ReDI School of
Digital Integration

Machine Learning Final Project

Obesity Levels
Based On Eating Habits and Physical Condition

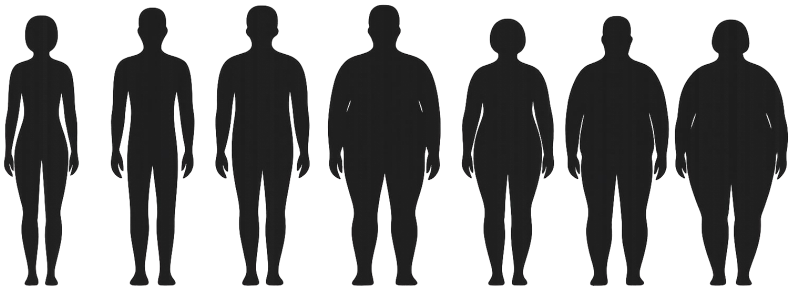
We use tech to connect human potential and opportunity with dignity & humility

Yuliia Fomenko

2025

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

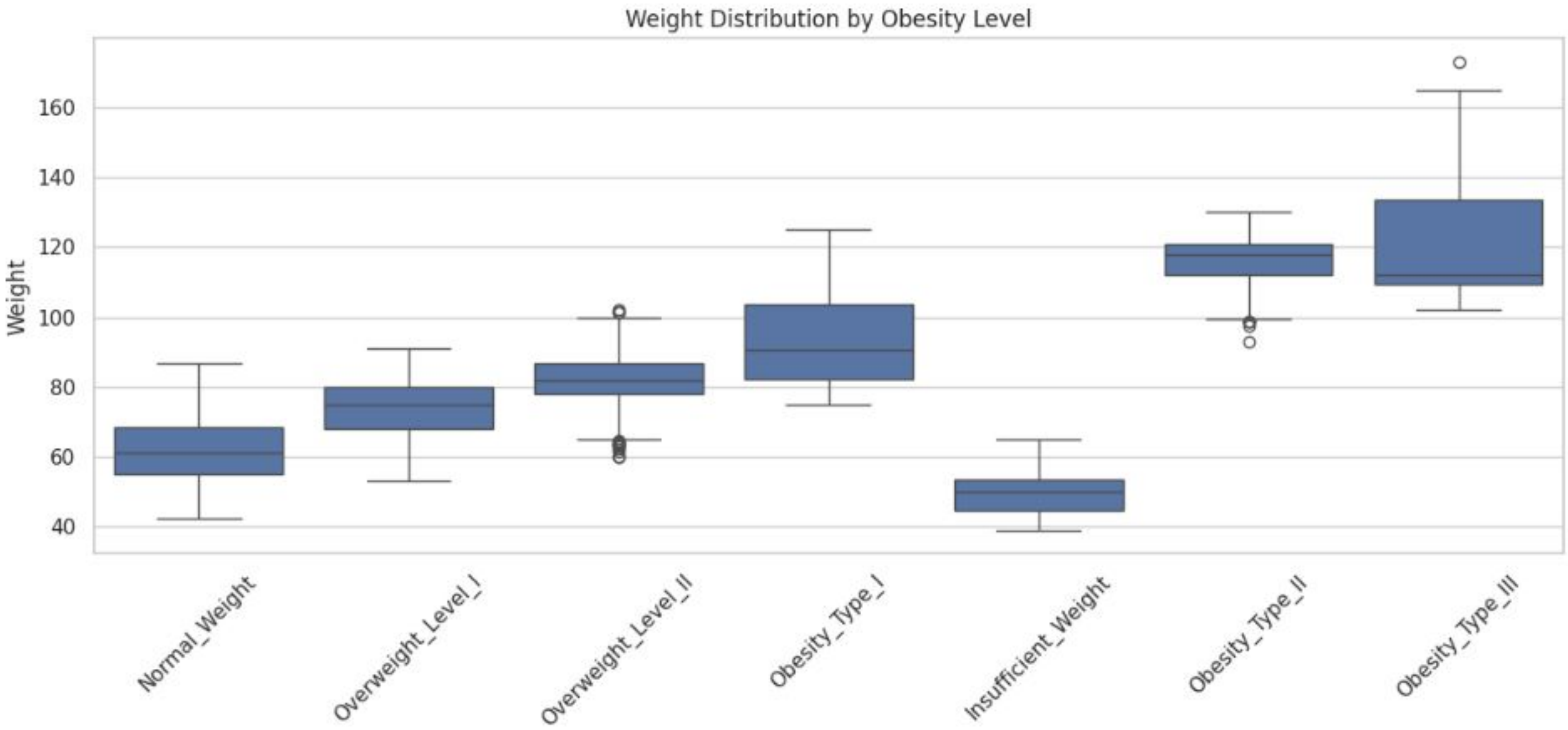
```python
# ==================================
# Load Dataset
# ==================================

# Load the dataset from CSV file
df = pd.read_csv('ObesityDataSet_raw_and_data_sinthetic.csv')

# Check dataset dimensions (rows, columns)
print("Dataset shape:", df.shape)

# Display summary information: column types, non-null counts
df.info()

# Generate summary statistics for numerical columns
df.describe()

# Check for missing values in each column
missing = df.isnull().sum()
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

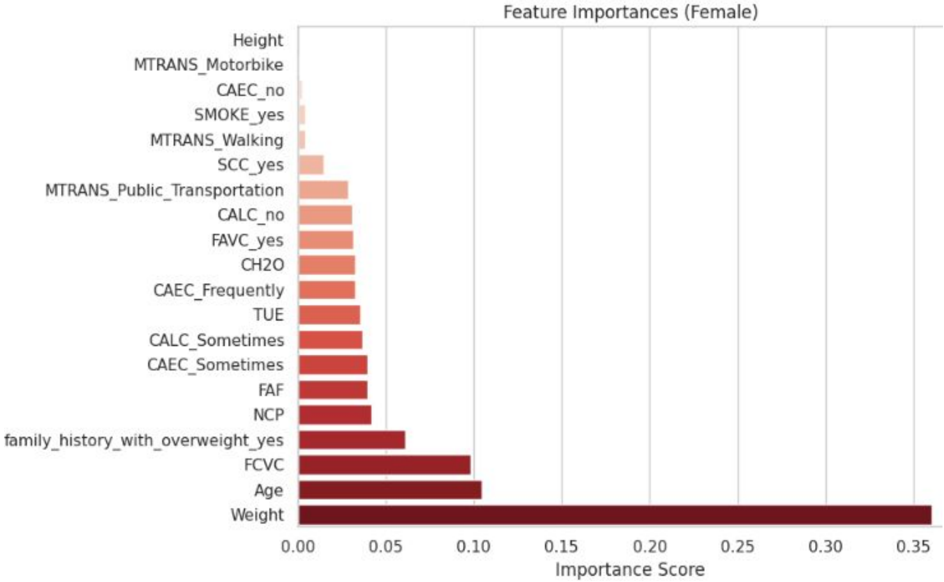


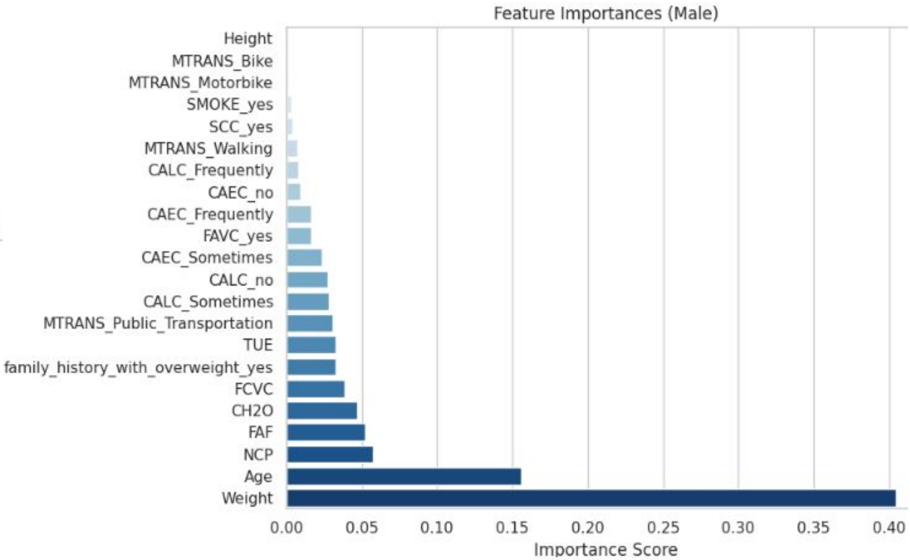
Statistical Test (ANOVA)

```
# ================================
# Statistical Test (ANOVA)
# ================================
# ANOVA checks if mean values of Weight differ significantly across obesity categories

groups = [df[df['NObeyesdad'] == label]['Weight'] for label in df['NObeyesdad'].unique()]
f_stat, p_val = f_oneway(*groups)

print(f"ANOVA for Weight across Obesity Levels: F={f_stat:.2f}, p={p_val:.4f}")

# Interpretation:
# - F-statistic: higher values indicate stronger differences between groups
# - p-value: probability that differences are random
#   If p < 0.05 → statistically significant differences
```

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

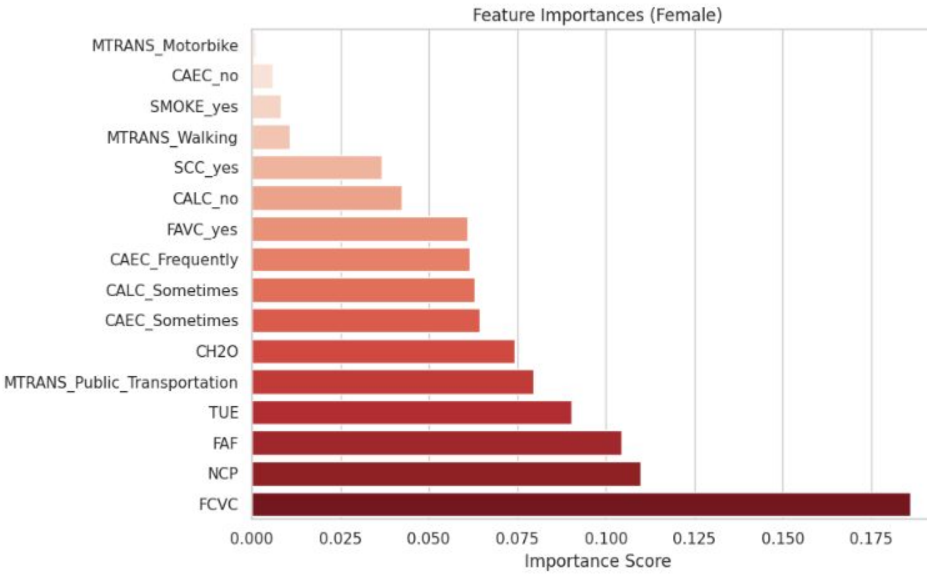
Feature Importance (Random Forest)

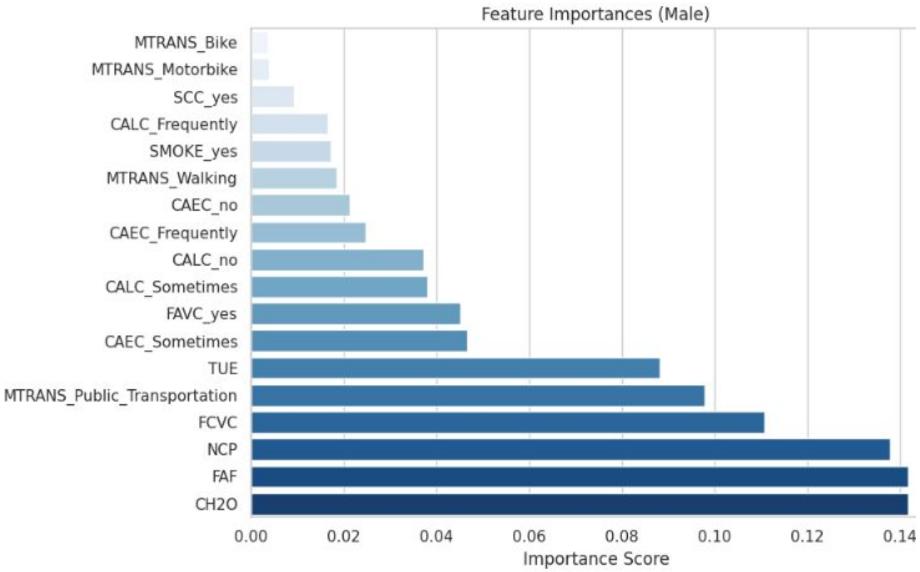




Feature Importance (Random Forest)

Without Weight, Height, Age, and family_history_with_overweight





Number of Records in Each Category

Women

NObeyesdad	count
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	count
Male	1068
Female	1043

Men

NObeyesdad	count
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

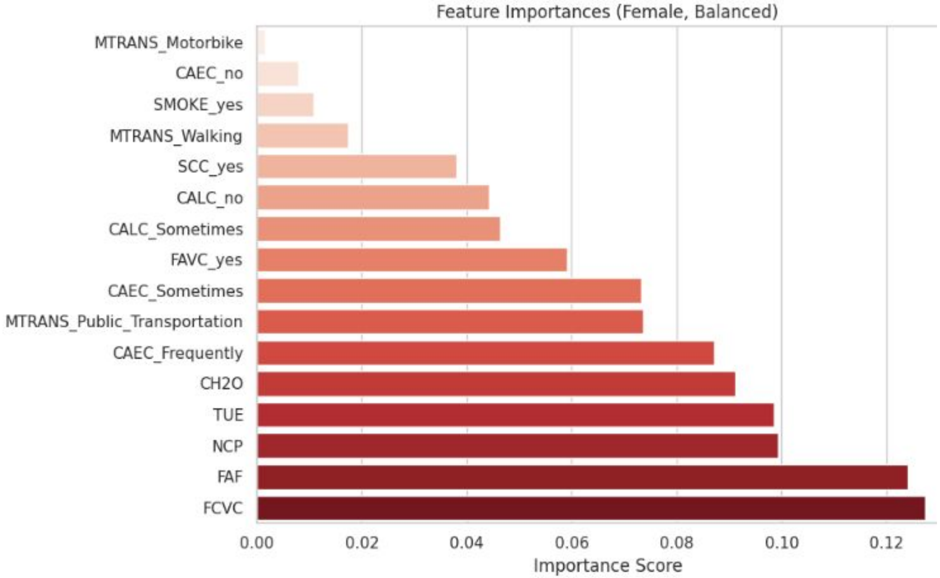
Women

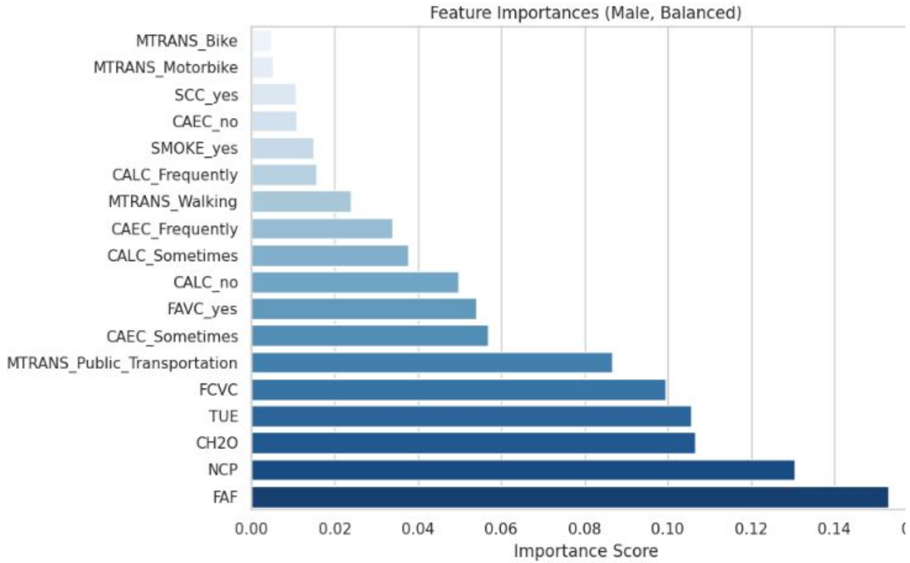
```
Female distribution:
Obesity_Grouped
Obesity                481
Overweight             248
Insufficient_Weight    173
Normal_Weight          141
Name: count, dtype: int64
```

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

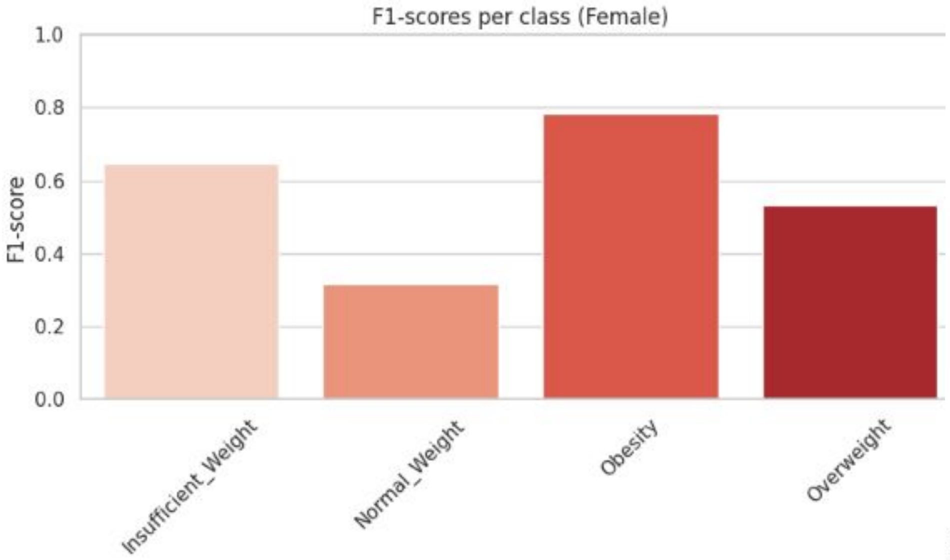
Classification Report (Female):

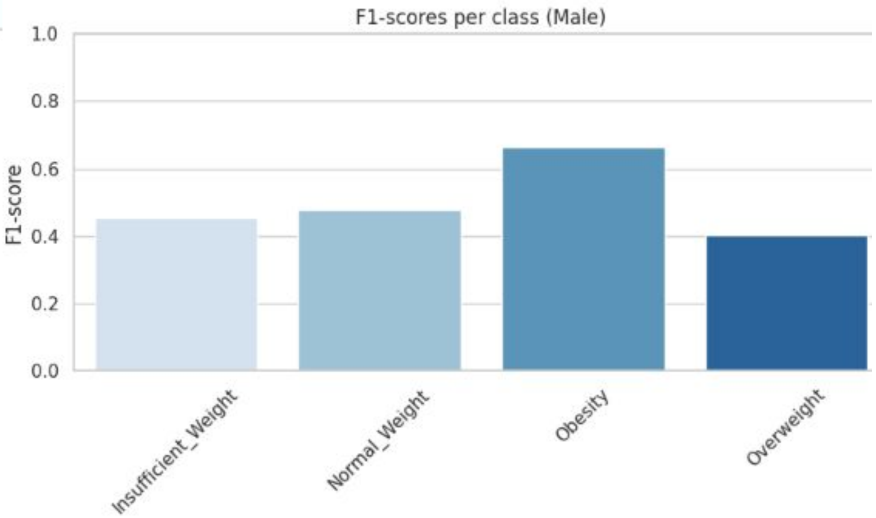
	precision	recall	f1-score	support
Insufficient_Weight	0.64	0.66	0.65	35
Normal_Weight	0.31	0.32	0.32	28
Obesity	0.84	0.74	0.78	96
Overweight	0.49	0.58	0.53	50
accuracy			0.63	209
macro avg	0.57	0.57	0.57	209
weighted avg	0.65	0.63	0.64	209

Classification Report (Male):

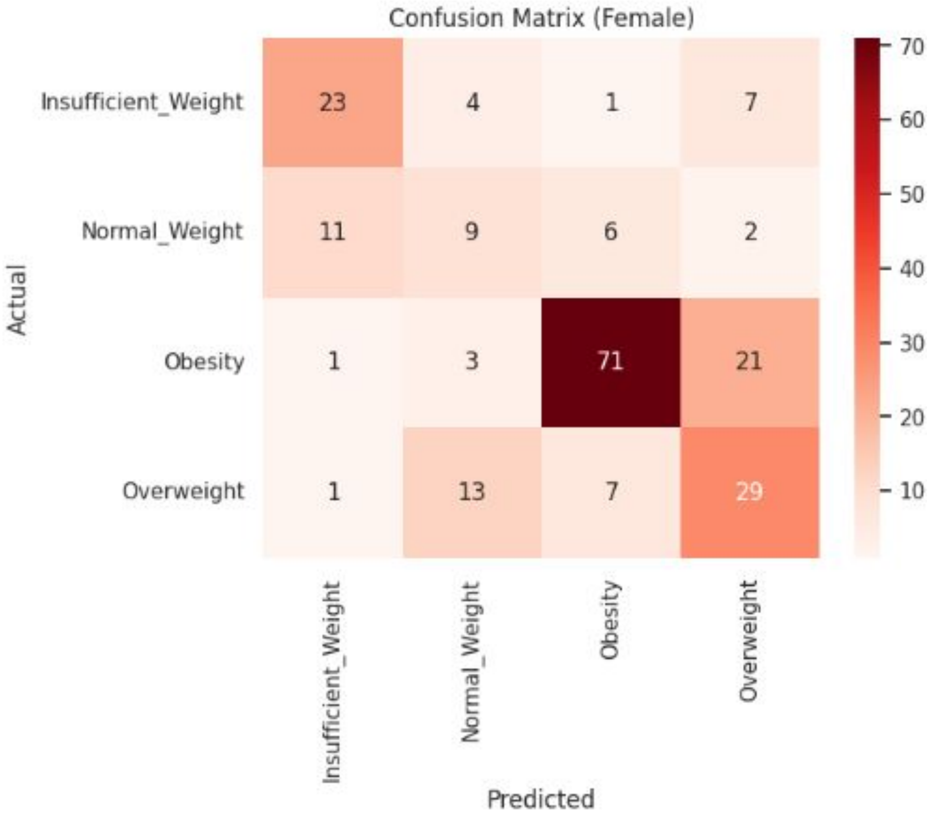
	precision	recall	f1-score	support
Insufficient_Weight	0.33	0.75	0.45	20
Normal_Weight	0.44	0.52	0.48	29
Obesity	0.68	0.64	0.66	98
Overweight	0.52	0.33	0.40	67
accuracy			0.54	214
macro avg	0.49	0.56	0.50	214
weighted avg	0.57	0.54	0.54	214

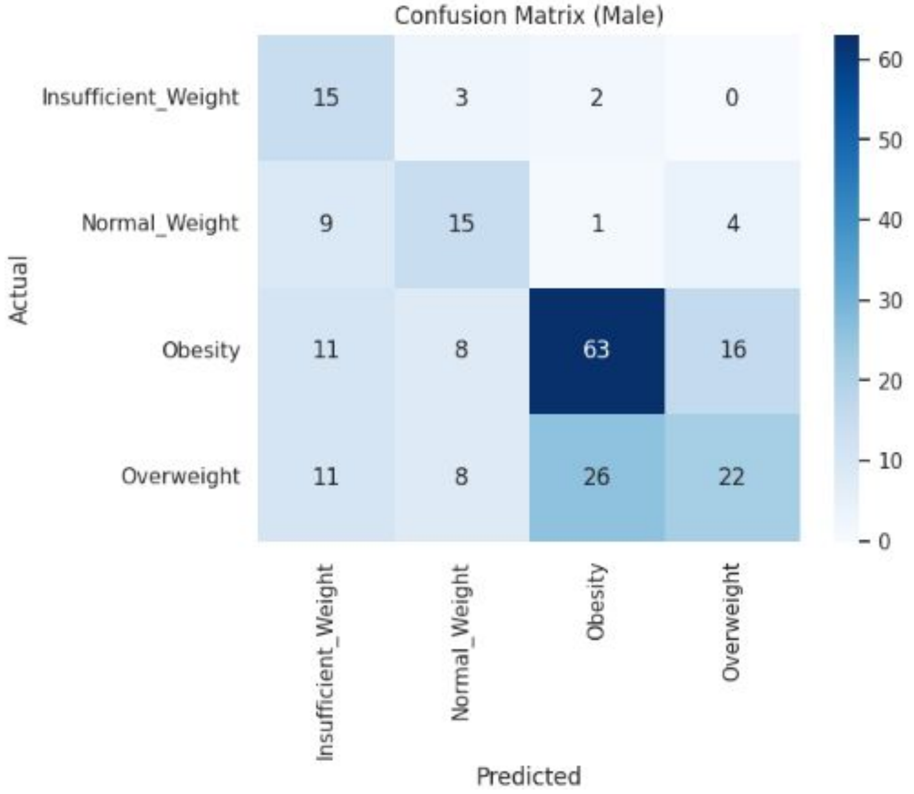
F1-Score Visualization





Confusion Matrix





Model Performance Overview

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

Women

- **Average f1 score:** 0.57.
- **Best prediction:** Obesity -> 0.78.
- **Worst prediction:** Normal_Weight -> 0.32.
- **Often confused:** Obesity with Overweight -> 21 cases out of 96.

Men

- **Average f1 score:** 0,4975.
- **Best prediction:** Obesity -> 0.66.
- **Worst prediction:** Overweight -> 0.40.
- **Often confused:** Overweight with Obesity -> 26 cases out of 67.

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!