kaggle competition 2024

Multi-Class Prediction

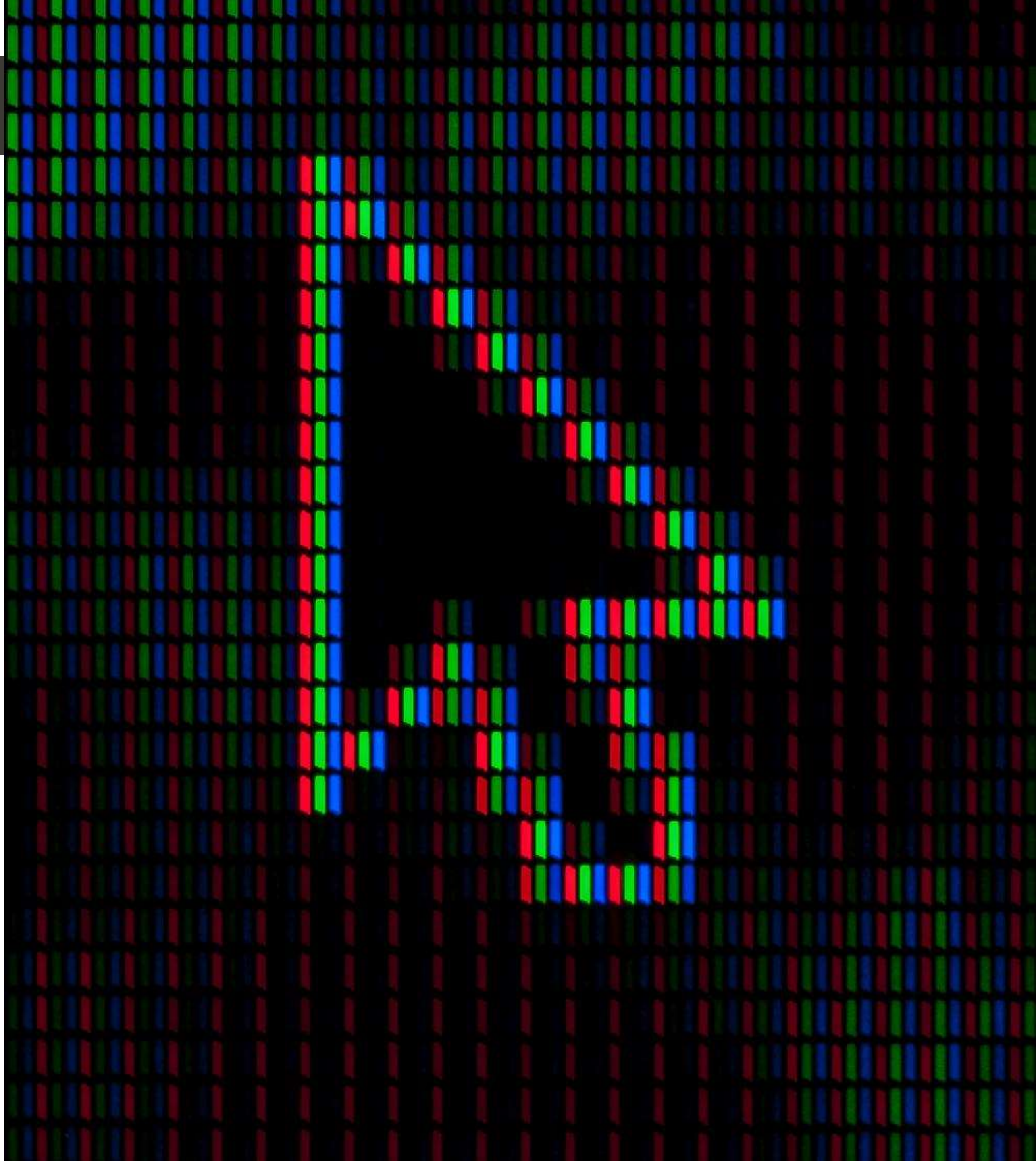
Obesity Risk

Lim Yewon | Yoon Jihoon | Weon Joosung

Role

- Data analysis & Model training:
  Lim Yewon, Yoon Jihoon, Weon Joosung
- Documentation: Lim Yewon
- Presentation: Lim Yewon, Yoon Jihoon

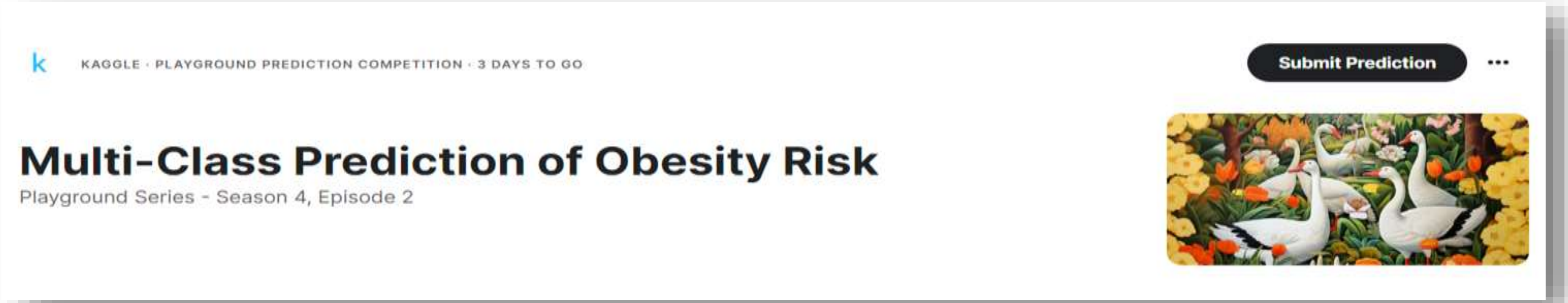
Table of Contents



Part 1 INTRODUCTION

ⓒSaebyeol Yu. Saebyeol's PowerPoint

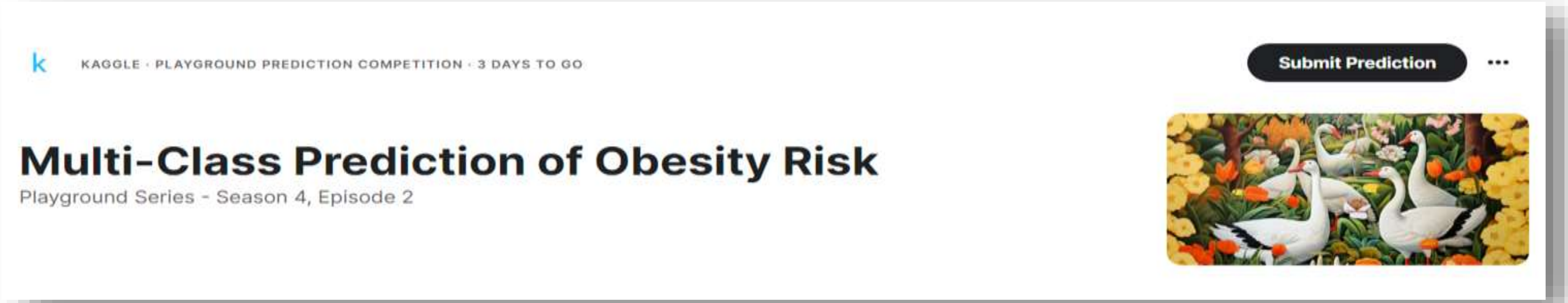
Overview of Competition



Competition Timeline	Feb 1, 2024 ~ Mar 1, 2024
Final Submission Deadline	Feb 29, 2024
Duration of participation	5 days (Feb 26 2024 ~ Mar 1 2024)
Evaluation	Submissions are evaluated using the accuracy score.
Kaggle Notebook	https://www.kaggle.com/competitions/playground-series-s4e2/overview

ⓒSaebyeol Yu. Saebyeol's PowerPoint

01 Overview of Competition

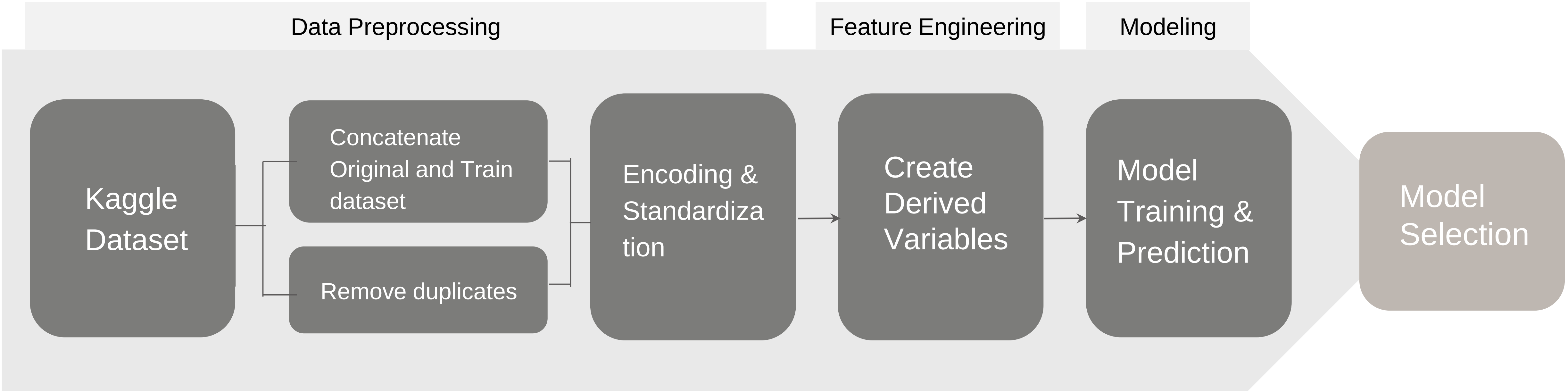


Goal:

The goal of this competition is to use various factors to predict obesity risk in individuals, which is related to cardiovascular disease.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

02 Process Flow



ⓒSaebyeol Yu. Saebyeol's PowerPoint

Introduction of Dataset

Train

(20758, 18)

Test

(13840, 17)

Original

(2111, 17)

Submission

(13840, 2)

ⓒSaebyeol Yu. Saebyeol's PowerPoint

03

Introduction of Dataset

◆ **Column Description**

- **'id'** : id
- **'Gender'** : Gender
- **'Age'** : Age
- **'Height'** : Height is in meter
- **'Weight'** : Weight is between 39 to 165
- **'family_history_with_overweight'** : family history with overweight yes or no
- **'FAVC'** : Frequent consumption of high calorie food yes or no
- **'FCVC'** : Frequency of consumption of vegetables yes or no
- **'NCP'** : Number of main meals
- **'CAEC'** : Consumption of food between meals
- **'SMOKE'** : yes or no
- **'CH2O'** : Consumption of water daily
- **'SCC'** : Calories consumption monitoring yes or no
- **'FAF'** : Physical activity frequency
- **'TUE'** : Time using technology devices "How long using technology devices to track your health"
- **'CALC'** : Consumption of alcohol
- **'MTRANS'** : Transportation used
- **'NObeyesdad'** : Target Obesity

ⓒSaebyeol Yu. Saebyeol's PowerPoint

03

Introduction of Dataset

Continuous Variables

- Age
- Height
- Weight
- FCVC
- NCP
- CH2O
- FAF
- TUE

= total 8

16 features in total

Categorical Variables

- Gender
- family_history_with_overweight
- FAVC
- CAEC
- SMOKE
- SCC
- CALC
- MTRANS

= total 8

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 2　Data Processing & EDA

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Data Preprocessing

Independent variables

Continuous variables
StandardScaler()

Categorical variables
OneHotEncoder()

Column Transformer →

Standardization

Dummy variables

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Data Preprocessing

Dependent Variable

Categorical Variable

'NObeyesdad'

Obesity_Type_III
Obesity_Type_II
Obesity_Type_I
Overweight_Level_II
Overweight_Level_I
Normal_Weight
Insufficient_Weight

Label Encoder →

Numeric Variable

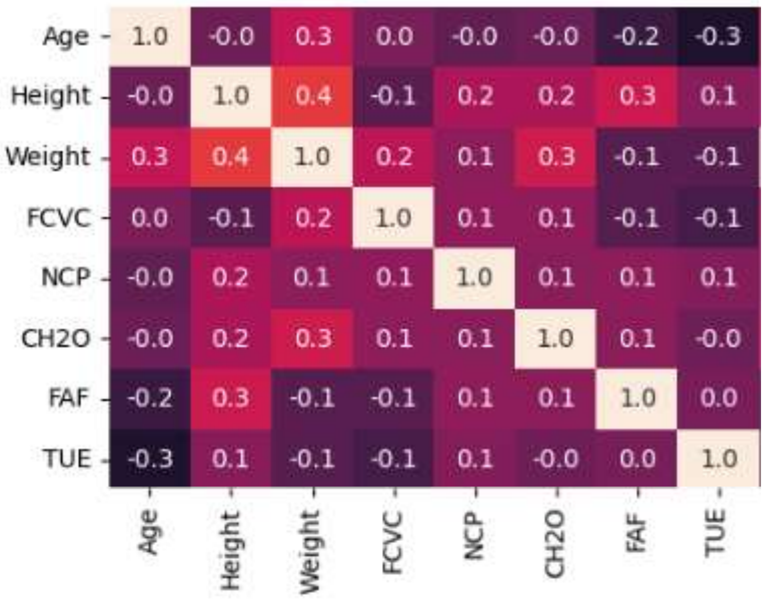
'NObeyesdad'

0
1
2
3
4
5
6

ⓒSaebyeol Yu. Saebyeol's PowerPoint

02 Feature Engineering

□ There is no feature with strong correlation
- we can avoid multicollinearity and proceed with feature engineering
- to better understand and classify the risk of obesity



Feature engineering →

BMI : Weight / Height
Tech Usage Score : TUE / Age
Meal Habits : FCVC / NCP

ⓒSaebyeol Yu. Saebyeol's PowerPoint

02 Feature Engineering

□ These features capture nuanced aspects of physical health, lifestyle choices, and nutritional habits.

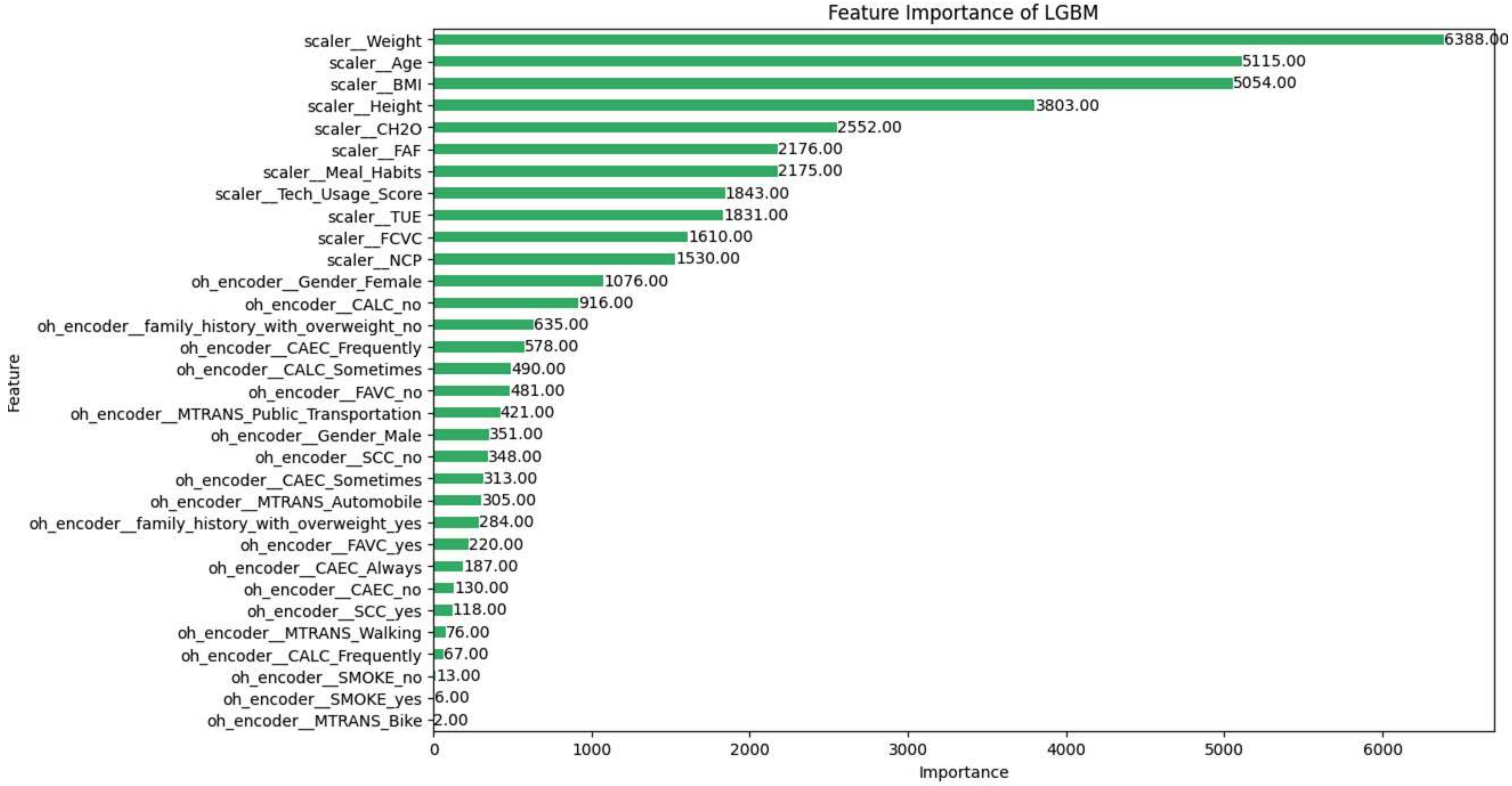
Derived feature	Description
Body Mass Index (BMI)	- The first step was to calculate BMI using 'Height' and 'Weight', a key metric for indicating obesity by accurately showing the weight-height relationship.
Meal Habits	- The 'Meal_Habits' feature combines 'FCVC' (Frequency of consumption of vegetables) and 'NCP' (Number of main meals), capturing overall dietary patterns by considering both variables.
Tech Usage Score	- 'Tech_Usage_Score' weights the frequency of technology usage ('TUE') by the individual's age, offering a nuanced perspective on technology habits by quantifying average time spent using technology relative to age.

Features above were created by Luca Massaron

ⓒSaebyeol Yu. Saebyeol's PowerPoint

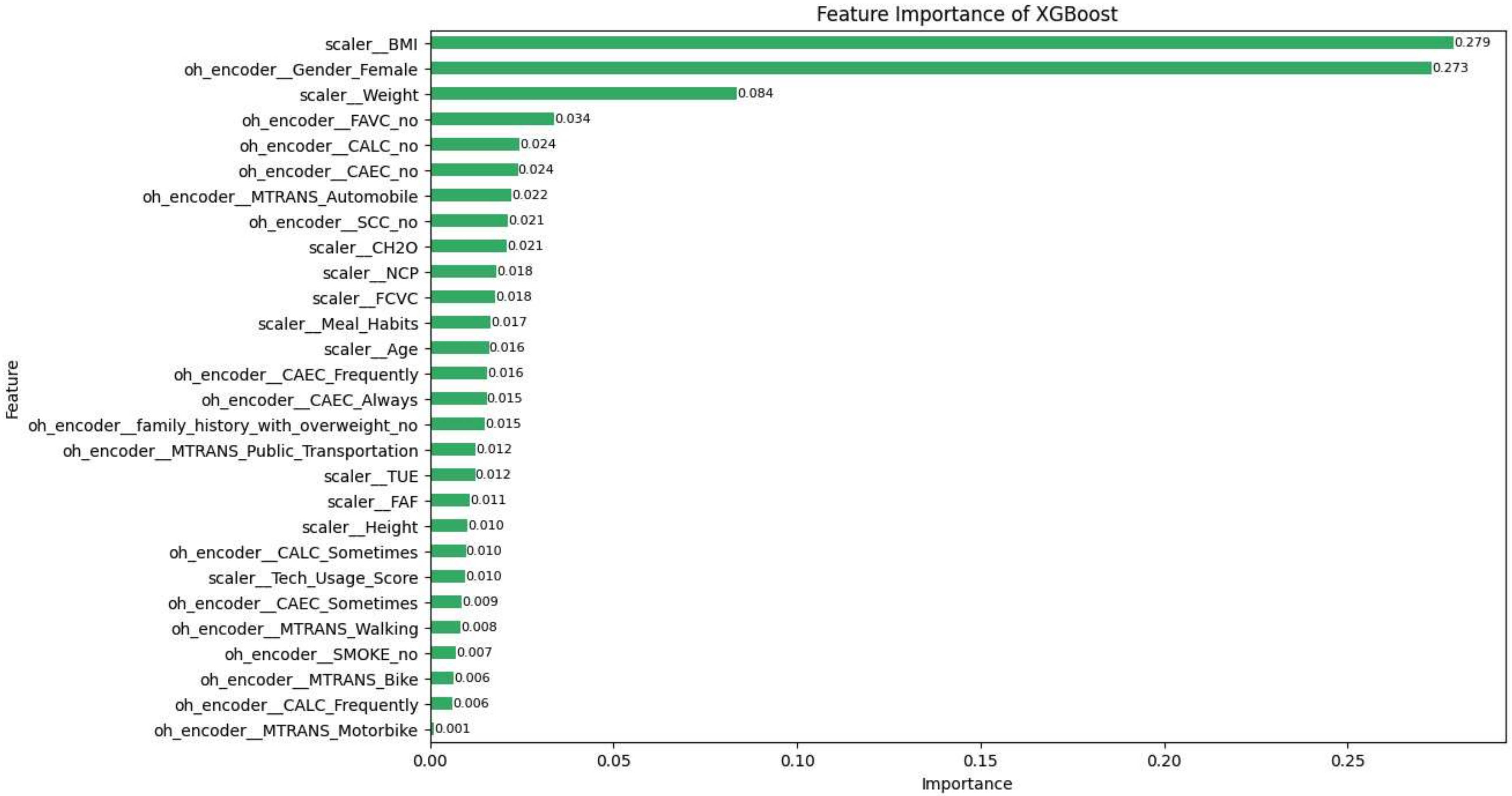
Feature Importance

- Different models are based on different learning algorithms
  - Each algorithm operates differently based on the characteristics of the data.



ⓒSaebyeol Yu. Saebyeol's PowerPoint

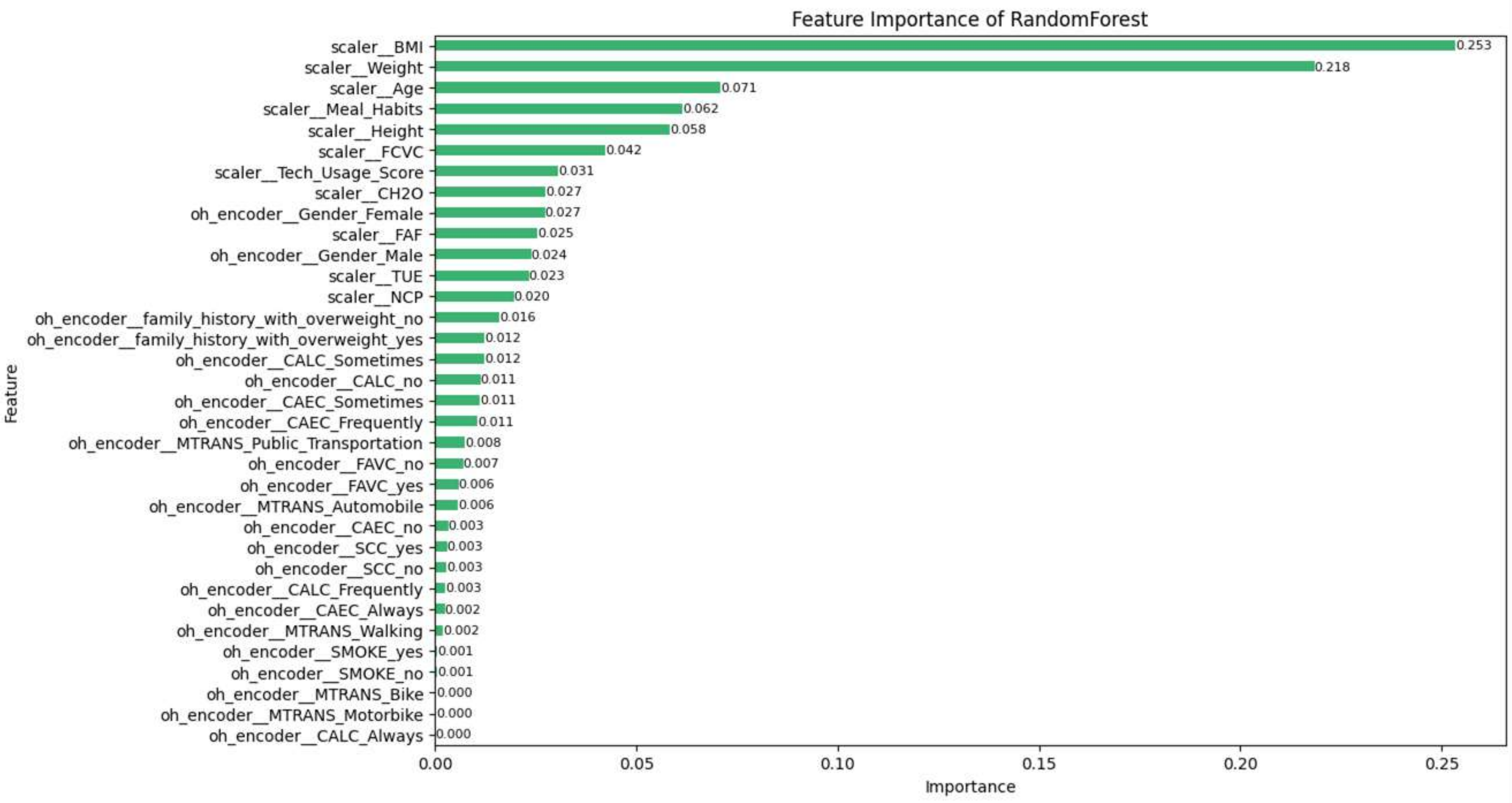
03 Feature Importance



ⓒSaebyeol Yu. Saebyeol's PowerPoint

03

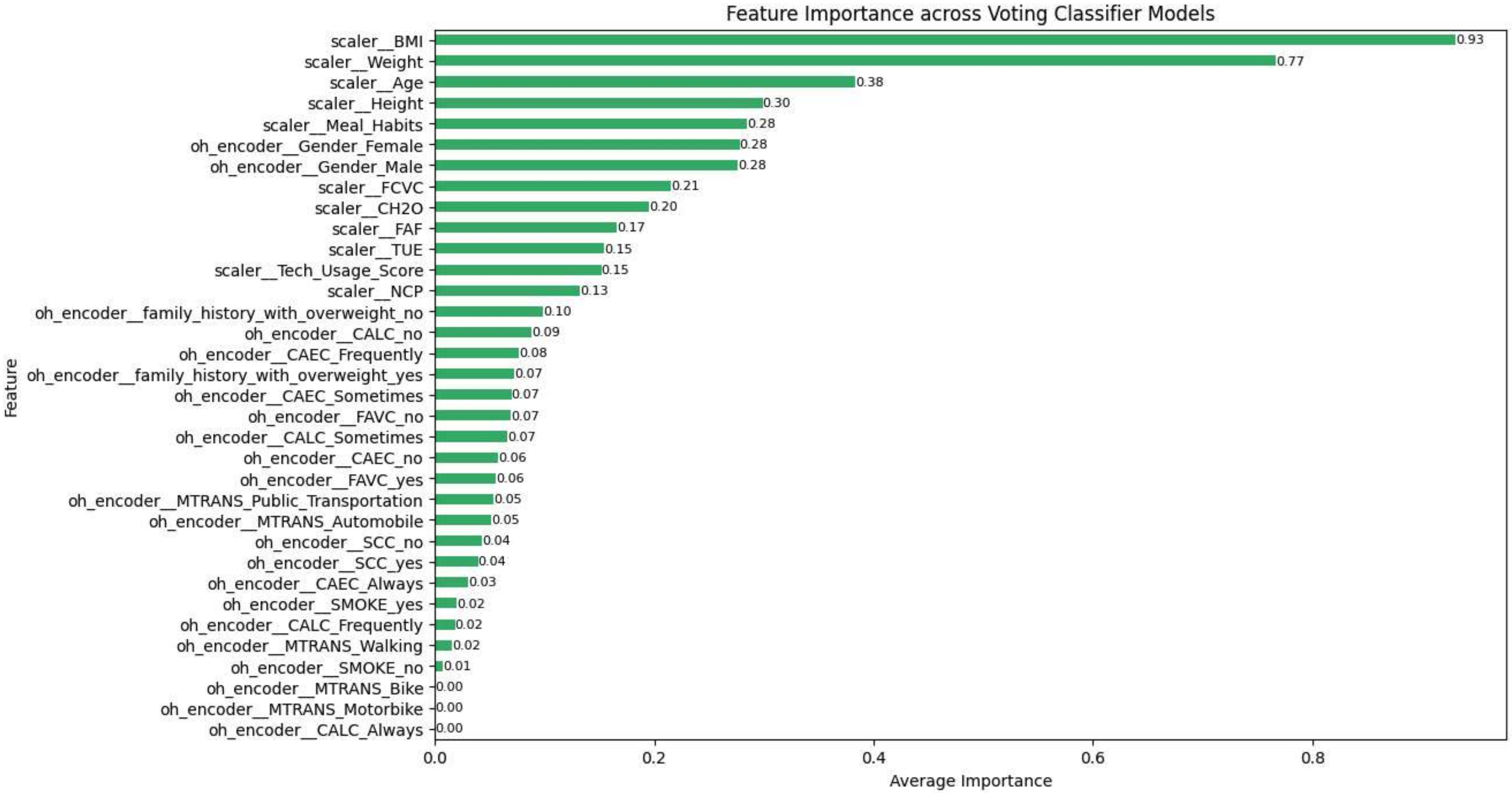
Feature Importance



ⓒSaebyeol Yu. Saebyeol's PowerPoint

03

Feature Importance

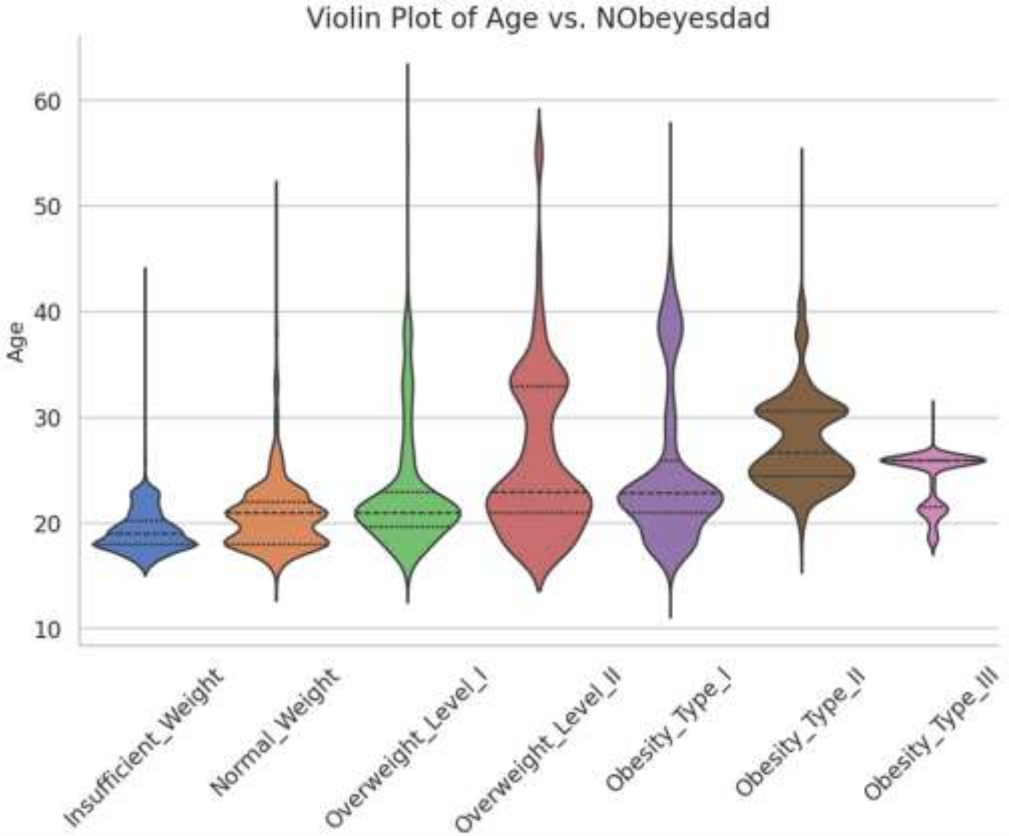


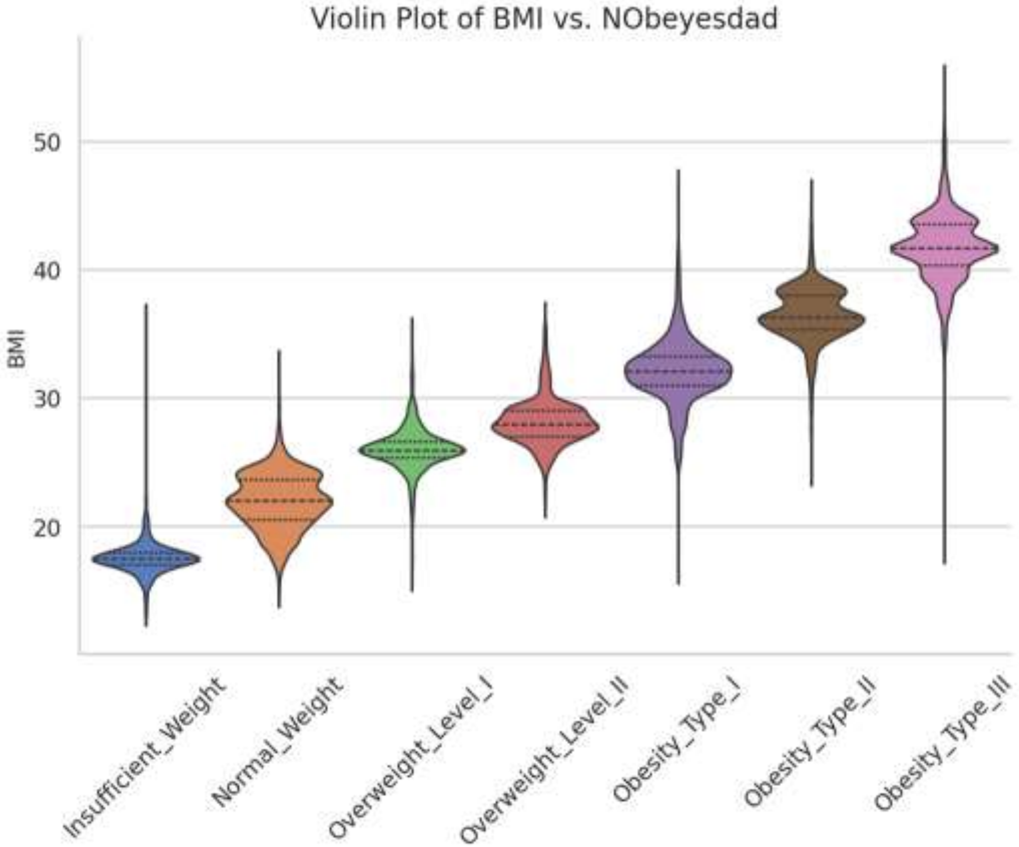
ⓒSaebyeol Yu. Saebyeol's PowerPoint

04 EDA with Important Features

 Top 5 important features from voting classifier were taken for visualization.

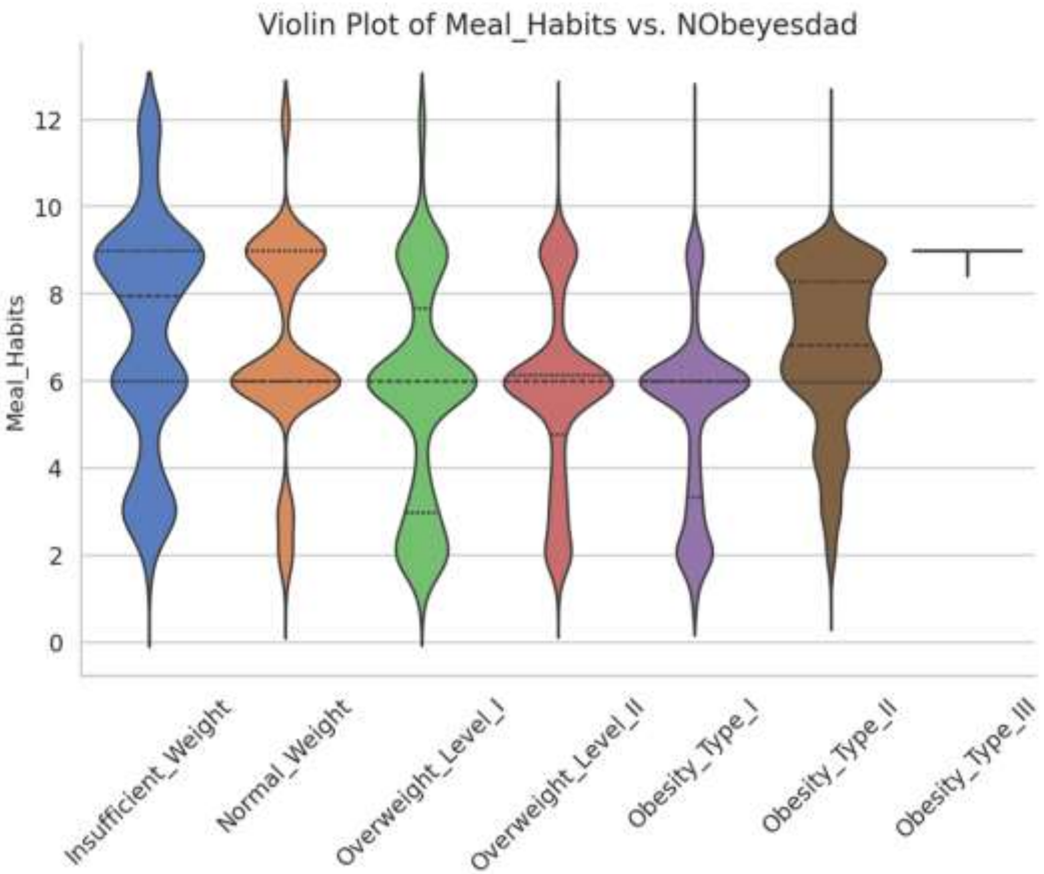
- BMI, age, meal_habits, CH2O, FAF

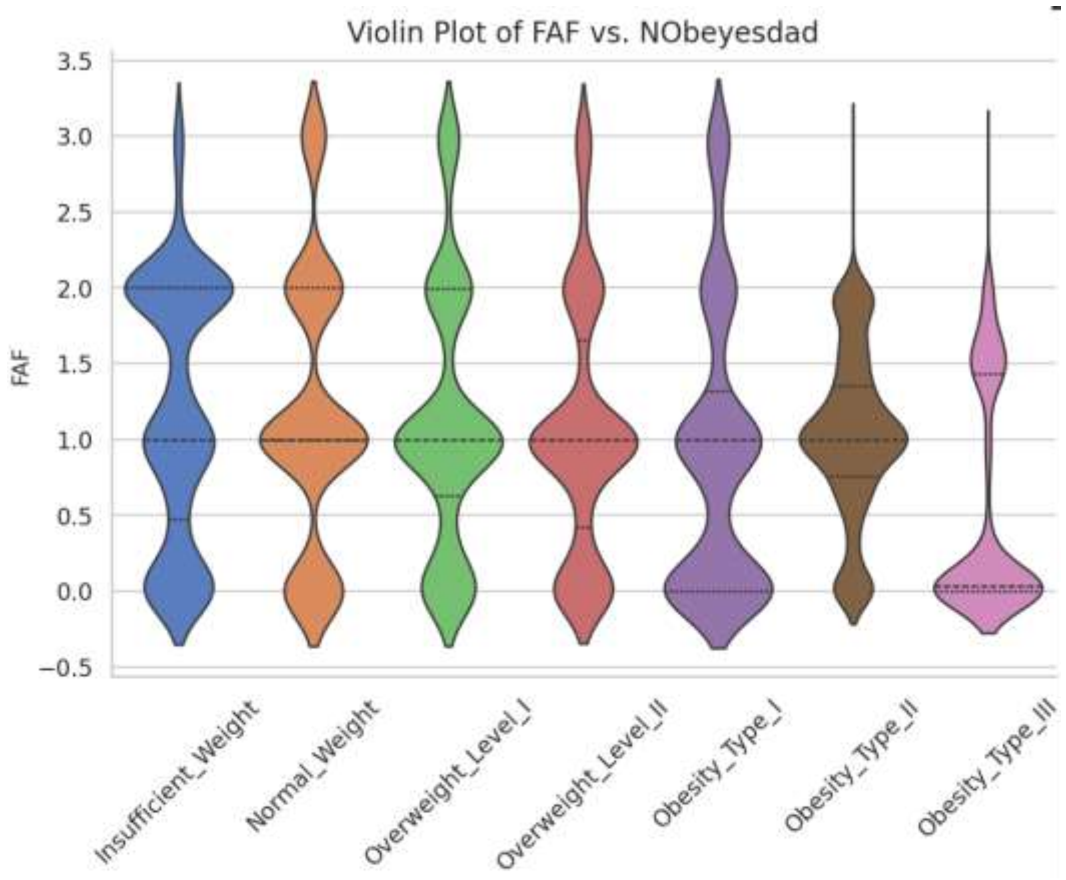




ⓒSaebyeol Yu. Saebyeol's PowerPoint

EDA with Important Features



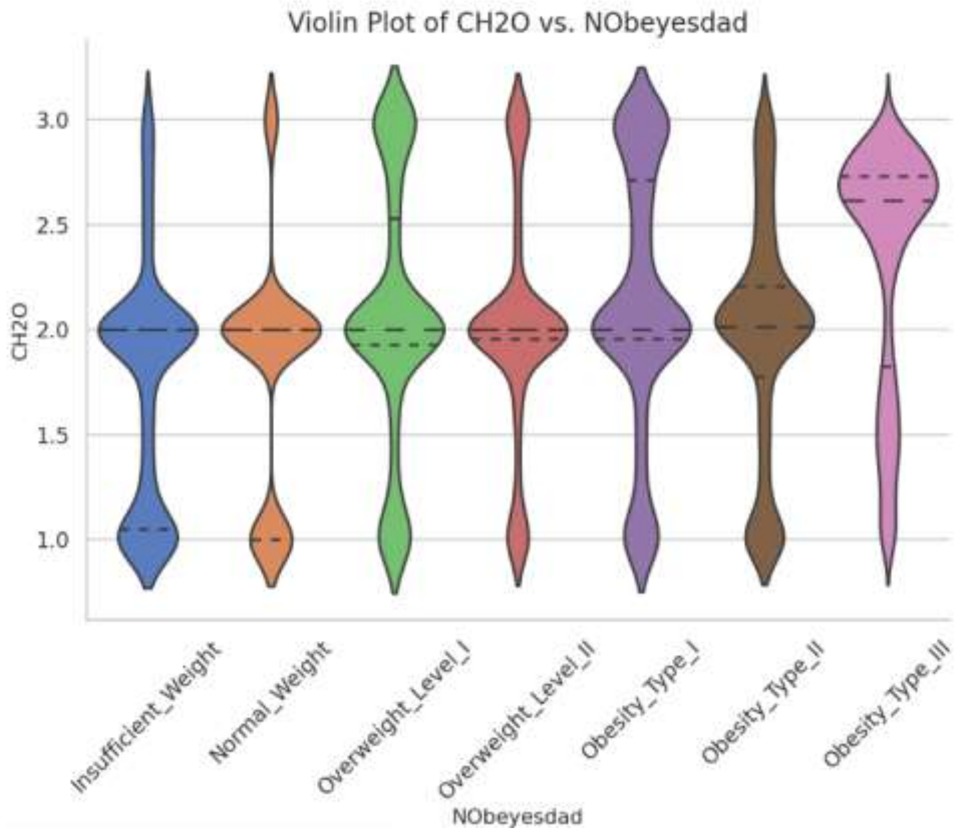


ⓒSaebyeol Yu. Saebyeol's PowerPoint

04

EDA with Important Features

- In CH2O, there's not much significant visible difference between groups.



ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 3 MODEL SELECTION

ⓒSaebyeol Yu. Saebyeol's PowerPoint

01 Pipeline

Gradient Boosting

LGBM

XGBoost

RandomForest

- We compared 4 baseline models to see which one shows the best performance.

- We followed the recommended pipeline method to prevent data leakage.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

01 Pipeline

Gradient Boosting

LGBM

XGBoost

RandomForest

- Utilizes a leaf-centered tree growth method, prioritizing leaf creation to minimize data loss
- Enables fast learning speed and efficient memory usage, suitable for high-dimensional and large-scale datasets
- Particularly effective for processing large amounts of data quickly due to its lightweight design

ⓒSaebyeol Yu. Saebyeol's PowerPoint

01

Pipeline

Gradient Boosting

LGBM

XGBoost

RandomForest

- Provides high speed and efficiency while maximizing prediction accuracy
- Includes a regularization function to prevent overfitting, ensuring robust performance on complex datasets
- Widely used in machine learning due to its optimization for processing large datasets and excellent prediction performance

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Key Factors

Feature Engineering

Parameter Tuning

Target Encoding

Cross-validation

ⓒSaebyeol Yu. Saebyeol's PowerPoint

03 LightGBM

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private score
LightGBM	X	X	X	X	0.91378	0.91401	0.90435
LightGBM	X	X	X	O	0.91308	0.91293	0.90354
LightGBM	O	O	X	O	0.907	0.9104	0.90408
LightGBM	O	O	O	O	0.935	0.9057	0.90534

- Based on the gap between public and private score, we can see that conducting cross-validation 5 times significantly reduced overfitting.

- Conducting target encoding and feature engineering helps improve the performance of the model.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

04 XGBoost

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private Score
XGBoost	O	O	X	X	0.909	0.91401	0.90245
XGBoost	O	O	X	O	0.907	-	-
XGBoost	O	O	O	O	0.937	0.90859	0.90652

- XGBoost showed better performance than LGBM under the same condition.

- Using the model with highest private score, we conducted further experiment to observe the impact of including derived variables in the model and optimizing the number of boosting iterations.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

04 XGBoost

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private Score
XGBoost	O	O	O	O(iter 1224)	0.937	0.90859	0.90652
XGBoost	O	O	O	O(iter 612)	0.914	0.90751	0.90706
XGBoost	O	X	O	O(iter 612)	0.909	0.91437	0.90697
XGBoost	O	X	O	O(iter 1224)	0.916	0.91437	0.90679

- Reducing the number of boosting iteration(n_estimators) may decrease the accuracy. However, this helped prevent overfitting and resulted in improved performance in both public score and private score.

- Also, including derived variables showed slightly better performance than the model without them.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

05 Final Model Selection

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private score
Soft Voting (lgbm, xgboost, randomforest)	O	X	O	O (iter 612)	0.9662	0.91221	0.90742
Soft Voting (lgbm, xgboost, randomforest)	O	O	O	O (iter 612)	0.9669	0.90823	0.90606

- We compared voting classifiers with reduced n-estimators for xgb model - one including derived variables and one that didn't.

- Although in the latter the gap between public and private score was lower which means less overfitting, private score was the highest in the former.

- However, overfitting was detected in voting model.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

06 Kaggle Ranking

➤ Final submission: 1580/3746 (Top 42%)

- we mistakenly chose this model due to its high public score.

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private score
LightGBM	X	X	X	X	0.91378	**0.91401**	**0.90435**

➤ Late submission: est. Top **18~19%**

- we went back to our previous model and made late submission.

Algorithm	Target encoding	Derived variables	Stratified K-fold	Hyperparameter tuning	Accuracy	Public score	Private score
Soft Voting (lgbm, xgboost, randomforest)	O	X	O	O (iter 612)	0.9662	**0.91221**	**0.90742**

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 4 Conclusion

ⓒSaebyeol Yu. Saebyeol's PowerPoint

01 Conclusion

Data Analysis	- In regards to feature importance, BMI, age, Meal_Habits, CH2O, FAF and such demonstrate high importance across three models. The features influencing performance vary depending on the structure and algorithm of the model.
Model Training	- Cross-validation and hyperparameter optimization is essential. - Choosing ideal number of cross-validation folds and boosting iterations is needed to prevent overfitting and improve model's generalization performance.
Feature Engineering	- Incorporating derived variables showed the best performance when using a single XGBoost model. However, when employing soft voting, the result was the opposite. - Reason for this remains unanswered, necessitating further study.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

02 Limitations

1	- Limitation in sorting out the best combination of hyperparameters
2	- Limitation in figuring out why incorporating derived variables showed different performance in different models.
3	- Limitation in identifying ways to reduce overfitting.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Thank you.