gender-specific modeling.

Used Resources



1. Dataset: [Estimation of Obesity Levels based on Eating Habits and Physical Condition \(UCI\)](#)
2. [Google Classroom materials](#): presentations and Jupyter notebooks (.ipynb files)
3. Google Search
4. [Stack Overflow](#)
5. [W3Schools Python](#) (documentation and tutorials)
6. AI tool Copilot (proofreading, brainstorming ideas, debugging code, and searching for the dataset)



Thank you!

