ML 1

# Multi-Class Prediction of Obesity Risk


Prepared by:

- Amina Mohamed
- Ashraf Mahmoud
- Nagham Ehab
- Shorouq Hossam

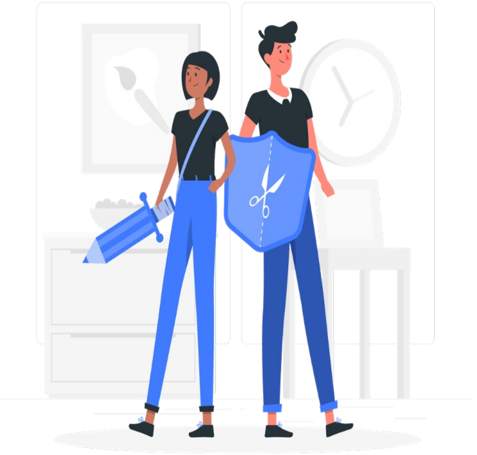
Under the Supervision of:

- Eng. Abdelrahman A. Eid

Information Technology Institute

# Table of Contents

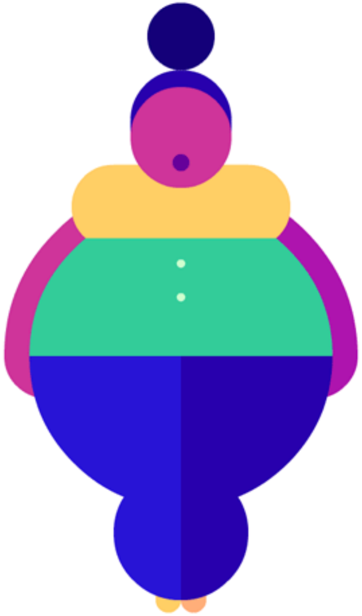
# 01. Problem

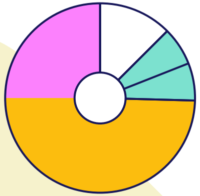


# Introduction

- The dataset we used was generated from a deep learning model trained on **the Obesity or CVD risk dataset.**

- The original dataset consist of the estimation of obesity levels in people from the countries of *Mexico, Peru and Colombia*, **77%** of the data was generated synthetically using the **Weka** tool and the **SMOTE** filter, **23%** of the data was collected directly from users through a web platform with a survey where anonymous users answered each question.






# Problem Statement

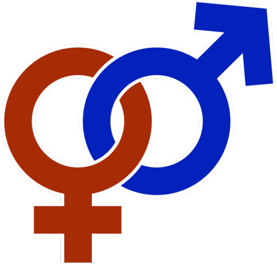
- The obesity burden has increased worldwide in recent decades. According to the **World Health Organization (WHO).**
- We will delve into the intricate relationship between lifestyle choices and weight management. By meticulously analyzing of *'Multi-Class Prediction of Obesity Risk'* data, we aim to unravel the key factors contributing to weight gain.
- Understanding of obesity classification using machine-learning techniques based on physical activity and nutritional habits.

# Meet our Dataset

The dataset contain **18** variables:

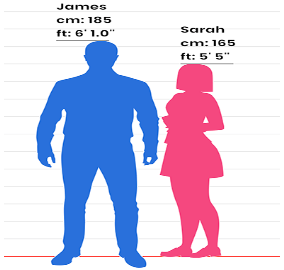

**Id**

person id



**Gender**

Male
Female




**Height**

1.45 to 1.98 (m)

**Weight**

39 to 165 (KG)

**Family history with**

Yes/No Q.

**FAVC**

Yes/No Q.

# Meet our Dataset, Cont'd

**FCVC**

(1-3 scale) Q.

**NCP**

Number of main meals

**CAEC**

Consumption of food between meals

**Smoke**

Yes/No Q.

**CH2O**

Daily water consumptio n

**SCC**

Do you monitor your calories consumption?

# Meet our Dataset, Cont'd

**Age**

14 to 61 years old

**FAF**

Physical activity frequency

**TUE**

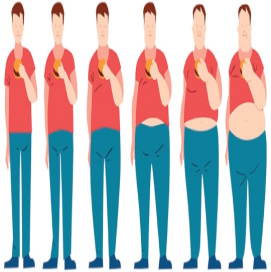
Time using technology devices

**CALC**

Alcohol consumption

**MTRANS**

Transportation used



**NObeyesdad**

7 types of obesity (Target)

# Profile Report

Overview | Variables | Interactions | Correlations | Missing values | Sample

Overview | Alerts (14) | Reproduction

| | |
|---|---|
| Gender is highly overall correlated with Height and 1 other fields | High correlation |
| Height is highly overall correlated with Gender | High correlation |
| NObeyesdad is highly overall correlated with Gender and 1 other fields | High correlation |
| Weight is highly overall correlated with family_history_with_overweight | High correlation |
| family_history_with_overweight is highly overall correlated with NObeyesdad and 1 other fields | High correlation |
| FAVC is highly imbalanced (57.9%) | Imbalance |
| CAEC is highly imbalanced (61.0%) | Imbalance |
| SMOKE is highly imbalanced (90.7%) | Imbalance |
| SCC is highly imbalanced (79.0%) | Imbalance |
| MTRANS is highly imbalanced (63.7%) | Imbalance |
| id is uniformly distributed | Uniform |
| id has unique values | Unique |
| FAF has 5044 (24.3%) zeros | Zeros |
| TUE has 6566 (31.6%) zeros | Zeros |

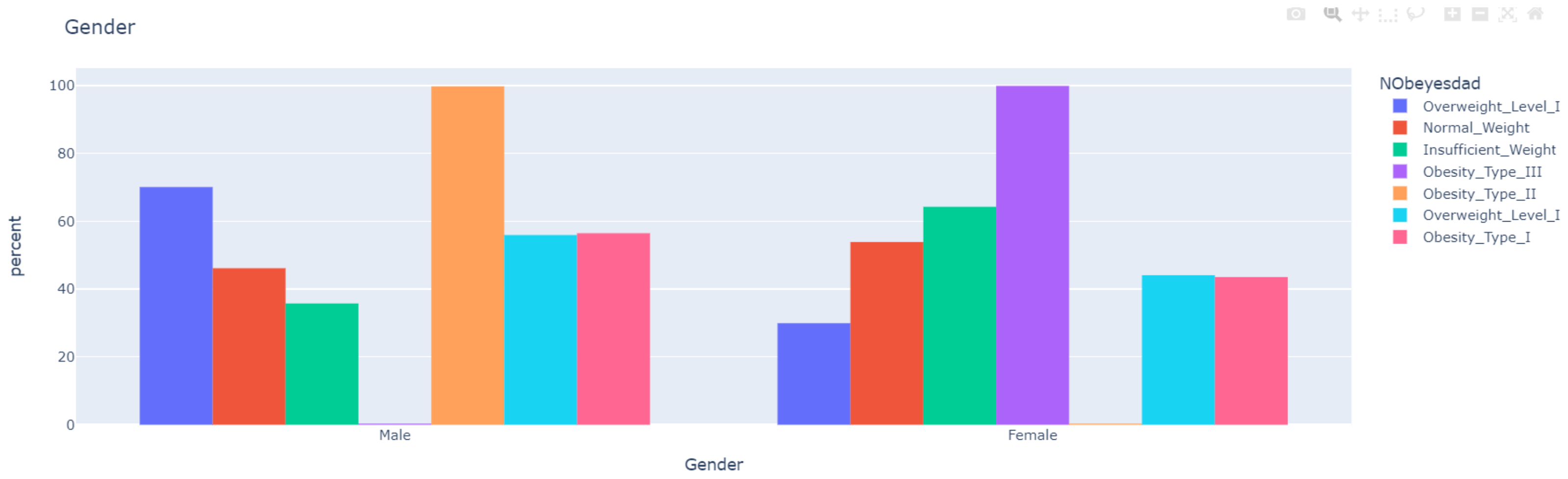
Report generated by YData.

02.

# EDA

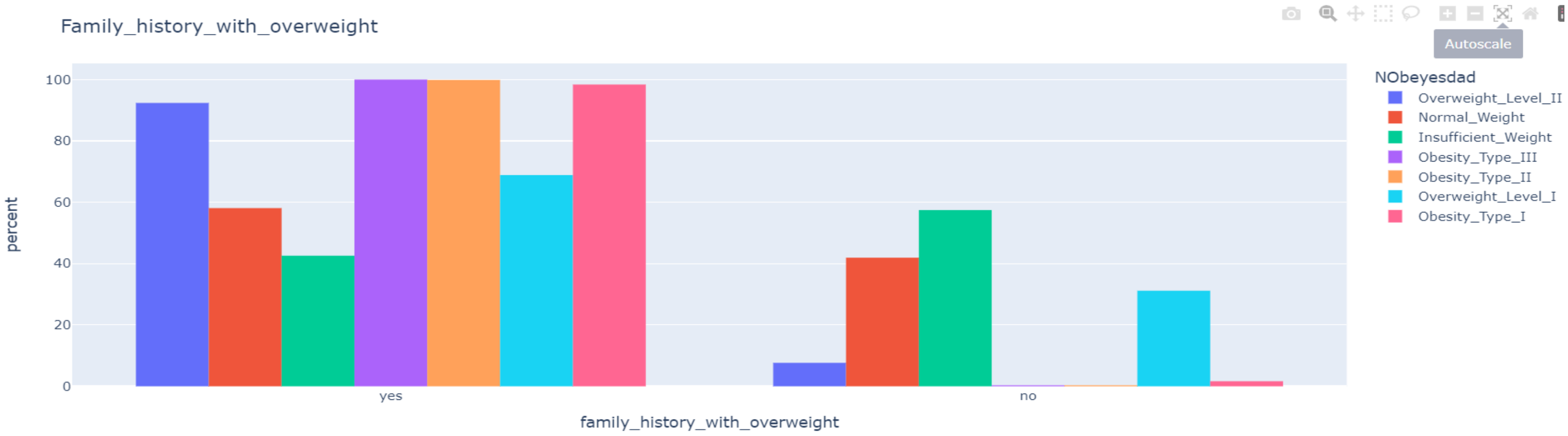
Every number tells a story

# Insights



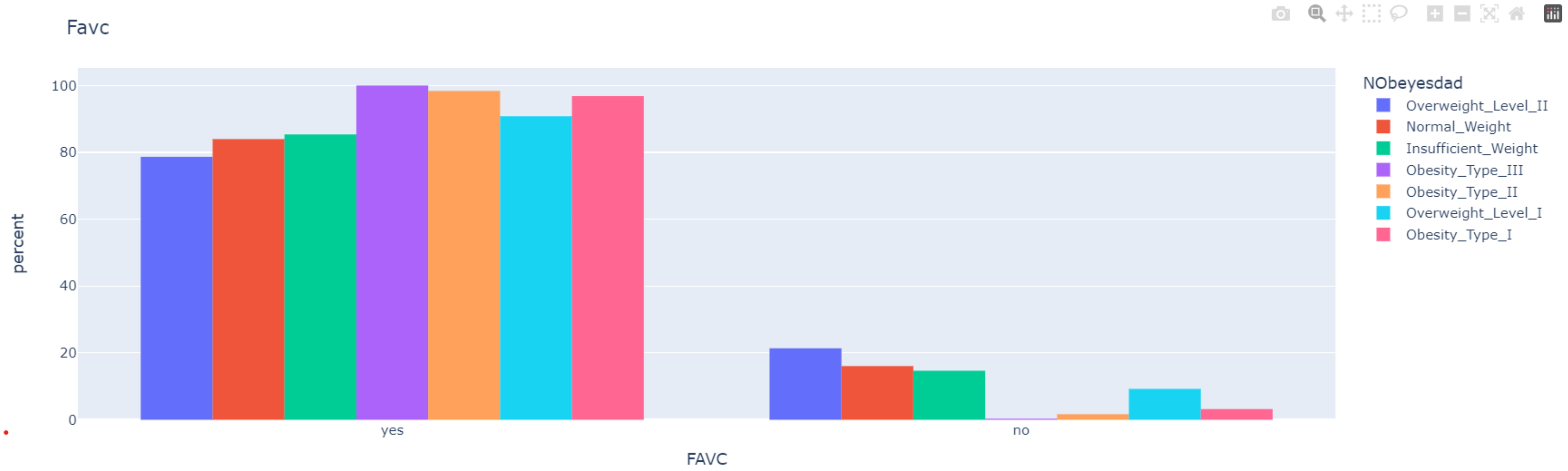

- 99.8%  of people suffering from obesity_type_3 are females.
- 99.7% of people suffering from obesity_type_2 are males.

# Insights, Cont'd



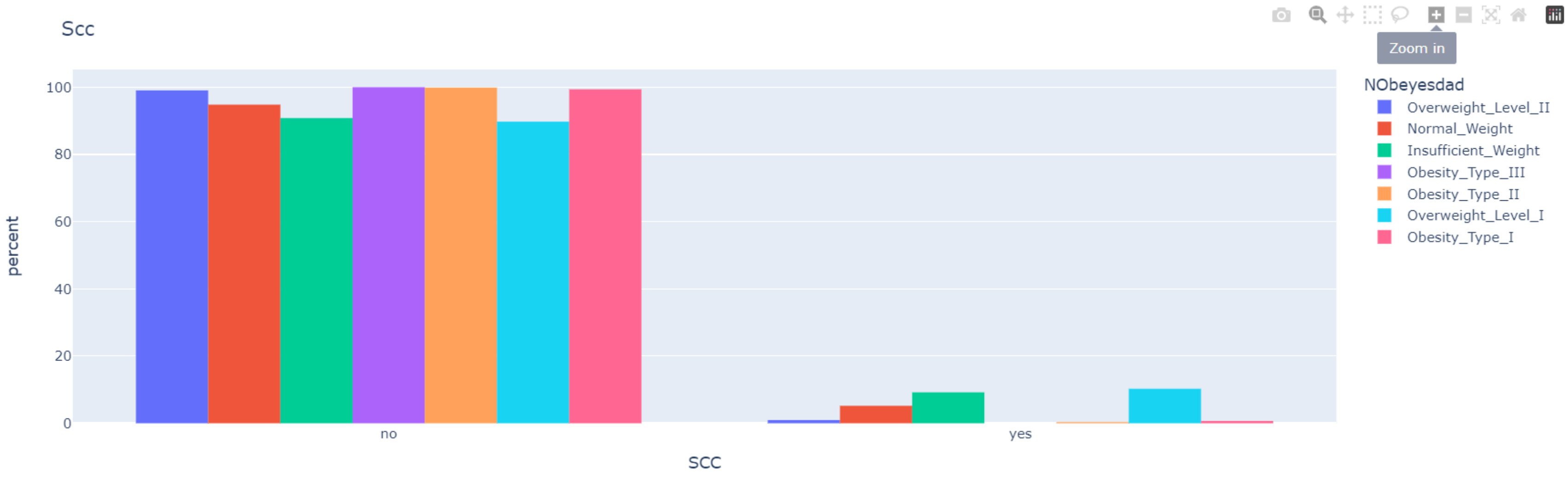

- Most of the people who are susceptible to obesity have a family history with overweight.

# Insights, Cont’d



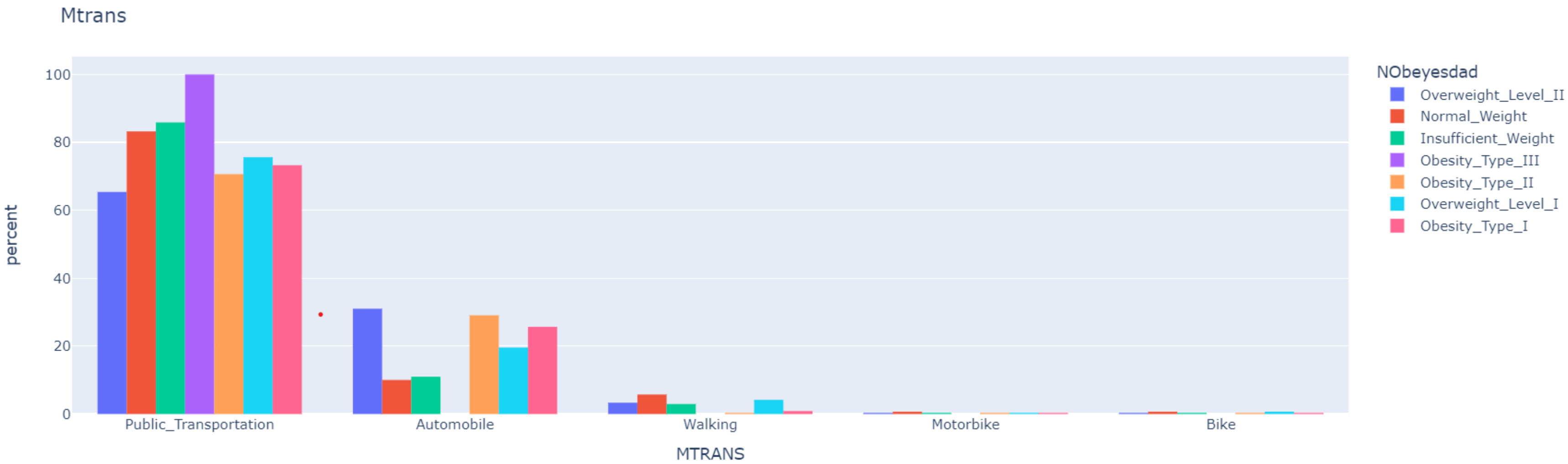

- people who consume food that contain a lot of calories are more susceptible to obesity.

# Insights, Cont'd



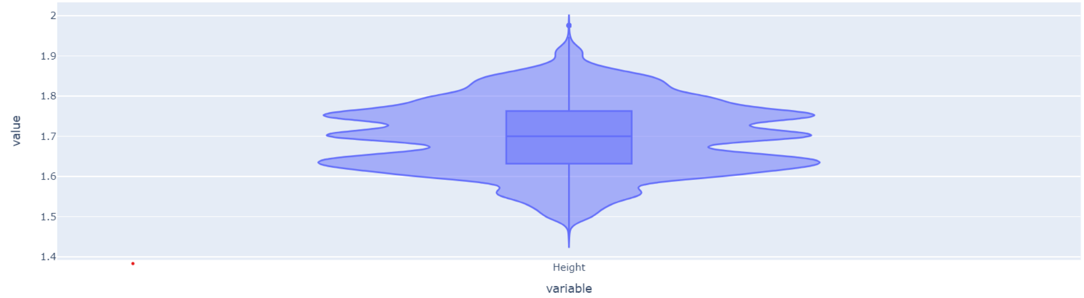

- people who track their calories consumption are less susceptible to obesity.

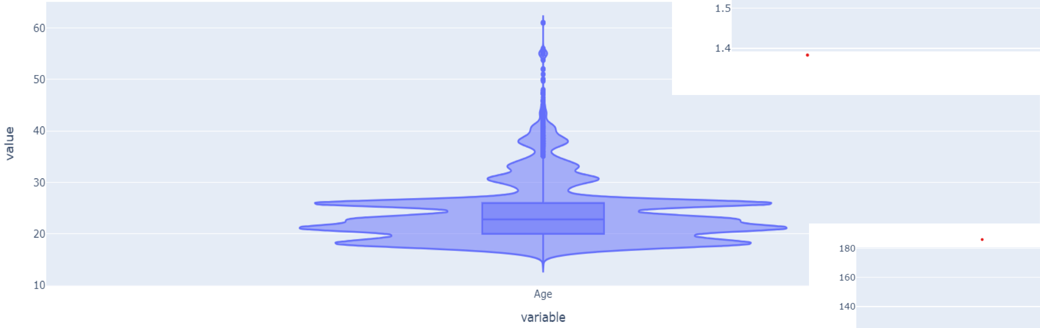
# Insights, Cont'd

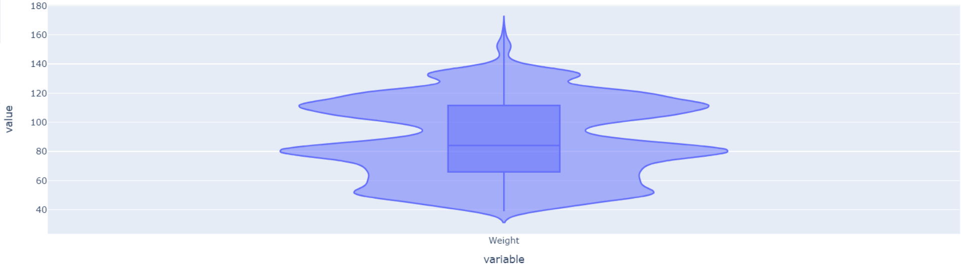



- people who prefer transportations that include physical activity are less susceptible to obesity.

# Insights, Cont'd

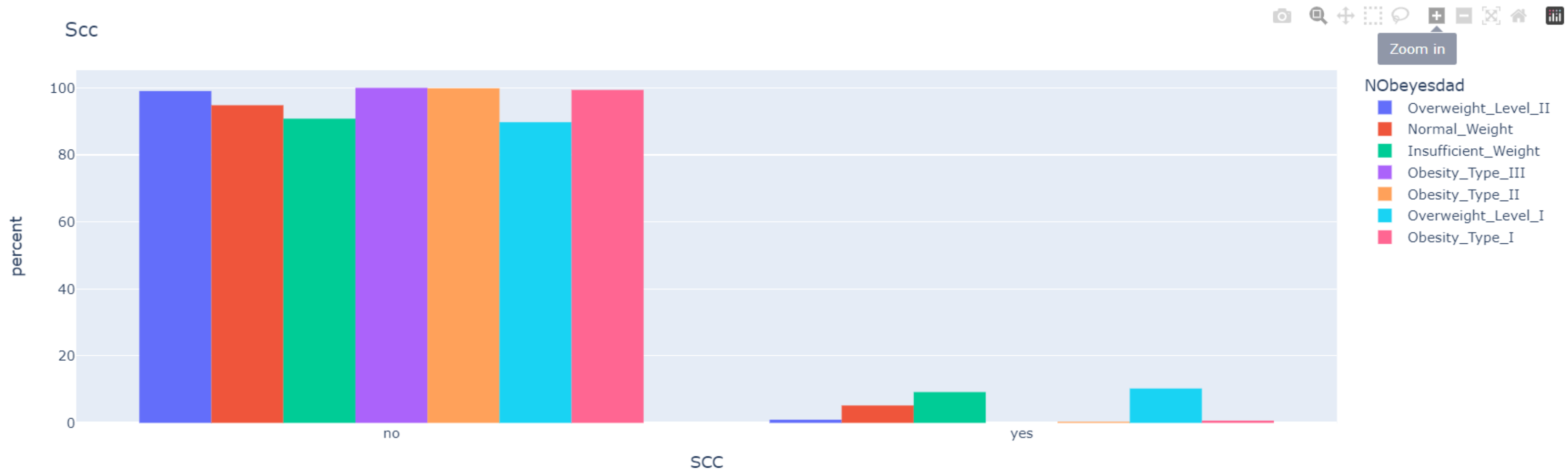


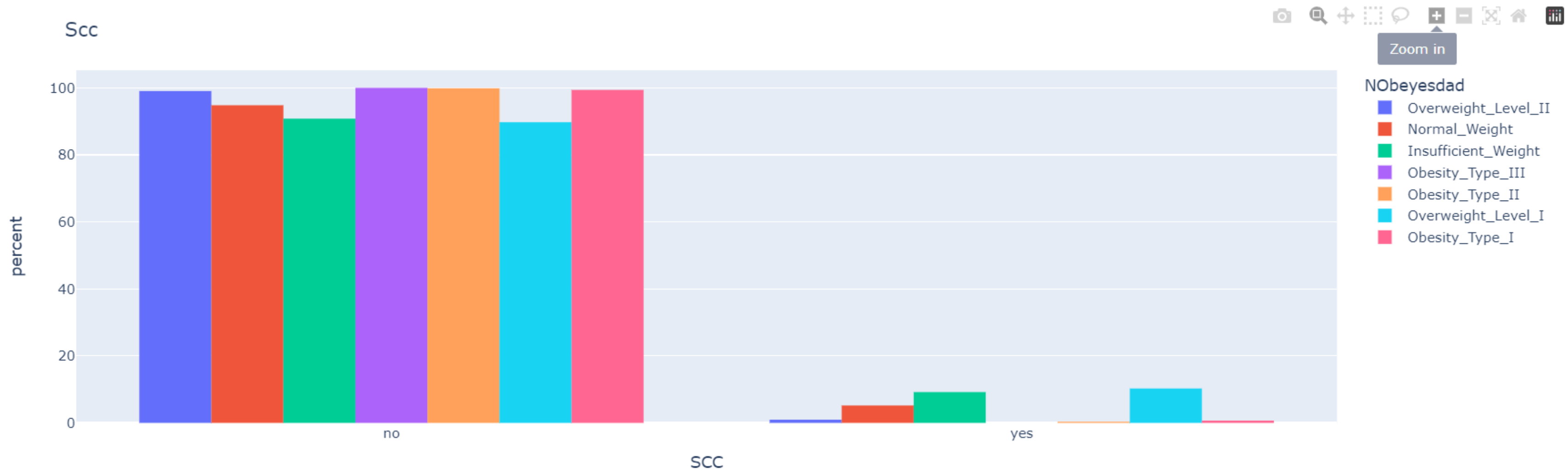

- There are Outliers present in Age.
- Age, height and Weight are normally distributed with some skewness.
- Some features are skewed.

# Insights, Cont'd



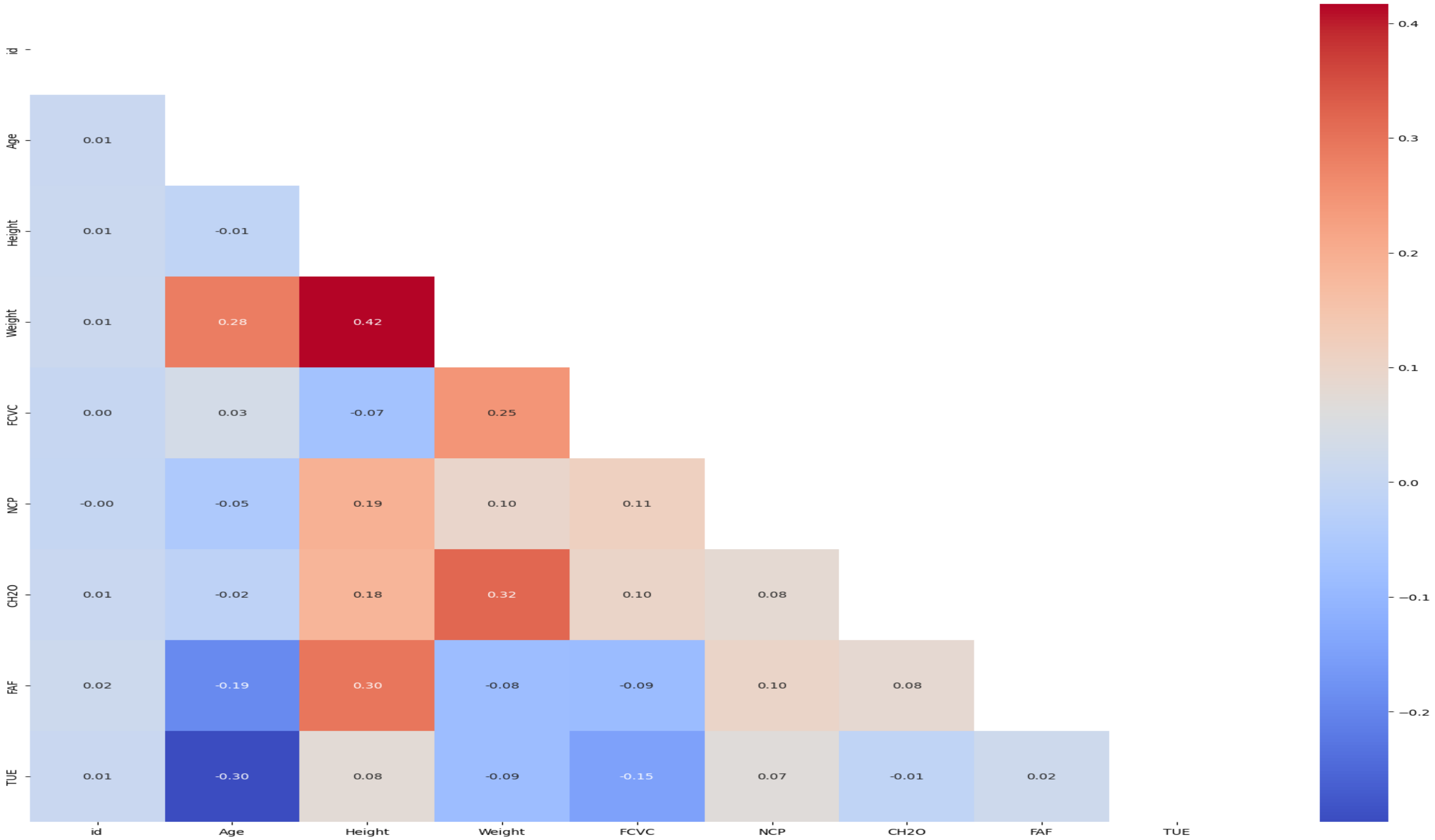

- people who track their calories consumption are less susceptible to obesity.

# Insights, Cont'd




- people who track their calories consumption are less susceptible to obesity.

# Insights, Cont'd



# 03. Training model on original data without any Features Extraction

# Data Preprocessing and Model Selection

## Data Preprocessing :

- Categorical Variables: One-Hot Encoding
- Numerical Variables: StandardScaler
- Target Variable: Label Encoding

## Model :

- Decision Tree
- No Feature Extraction
- Accuracy: 84%

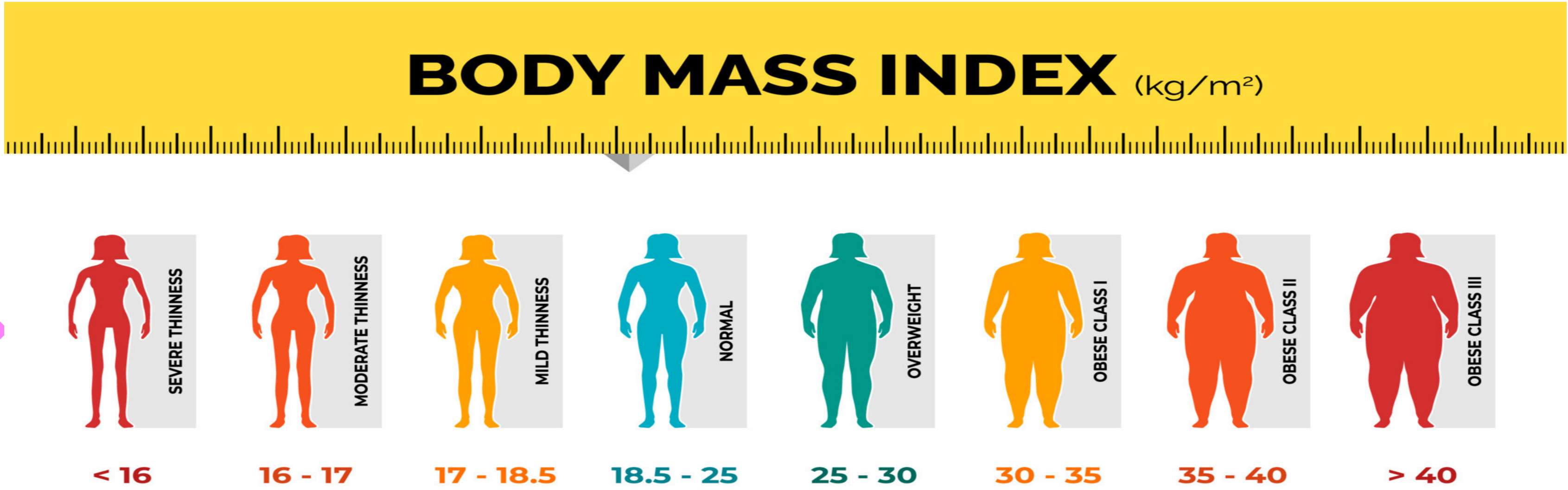
# Preprocessing

1. New Features Extraction

# New Features

After understanding the data very well, we could extract new features that can help us get more insights about the data and improve the performance of the models.

## 1. Body Mass Index (BMI)

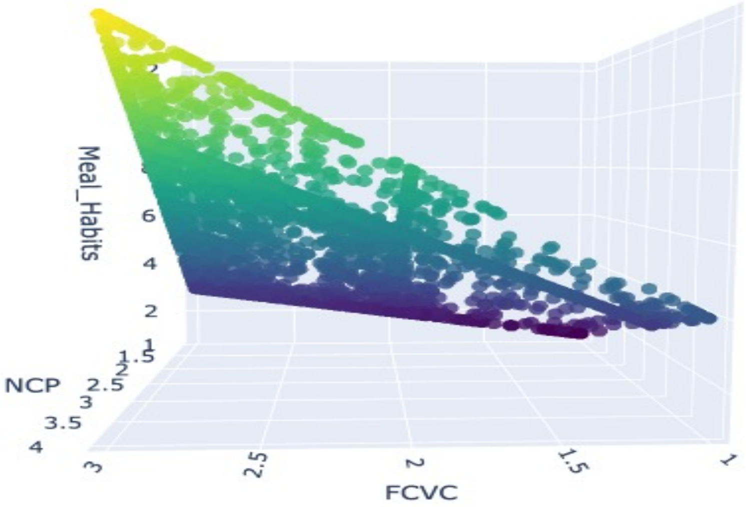


# BMI Categories

# New Features, Cont'd

## 2. Meal Habits

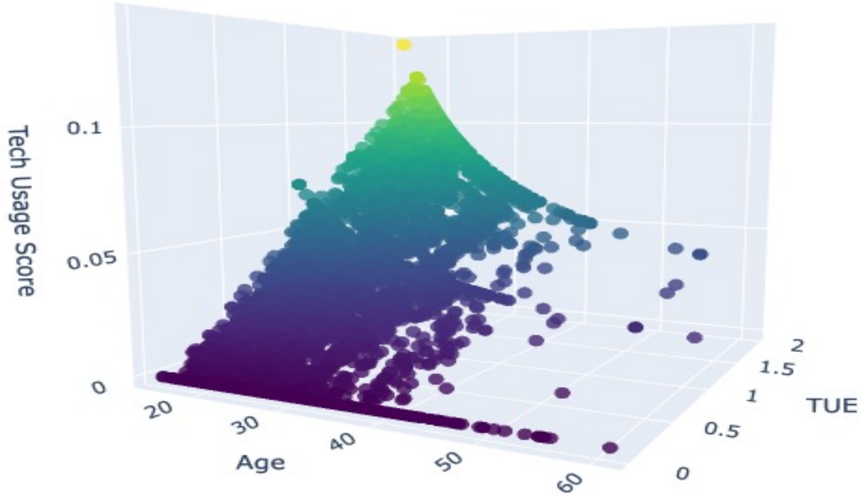
- Combination of **'FCVC'** (Frequency of consumption of vegetables) and **'NCP'** (Number of main meals) created the 'Meal_Habits' feature.
- This feature seeks to encapsulate overall dietary patterns, considering both the frequency of vegetable consumption and the number of main meals.



# New Features, Cont'd

## 3. Tech Usage Score

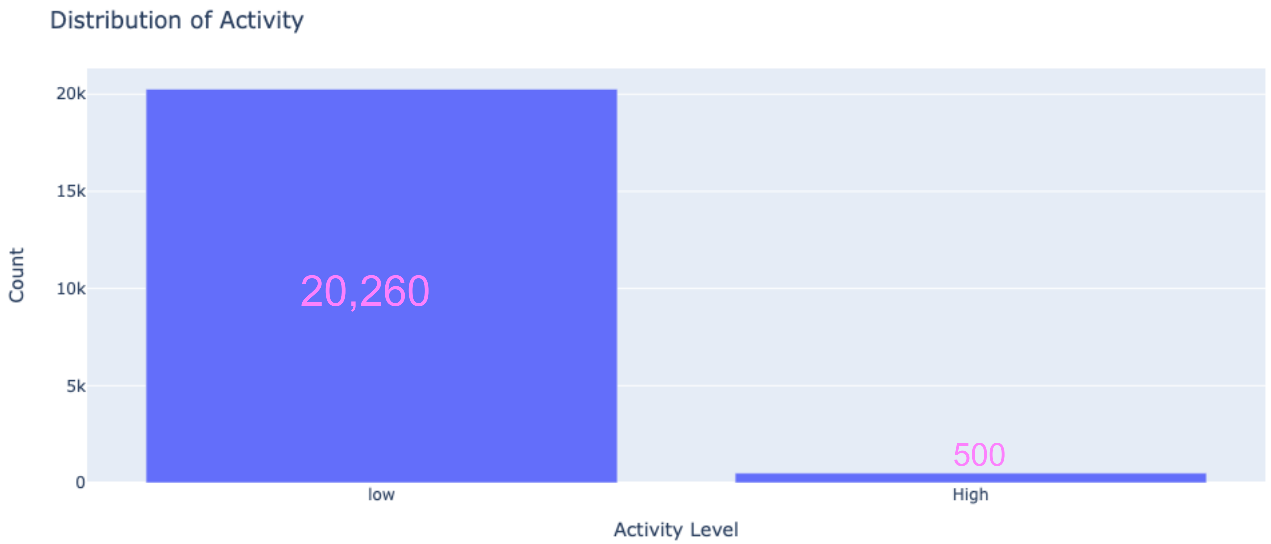
- A comprehensive score was crafted by weighting the frequency of technology usage **'TUE'** by the individual's **age**.
- The resulting 'Tech_Usage_Score' aims to quantify the average time spent using technology relative to the person's age, providing a nuanced perspective on technology habits.



# New Features, Cont'd

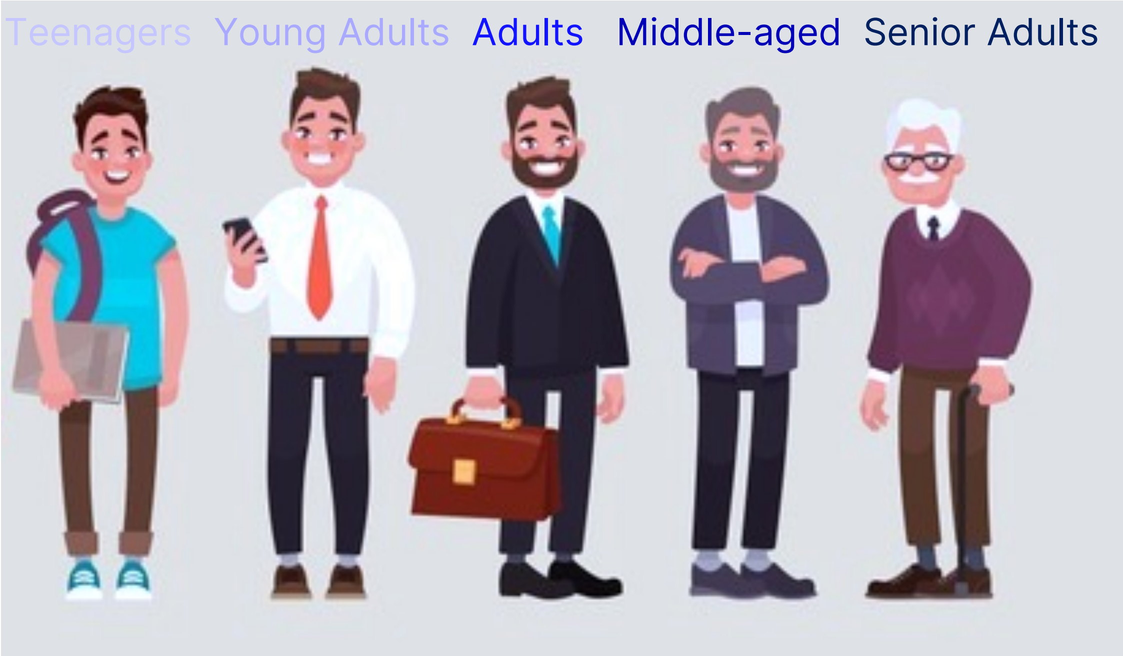
## 4. Activity

- If the person use bike or walk, his activity will be **high**.
- If the person use transportation, car, motorbike, activity will be **low**.

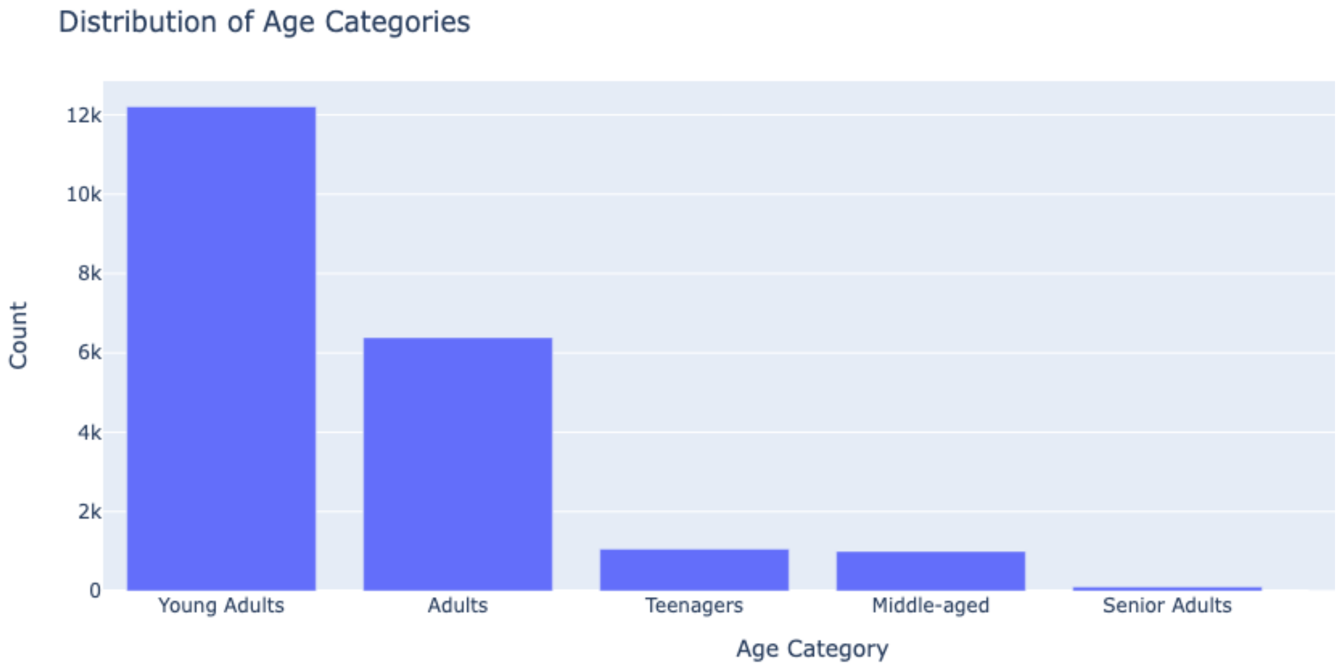


# New Features, Cont'd

5. **Age Categories**



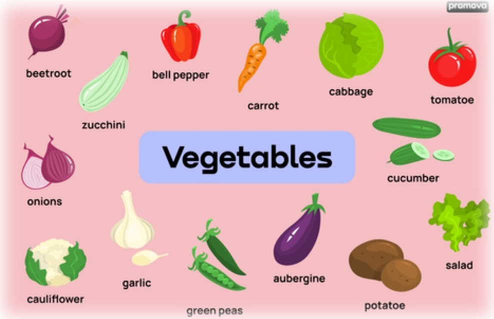
# Age Categories

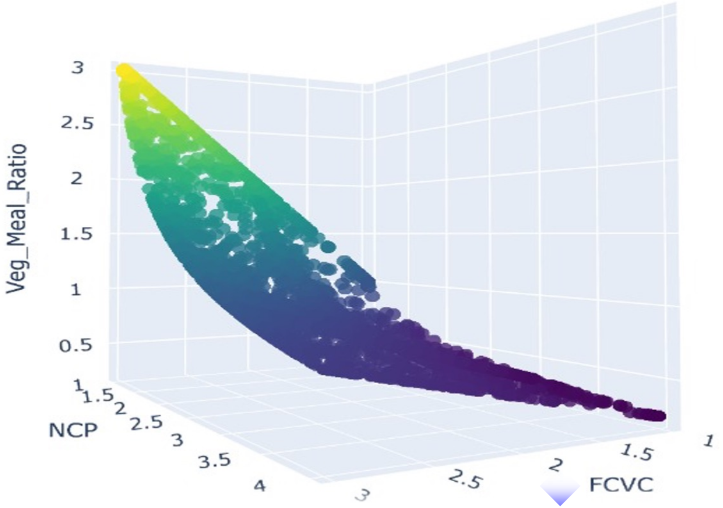


# New Features, Cont'd

## 6. Vegetable meal ratio(Veg_Meal_Ratio):

- X_['Veg_Meal_Ratio'] = X_['FCVC'] / X_['NCP'] .
- This ratio helps in understanding the relationship between the frequency of vegetable consumption and the number of main meals.



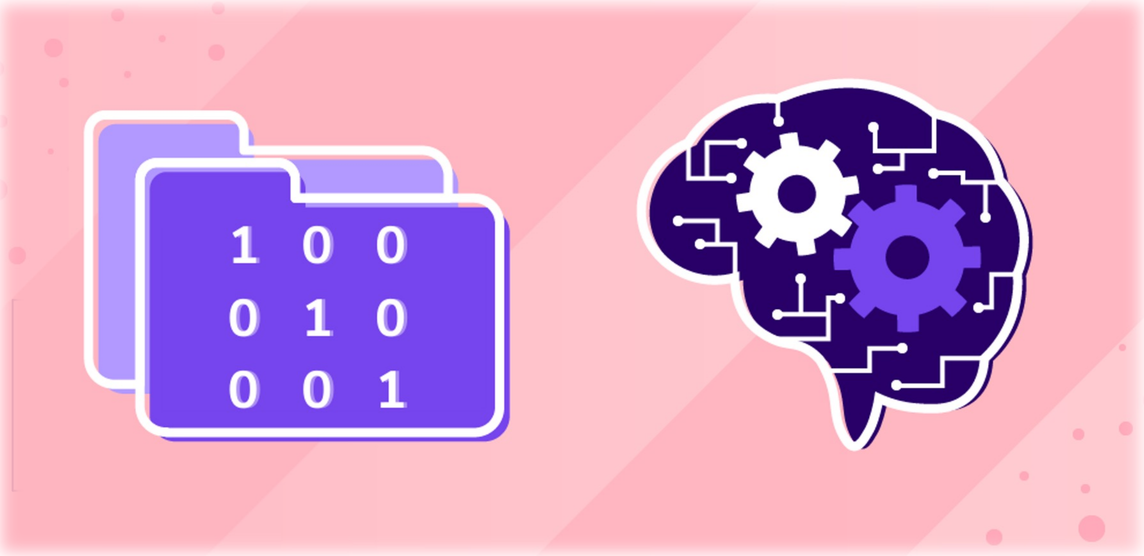



# Preprocessing , Cont’d

## 2.Custom One-Hot Encoding

# 2.Custom One-Hot Encoding

- Handles categorical variables using a custom one-hot encoding approach.

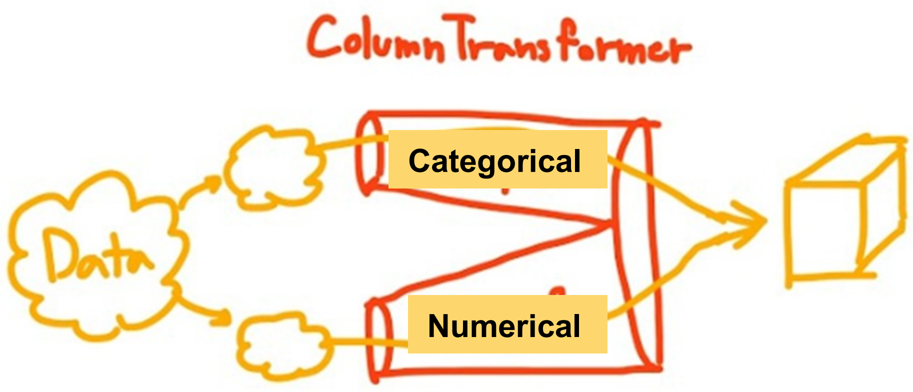


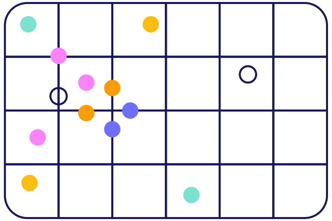
# Preprocessing , Cont'd

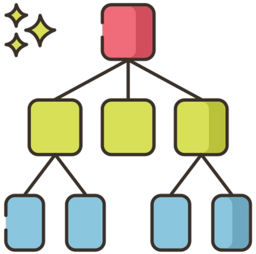
## 3. Column Transformation

# 3. Column Transformation:

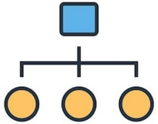
- **Numerical Columns Pipeline:**
  - Imputes missing values using median strategy.
- **Categorical Columns Pipeline:**
  - Imputes missing values using the most frequent strategy.
  - Applies custom one-hot encoding.

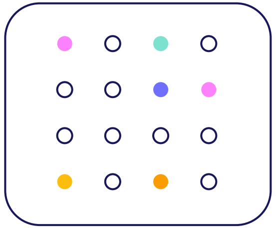






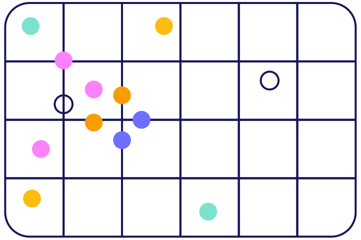
# Model

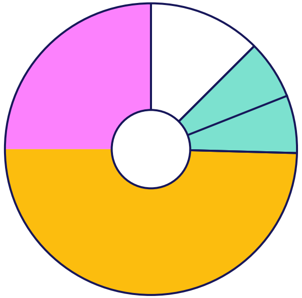




# Models:

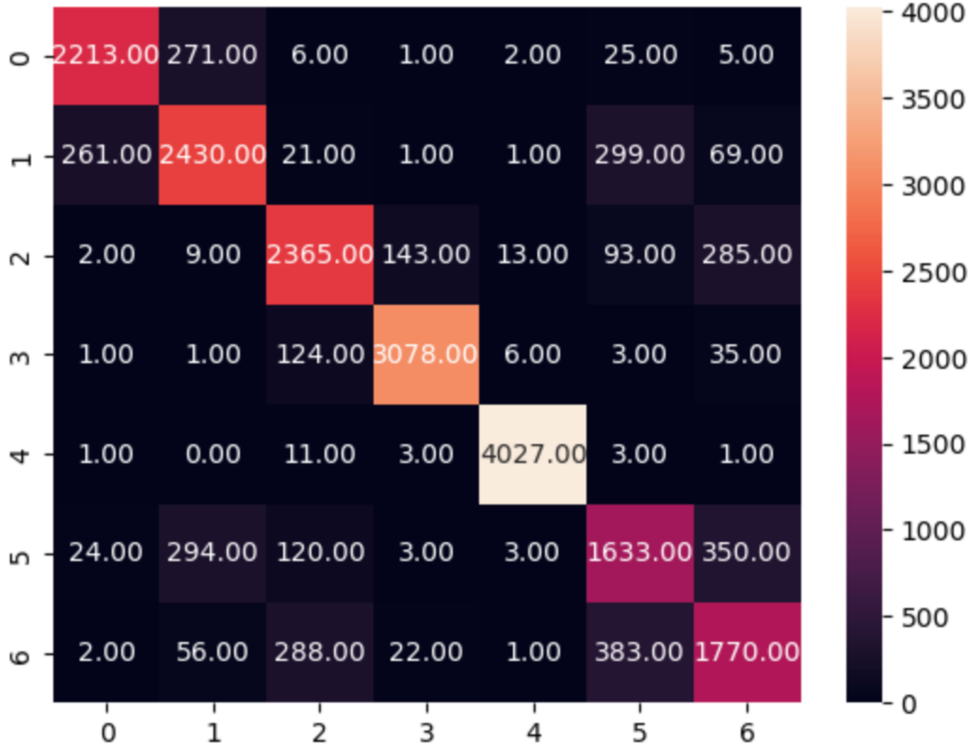
| Model | Validation Accuracy |
|---|---|
| Logistic Regression | 0.873 |
| SVM | 0.875 |
| Decision Tree | 0.845 |
| Random Forest | 0.890 |
| KNN | 0.837 |
| Adaboost | 0.878 |
| LGBM | 0.906 |
| Catboost | 0.902 |
| XGBoost | 0.903 |





# LGBM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.93 | 0.93 | 2535 |
| 1 | 0.89 | 0.87 | 0.88 | 3147 |
| 2 | 0.88 | 0.89 | 0.89 | 2876 |
| 3 | 0.97 | 0.97 | 0.97 | 3266 |
| 4 | 1.00 | 1.00 | 1.00 | 4046 |
| 5 | 0.77 | 0.80 | 0.79 | 2319 |
| 6 | 0.82 | 0.80 | 0.81 | 2569 |
| accuracy | | | 0.90 | 20758 |
| macro avg | 0.89 | 0.89 | 0.89 | 20758 |
| weighted avg | 0.91 | 0.90 | 0.90 | 20758 |

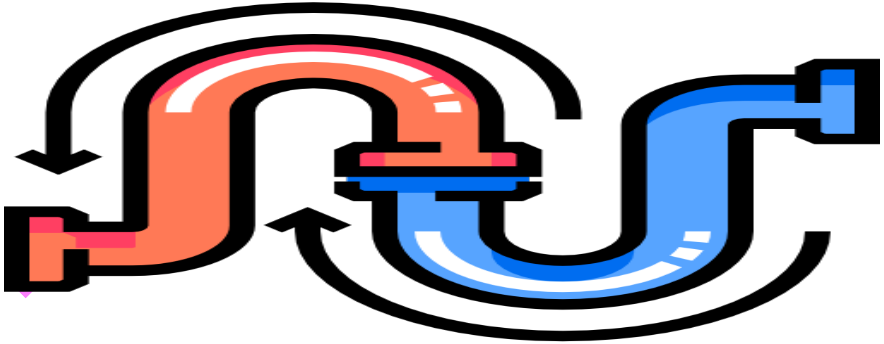


# Preprocessing, Cont'd

## 4. Full Pipeline

# Full Pipeline

- Features Extraction Pipeline: Performs feature engineering.
- Preprocessing Pipeline: Applies column transformation to numerical and categorical columns.
- StandardScaler: Standardizes features.
- LGBMClassifier: Fits an LGBM classifier using specified parameters (after tuning using optuna ).



# Resources

- https://www.kaggle.com/competitions/playground-series-s4e2
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9887184/
- https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

# Thanks!