

# Detecting Obesity Level

With Predictive Modeling

BUS AN 516 | Team 9 | Hyewon Jeong, Samvit Patankar,  
Lucy Tsai, Weihan Weng, Yoey Zhang

# Table of Contents

Introduction

Research Purpose and Questions

Data Overview & Details

Data Exploration

Predictive Models and Findings

Summary

# Introduction

- Project Overview:  
Analyzing factors affecting obesity levels across Mexico, Peru, and Colombia by examining dietary habits, physical activities, lifestyle, and demographic factors. Constructing predictive models aimed to identify individuals at high risk of obesity.
- Motivation & Audience :  
Obesity is a significant global health concern, impacting individuals' well-being. By understanding the factors that cause obesity, healthcare professionals and policymakers can make informed decisions to prevent and manage the issue effectively.

# Research Purpose and Questions

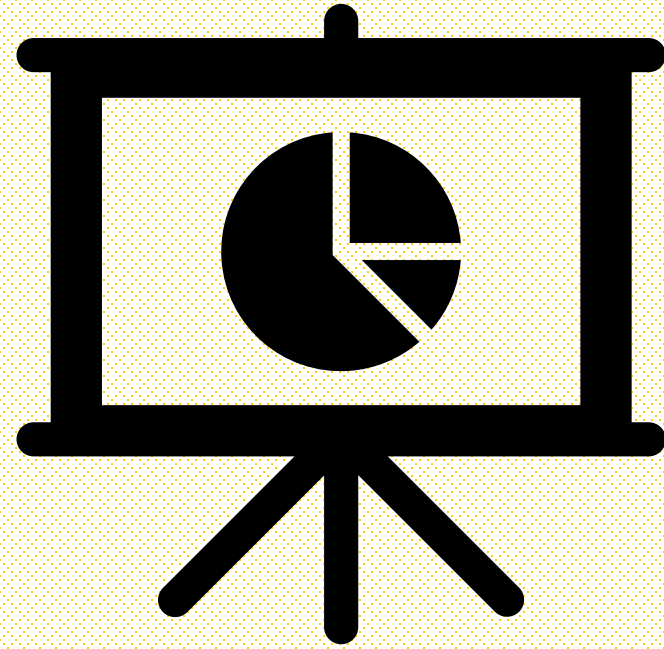


# Research Purpose & Questions

Can we accurately predict obesity levels based on dietary habits, physical activity, lifestyle, and demographic factors?

What are the key factors influencing obesity levels?

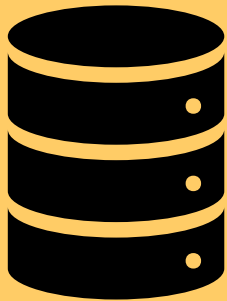
# Data Overview & Details



# Dataset Overview

Total number of records: 2111, no missing values

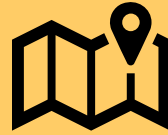
## Source



[Kaggle](#)

Collected via survey  
at a web platform

## Scope



Mexico, Peru, and  
Colombia



Health condition

## Timespan



Assumed to  
represent a snapshot  
at the time of data  
collection

# Dataset Details

Total variables: 16

<i>Demographics &amp; Health Factors</i>	<i>Eating Habits</i>	<i>Physical Condition</i>
Gender	Frequent consumption of high caloric food (FAVC)	Calories consumption monitoring (SCC)
Age	Frequency of vegetables consumption (FCVC)	Physical activity frequency (FAF)
Height	Number of main meals (NCP)	Time using technology devices (TUE)
Weight	Consumption of food between meals (CAEC)	Transportation mode used (MTRANS)
Family history	Consumption of water daily (CH20)	
Smoke	Consumption of alcohol (CALC)	

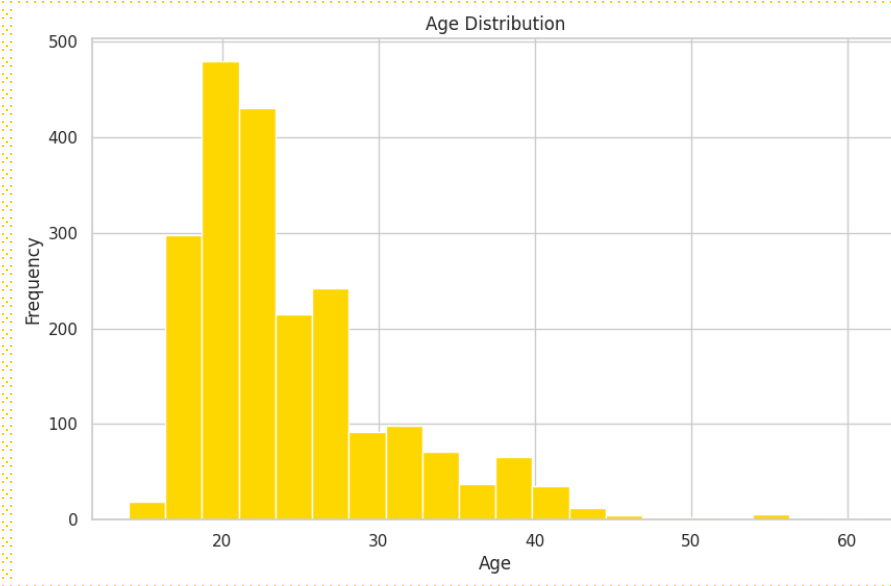
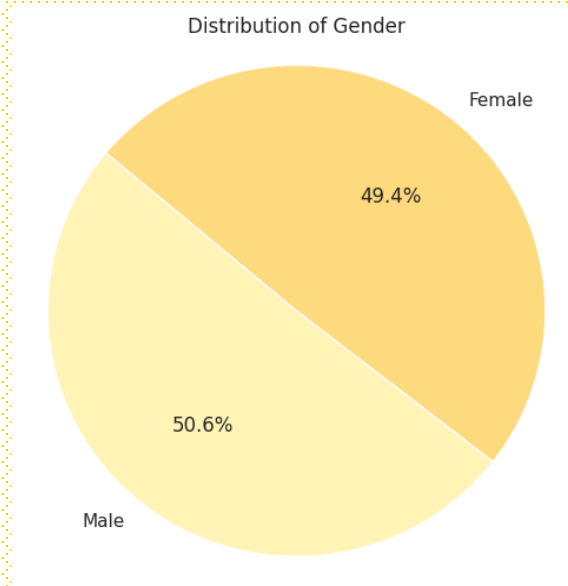


# Data Exploration



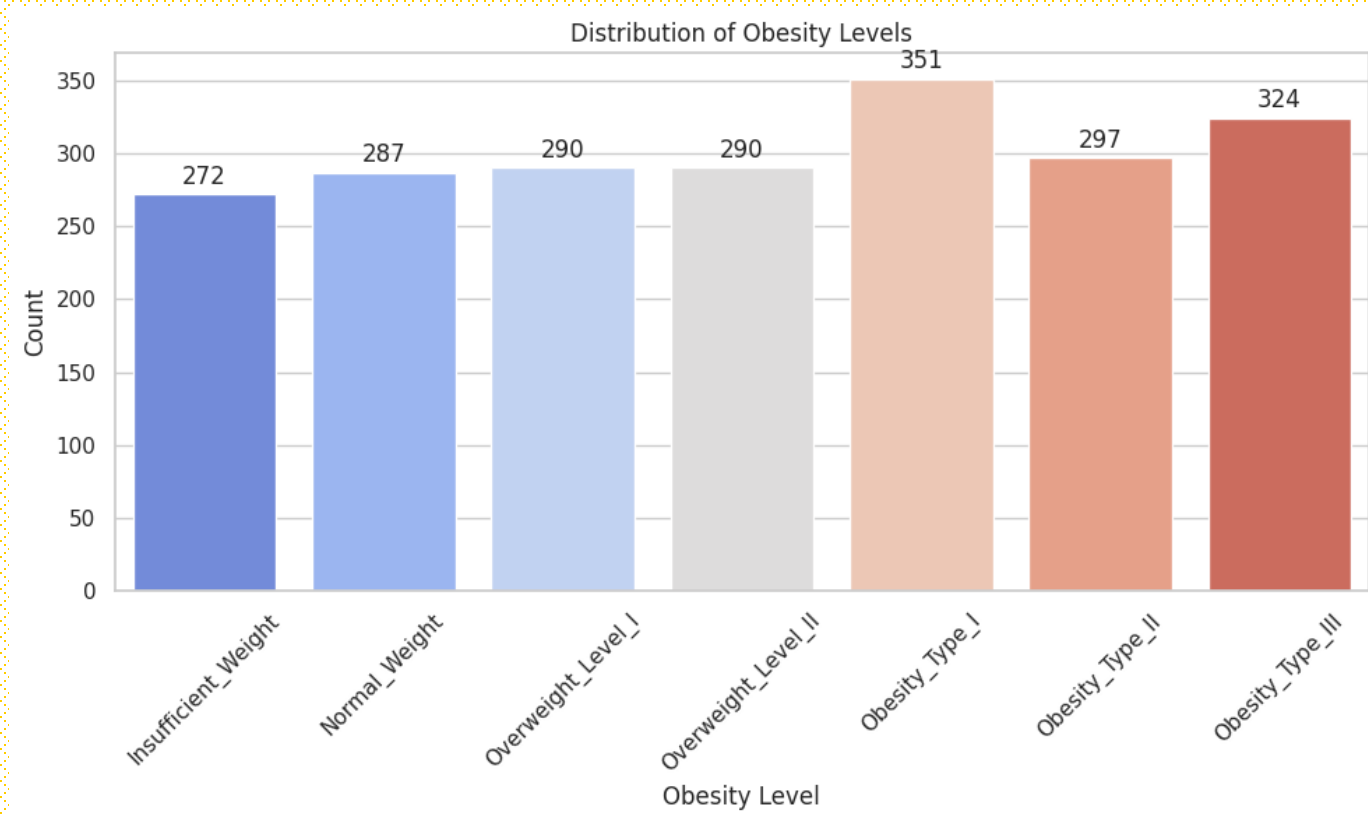
# Exploration – Descriptive Statistics

- Genders are evenly distributed.
- Age ranges from 14 to 61, with a mean age of 24.3.
- The dataset predominantly comprises younger individuals, possibly due to the online survey method.



# Exploration – Descriptive Statistics

- Distribution of weight level



- BMI < 18.5 → Insufficient
- $18.5 \leq \text{BMI} < 25$  → Normal
- $25 \leq \text{BMI} < 30$  → Overweight
- BMI  $\geq 30.0$  → Obesity

# Exploration – Descriptive Statistics

- **Insights:** Splitting the gender categories for obesity distribution reveals a higher prevalence of males in other obesity categories, with noticeable gender bias observed in obesity types II and III.
- **Conclusion:** The data suggests a higher count of females in lower obesity levels. Combining obesity type II and III into one category, as well as grouping overweight categories together, would reduce model complexity.



# Exploration – Descriptive Statistics

- **Insights:** Height and Weight show a strong positive correlation, which is expected as taller individuals generally weigh more.
- **Conclusion:** Either Height or Weight can be dropped from our key variables, and we decided to remove Weight. Height tends to remain relatively constant for adults, while weight can fluctuate due to various factors.

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Age	1.000000	-0.025958	0.202560	0.016291	-0.043944	-0.045304	-0.144938	-0.296931
Height	-0.025958	1.000000	0.463136	-0.038121	0.243672	0.213376	0.294709	0.051912
Weight	0.202560	0.463136	1.000000	0.216125	0.107469	0.200575	-0.051436	-0.071561
FCVC	0.016291	-0.038121	0.216125	1.000000	0.042216	0.068461	0.019939	-0.101135
NCP	-0.043944	0.243672	0.107469	0.042216	1.000000	0.057088	0.129504	0.036326
CH2O	-0.045304	0.213376	0.200575	0.068461	0.057088	1.000000	0.167236	0.011965
FAF	-0.144938	0.294709	-0.051436	0.019939	0.129504	0.167236	1.000000	0.058562
TUE	-0.296931	0.051912	-0.071561	-0.101135	0.036326	0.011965	0.058562	1.000000

# Exploration – Key Variables

Total Key Variables: 15/16

<i>Demographics &amp; Health Factors</i>	<i>Eating Habits</i>	<i>Physical Condition</i>
Gender	Frequent consumption of high caloric food (FAVC)	Calories consumption monitoring (SCC)
Age	Frequency of vegetables consumption (FCVC)	Physical activity frequency (FAF)
Height	Number of main meals (NCP)	Time using technology devices (TUE)
<del>Weight</del>	Consumption of food between meals (CAEC)	Transportation mode used (MTRANS)
Family history	Consumption of water daily (CH20)	
Smoke	Consumption of alcohol (CALC)	

# Approach 1:

## Predict obesity level by predicting BMI

Model 1: Linear Regression

# Model 1: Linear Regression for BMI predication

## Model Methodology

- BMI is calculated using the height and weight parameters present in the dataset in our original dataset.
- This new BMI parameter is used as the target variable to predict the Obesity level of a person.
- Linear regression is used to predict BMI with the independent variables defined.
- As BMI can be mapped to Obesity levels, we attempted to use a linear model to gain better accuracy predicting a numerical target variable (BMI) and then check the classification of the actual target variable (Obesity).



# Model 1: Linear Regression for BMI predication

## Model Training

### Independent Variables

gender, age, height,  
family\_history\_with\_overweight,  
requecy\_consumption\_of\_high\_calorie\_food,  
frequency\_consumption\_of\_vegetables,  
number\_of\_main\_meals,  
consumption\_of\_food\_between\_meals, smoker,  
consumption\_of\_water\_daily,  
calorie\_consumption\_monitoring,  
physical\_activity\_frequency,  
time\_using\_technology\_devices,  
consumption\_of\_alcohol

### Target Variable

BMI

## Model Evaluation

### BMI to Obesity Mapping

Insufficient\_Weight : BMI < 18.5

Normal\_Weight :  $18.5 \leq \text{BMI} < 24.9$

Overweight\_Level\_I\_and\_II:  $24.9 \leq \text{BMI} < 29.9$

Obesity\_Type\_I:  $29.9 \leq \text{BMI} < 34.9$

Obesity\_Type\_II\_and\_III:  $34.9 < \text{BMI}$

### Classification Report

Accuracy, Precision, Recall

# Model 1: Linear Regression for BMI predication

## Model Assumptions

- **Linearity:** The relationship between the independent variables (predictors) and the dependent variable (response) should be **linear**. This means that the change in the response variable is proportional to the change in the predictor variables.
- **No Multicollinearity:** The predictor variables **should not be highly correlated with each other**. Multicollinearity can lead to unstable estimates of the regression coefficients and makes it difficult to interpret the effects of individual predictors.

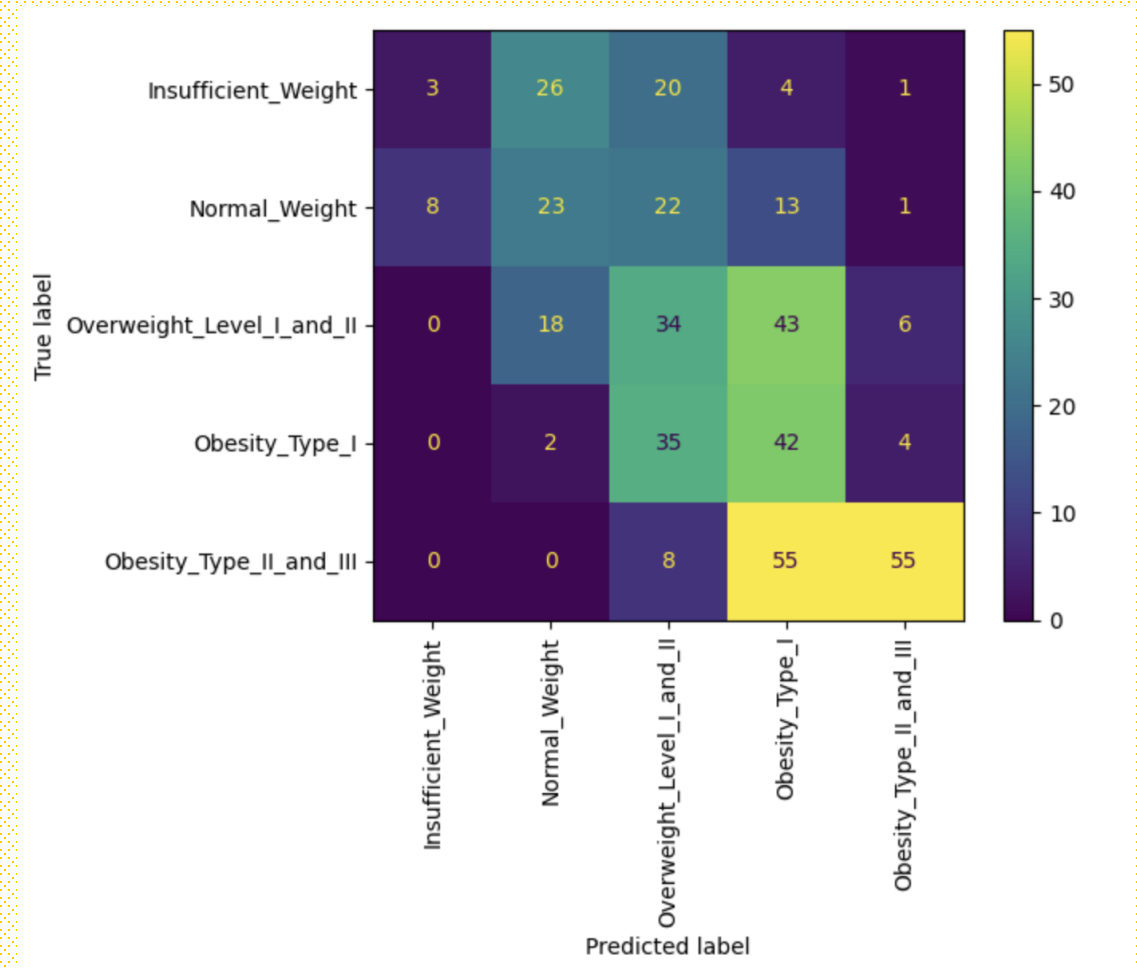
# Linear Model - Evaluation

Model Accuracy: **37%**

Weighted Precision: **43%**

Weighted Recall: **37%**

- The model performs well in predicting only **Obesity Type I**
- The model performs relatively poorly predicting **all other variables**.



# Approach 2:

## Classify the different obesity levels

Model 2: Ordinal Logistic Regression

Model 3: CART Model

# Model 2: Ordinal Logistic Regression

## Model Methodology

- The target variable of Obesity Level is converted into an ordinal variable in the following order: Insufficient weight , Normal Weight, Overweight level 1 and 2, Obesity level 1 and Obesity level 2 and 3.
- We used **stratify** option while splitting the data in train and test to make sure that there is not class imbalance in the test set.
- The model was chosen to classify with a better accuracy as the target variable can be converted in an ordinal fashion without losing any meaning.
- We used the following parameters for logistic regression:
  - Method: BFGS (Algorithm that iteratively tries to reduce the loss function)

# Model 2: Ordinal Logistic Model for Obesity Classification

## Model Training

### Independent Variables

gender, age, height,  
family\_history\_with\_overweight,  
requecy\_consumption\_of\_high\_calorie\_food,  
frequency\_consumption\_of\_vegetables,  
number\_of\_main\_meals,  
consumption\_of\_food\_between\_meals, smoker,  
consumption\_of\_water\_daily,  
calorie\_consumption\_monitoring,  
physical\_activity\_frequency,  
time\_using\_technology\_devices,  
consumption\_of\_alcohol

### Target Variable

Obesity Level (Ordinal)

## Model Evaluation

### Classification Report

Accuracy, Precision, Recall

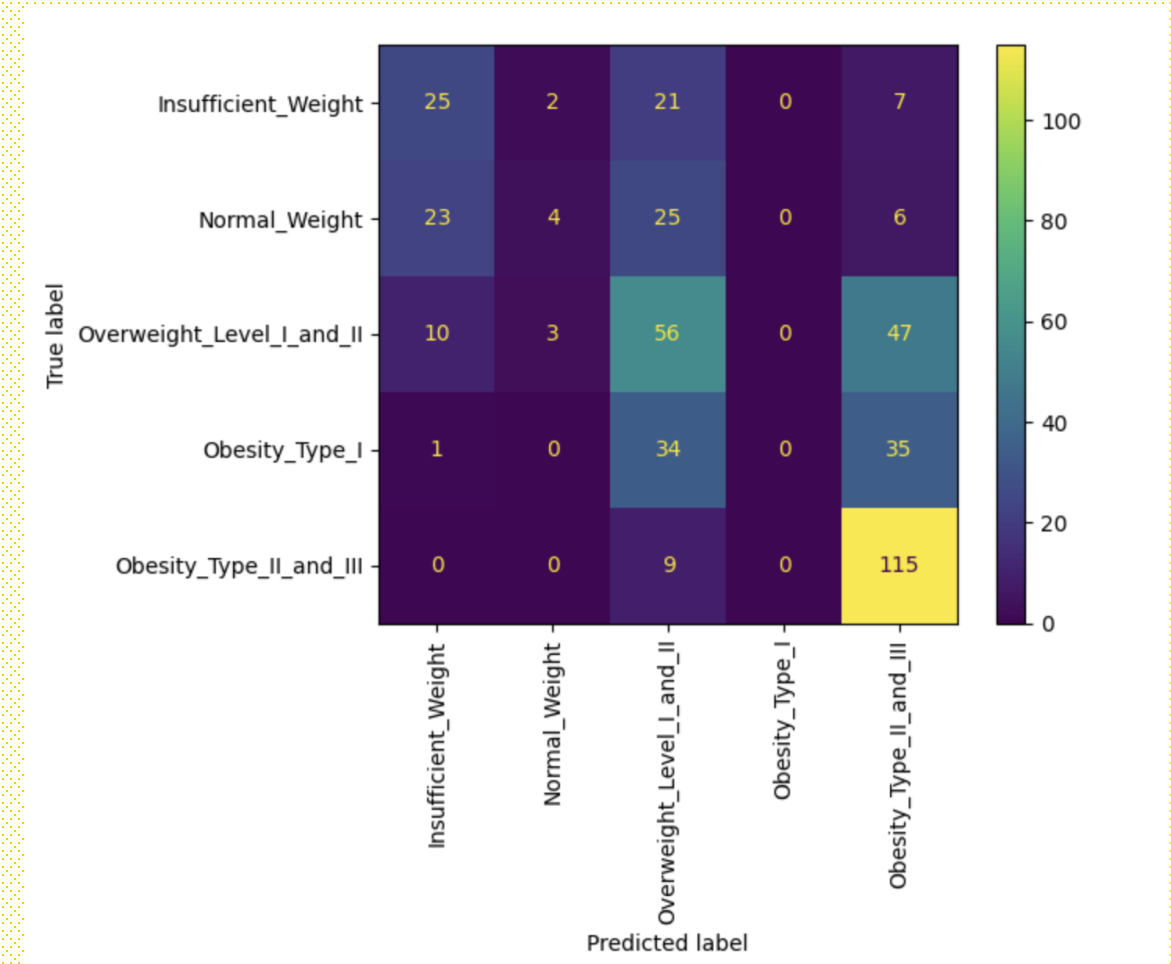
# Ordinal Logistic Regression- Evaluation

Model Accuracy: **47%**

Weighted Precision: **38%**

Weighted Recall: **47%**

- The model performs well in predicting **Insufficient Weight, Obesity Type II and III & Overweight Level I and II**
- The model performs relatively poorly predicting **Normal Weight and Obesity level I**



# Model 3: CART Model

## Model Methodology

### Decision Tree Model

- Excluding BMI and Weight, the remaining features are used as independent variables
- Obesity Level as dependent variable
- Set 20% of total rows as test set, and applied stratify option while splitting the data in train and test
- Standardized independent variables that are in numerical format, that is, scaling those to a standard normal distribution with a mean of 0 and a variance of 1

### Parameter Tuning

- Use Grid Search to find the best decision tree model
- During this process, we tuned three parameters:
  - `max_depth`: The maximum depth of the tree, used to control the growth depth of the tree
  - `min_samples_split`: The minimum number of samples required for node splitting
  - `min_samples_leaf`: the minimum number of samples required for leaf nodes
- The train set is divided into 5 parts, and each part is used as a validation set in turn for cross-validation.



# Model 3: CART Model

## Rebuild the model

- Rebuild the model with the best parameter combination obtained.
- The best parameters are:
  - max\_depth: 15
  - min\_samples\_leaf: 1
  - min\_samples\_split: 2
- Use the test set to evaluate the model – Run Classification Report

# Model 3: CART Model

## Model Assumptions

- Independent variables are continuous features and categorical features.
- Features are **independent** of each other, that is, the splitting of features is not affected by other features.
- There is **no linear relationship** between features.

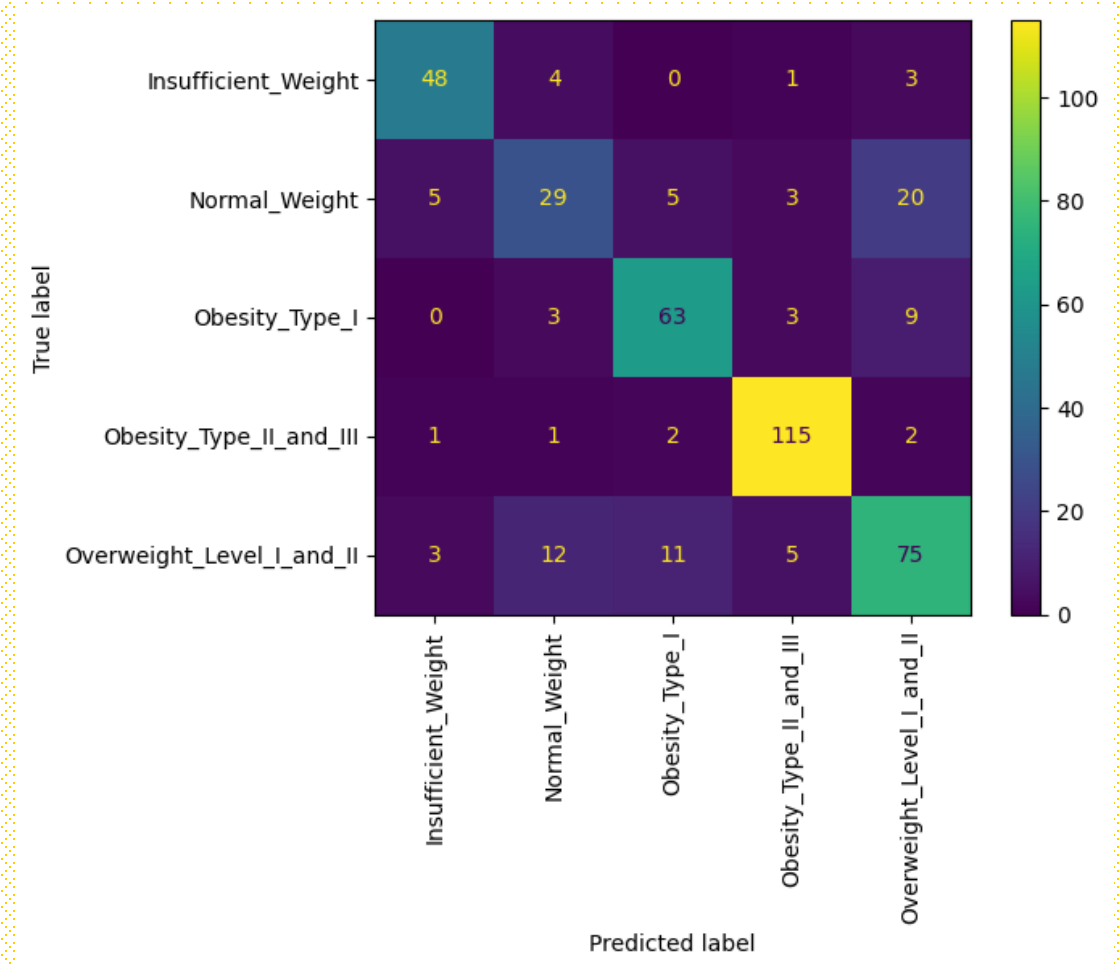
# CART Model - Evaluation

Model Accuracy: **78%**


Weighted Precision: **77%**

Weighted Recall: **78%**

- The model performs well in predicting **Insufficient Weight, Obesity Type I, II and III**
- The model performs relatively poorly in predicting **Normal Weight & Overweight Level I and II**

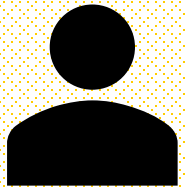


# Summary

	Linear Regression	Ordinal Logistic Regression	 CART
Accuracy	37%	47%	<b>78%</b>
Weighted Precision	43%	38%	<b>77%</b>
Weighted Recall	37%	47%	<b>78%</b>

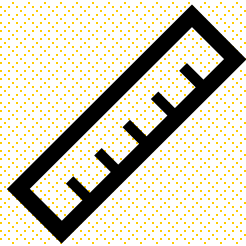
# Summary

## Top 5 key factors influencing obesity levels



**Age**

Importance: 15.4%



**Height**

Importance: 14.5%



**Frequency  
consumption of  
vegetables**

Importance: 12.5%



**Number of main  
meals**

Importance: 10.4%



**Time using  
technology devices**

Importance: 8.9%

Thank you! Any questions?

