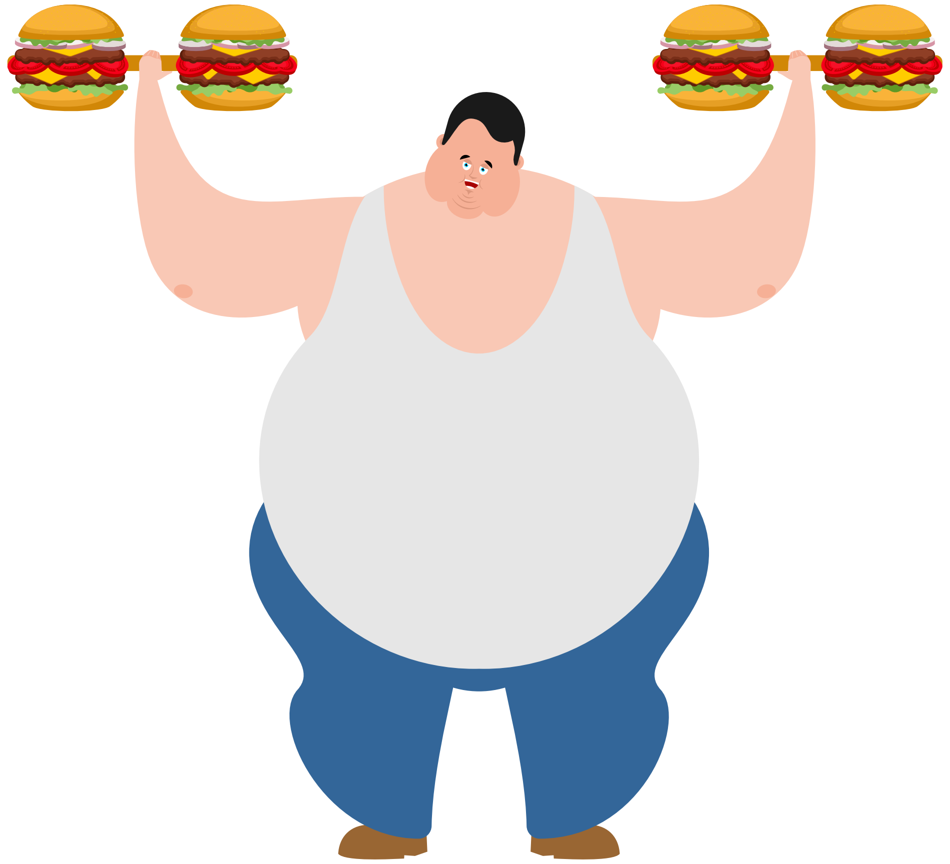
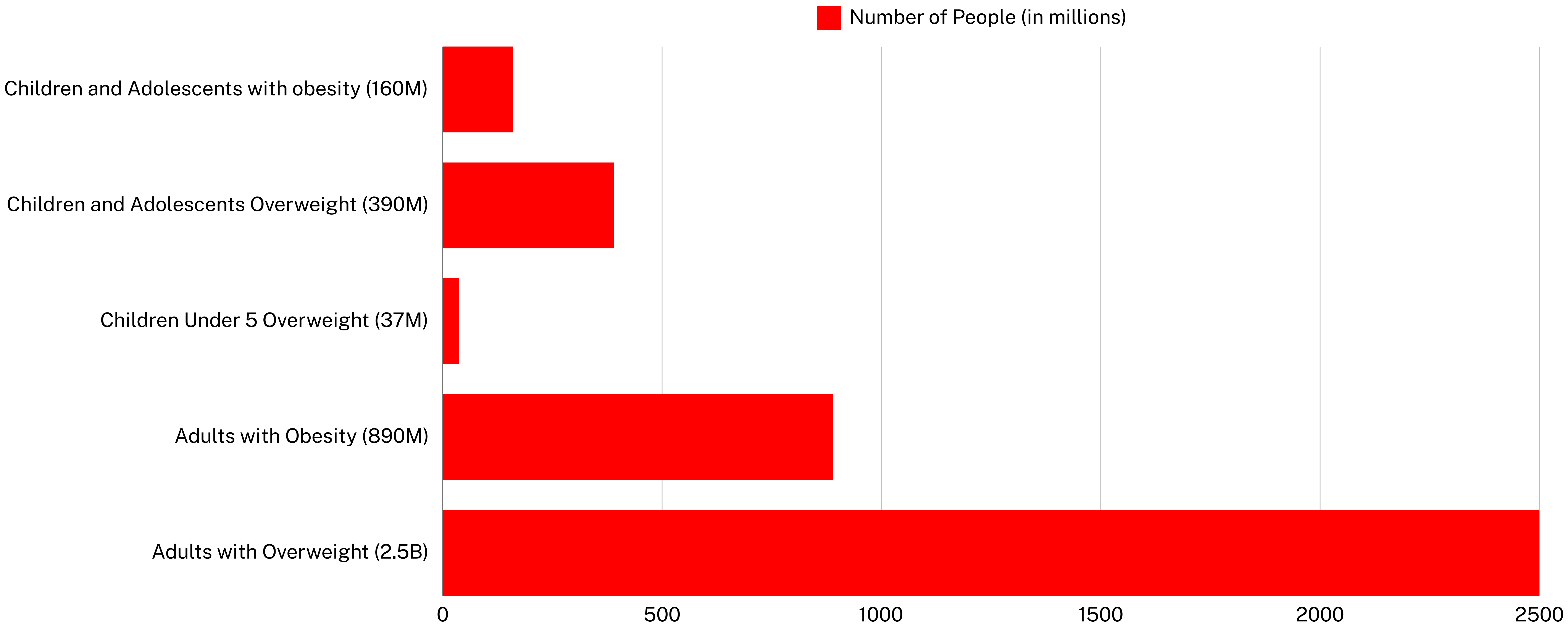
Predicting **Obesity** Levels by Linking **Personal Information** and **Lifestyle Factors**

Data Mining, Fall 2024
Team11



313553053 康峻瑋
313554059 林滋隆
313554046 許茗鈞
413551001 黃正鵬

BACKGROUND



Resource: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

MOTIVATION AND RESEARCH AIMS

Motivation:

a. early interventions and promoting healthier living habits

（提早干涉和提倡健康生活習慣）

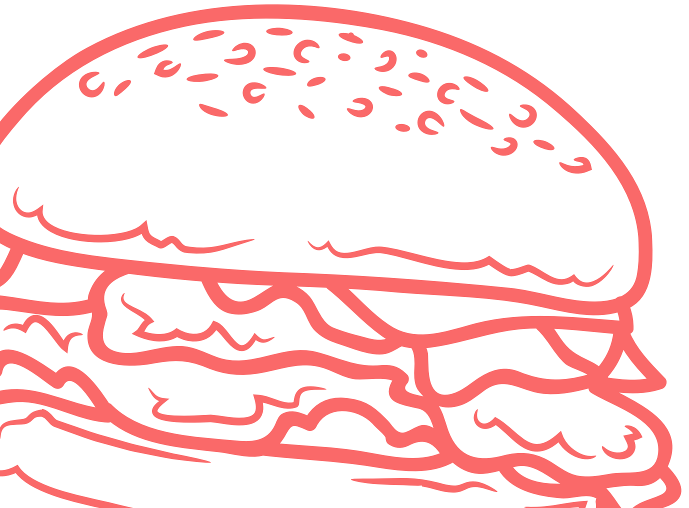
b. detecting causes of obesity.

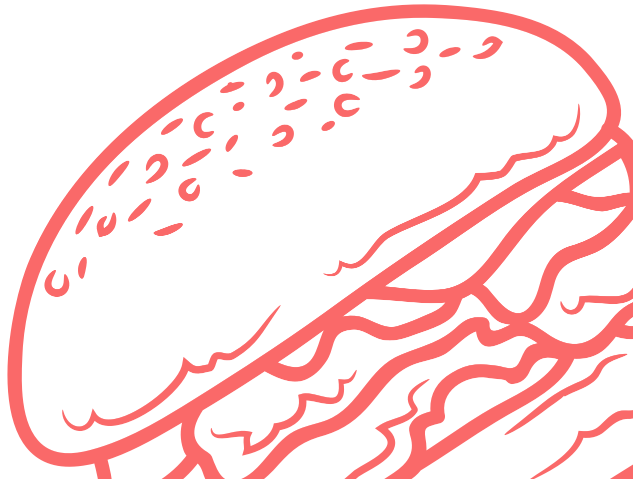
（觀測肥胖原因）

Research Aims:

1. data mining techniques → predict obesity levels

2. predictive models → high-risk individuals → provide actionable insights





PROBLEM DESCRIPTION

Predicting Obesity Levels by Linking Personal and Lifestyle Factors

（通過個人和生活方式因素的關聯來預測肥胖程度）

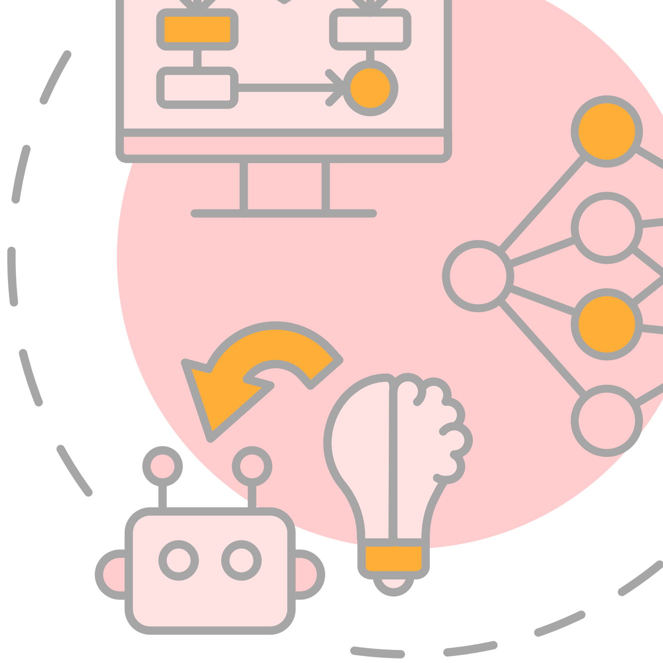
- **Input** ：資料集包含16個欄位，涵蓋個人因素（如年齡、身高、體重等）和生活方式因素（如每日餐數、吸菸狀態等）。

- **Process** ：1. 肥胖水平預測
  A. 使用分類模型（DECISION TREES, RANDOM FOREST, GRADIENT BOOSTING）來預測肥胖水平

  2. 將生活方式因素與肥胖水平聯繫
  A. 預測某些生活方式因素（例如HIGH FAST FOOD CONSUMPTION, LOW PHYSICAL ACTIVITY）導致肥胖的可能性（使用LOGISTIC REGRESSION, COLLABORATIVE FILTERING等方法）

- **Ouput** ：預測：WEIGHT LEVEL(4個等級)、OBESITY LEVEL(3種類型)
  關聯：生活方式與OBESITY LEVEL之間的「關聯」

TARGET PERFORMANCE



TARGET

- Accuracy 90% up
- AUC 80% up

Authors	Research Objective	Results
(Thamrin, Arsyad, et al., 2021)	Predicting obesity in adults	Accuracy: 72%, AUC: 79%
(Cheng et al., 2021)	Prediction of the effect of physical activity on obesity	Accuracy: 67% AUC: 64%
(Santisteban Quiroz, 2022)	Identifying obesity levels based on lifestyle through ML techniques	Accuracy: 97.45% AUC: 99.90%

Thamrin, Sri Astuti, et al. "Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018." Frontiers in nutrition 8 (2021): 669155.
Quiroz, Juan Piero Santisteban. "Estimation of obesity levels based on dietary habits and condition physical using computational intelligence." Informatics in Medicine Unlocked 29 (2022): 100901.
Cheng, Xiaolu, et al. "Does physical activity predict obesity — A machine learning and statistical method-based analysis." International Journal of environmental research and public Health 18.8 (2021): 3966.

DATA DESCRIPTION

16 + 1 CLASS VARIABLE **FEATURES**

INSTANCES 2111 RECORDS

Data acquired by : survey

ORIGINAL SOURCE:
UC IRVINE MACHINE LEARNING

Repository(Donated on 8/26/2019) https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

DOWNLOAD SOURCE:KAGGLE

https://www.kaggle.com/datasets/jayitabhattacharyya/estimation-of-obesity-levels-uci-dataset/

DATA DESCRIPTION (CONT.)

Variable Name	Role	Type	Range	Description	Missing Values
Gender	Feature	Categorical	2類		no
Age	Feature	Continuous	14~61歲		no
Height	Feature	Continuous	1.45~1.98公尺		no
Weight	Feature	Continuous	39~173公斤		no
family_history_with_overweight	Feature	Binary	TorF	Has a family member suffered or suffers from overweight?	no
FAVC	Feature	Binary	TorF	Do you eat high caloric food frequently?	no

FCVC	Feature	Integer	never(1)
			sometimes(2)
			always(3)
				Do you usually eat vegetables in your meals?	no

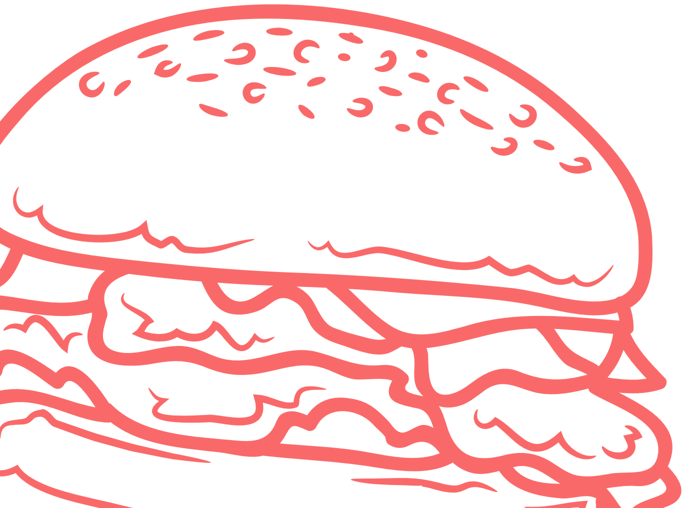
NCP	Feature	Continuous	1~4餐	How many main meals do you have daily?	no
CAEC	Feature	Categorical	Sometimes Frequently Other	Do you eat any food between meals?	no

DATA DESCRIPTION (CONT.)

SMOKE	Feature	Binary	TorF	Do you smoke?
CH2O	Feature	Continuous	1~3(L)	How much water do you drink daily?
SCC	Feature	Binary	TorF	Do you monitor the calories you eat daily?
FAF	Feature	Continuous	0~3天	How often do you have physical activity?
TUE	Feature	Integer	0~2小時	How much time do you use technological devices such as cell phone, videogames, television, computer and others?
CALC	Feature	Categorical	Sometimes no Other	How often do you drink alcohol?
MTRANS	Feature	Categorical	Public_Transportation Automobile Other	Which transportation do you usually use?
NObeyesdad	Target	Categorical	**Obesity_Type_I** **Obesity_Type_III** **Other**	Obesity level

ENVIRONMENT

作業系統	MacOS 14+
程式語言	Python
工具	Jupyter Notebook
函式庫	Scikit-learn



ANALYSIS WORKFLOW

1 **Data Processing**

- UNUSUAL DATA
- INCONSISTENT DATA
- OUTLIERS
- MISSING VALUES.

2 **Data Transformation**

- NUMERICAL REPRESENTATION
- SCALE BY NORMALIZATION/STANDARDIZATION

3 **Feature Engineering**

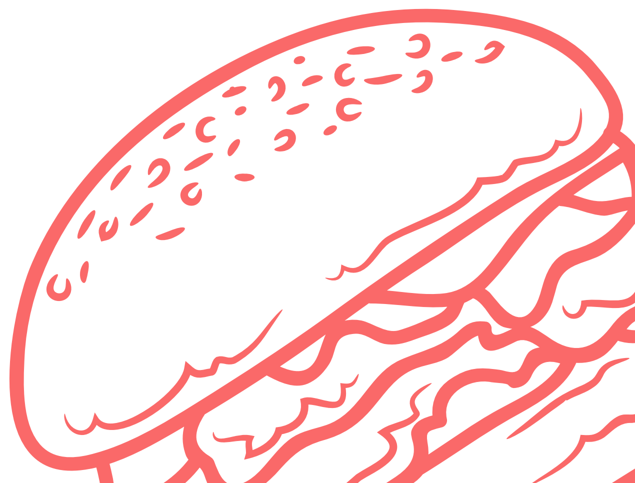
- BINNING
- FEATURE COMBINATION
- FEATURE SELECTION

4 **Modeling**

- CLASSIFICATION: DECISION TREES, RANDOM FOREST, GRADIENT BOOSTING (E.G., XGBOOST, LIGHTGBM, CATBOOST), SUPPORT VECTOR MACHINES (SVM), K-NEAREST NEIGHBORS (K-NN)
- LINK ANALYSIS: LOGISTIC REGRESSION, COLLABORATIVE FILTERING, MATRIX FACTORIZATION

5 **Evaluation**

- USE CROSS-VALIDATION TO ASSESS MODEL PERFORMANCE.
- CLASSIFICATION: ACCURACY, F1-SCORE, PRECISION, RECALL
- LINK PREDICTION: AUC-ROC



PROBLEM DESCRIPTION

1 Confusion Matrix

Helps identify how well the model predicts different obesity types or weight levels and highlights common errors.

2 AUROC

Measures the model's ability to classify obesity types or weight levels across all classes

3 Accuracy

The percentage of correct predictions for obesity types or weight levels, though it can be misleading with class imbalances.

4 Precision

Of those predicted to be a certain obesity type, how many are actually correct.

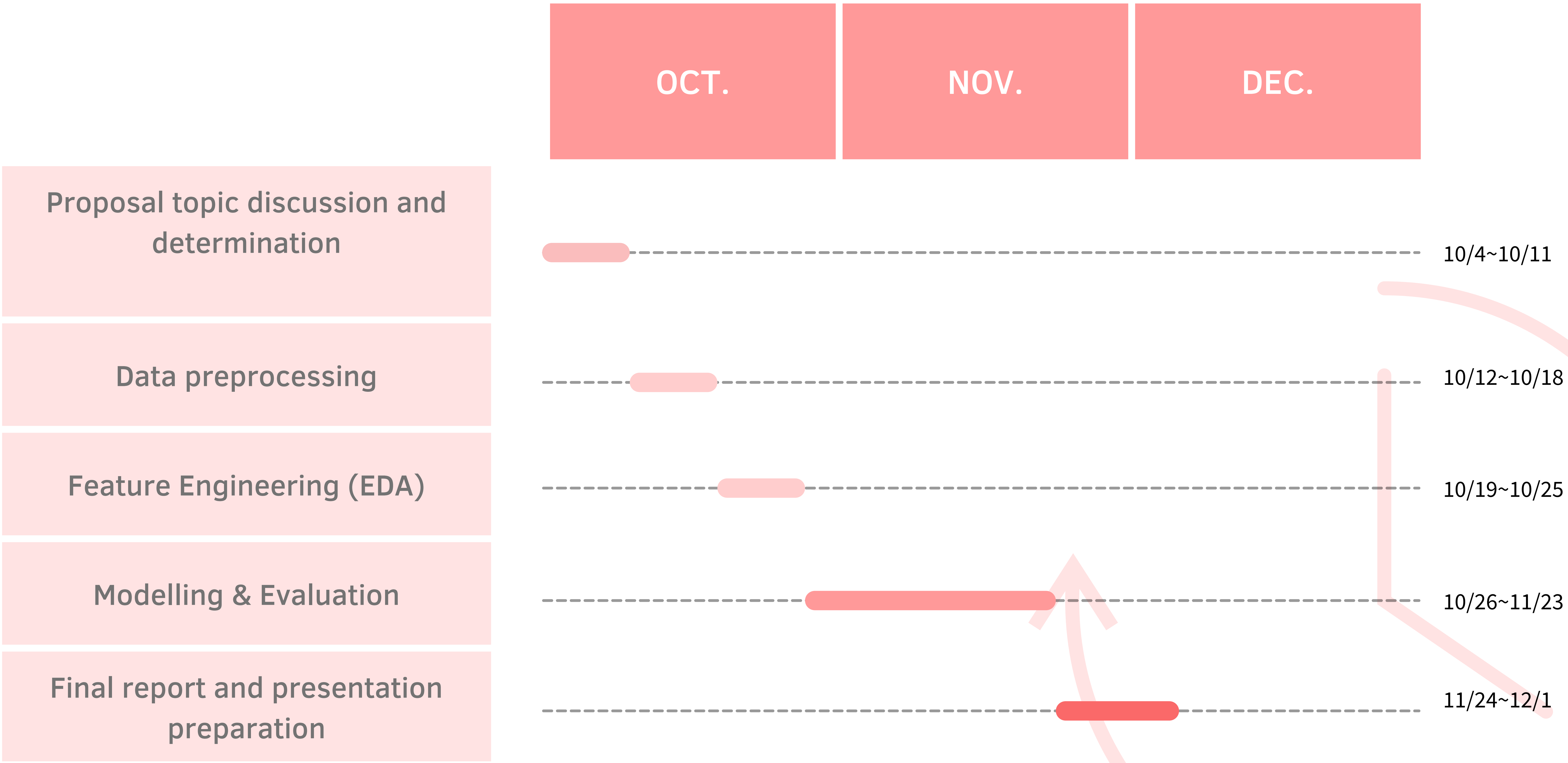
5 Recall

Of those who truly belong to a certain obesity type, how many are correctly identified.

6 F1 Score

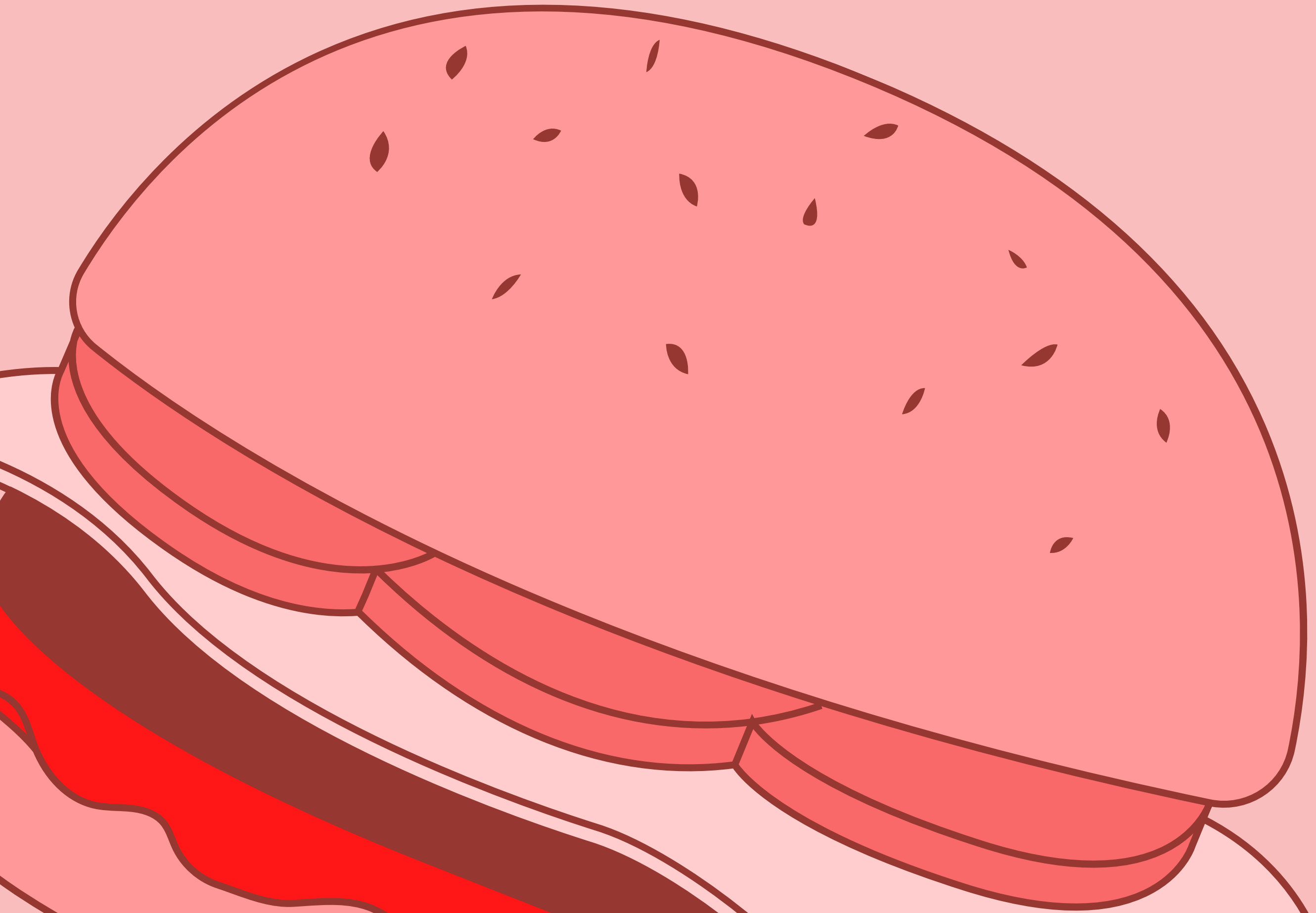
Balances precision and recall to evaluate overall prediction performance.

TENTATIVE SCHEDULE



Proposal topic discussion and determination	10/4~10/11
Data preprocessing	10/12~10/18
Feature Engineering (EDA)	10/19~10/25
Modelling & Evaluation	10/26~11/23
Final report and presentation preparation	11/24~12/1

THANK YOU



313553053	康峻瑋
313554059	林滋隆
313554046	許茗鈞
413551001	黃正鵬

TEAM11

313553053	康峻瑋
313554059	林滋隆
313554046	許茗鈞
413551001	黃正鵬

TEAM11