

Kaggle 다중 분류 모델 대회


정소영

데이터 분석&엔지니어(Python) 34회차

목차

1. 대회 소개
2. 탐색적 자료분석
3. 머신러닝 주요 알고리즘 소개
4. 모델 평가
5. 결론 및 리뷰

1. 대회 소개

 KAGGLE · PLAYGROUND PREDICTION COMPETITION · 3 DAYS TO GO

Submit Prediction ...

Multi-Class Prediction of Obesity Risk


Playground Series - Season 4, Episode 2

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Overview

Welcome to the 2024 Kaggle Playground Series! We plan to continue in the spirit of previous playgrounds, providing interesting and approachable datasets for our community to practice their machine learning skills, and anticipate a competition each month.

Your Goal: The goal of this competition is to use various factors to predict obesity risk in individuals, which is related to cardiovascular disease. Good luck!

Competition Host
Kaggle 

Prizes & Awards
Swag
Does not award Points or Medals

Participation



1. 대회 소개

◆ 대회 개요

대회 타임라인	February 1, 2024~February 29, 2024, 11:59 PM UTC
상금	Choice of Kaggle merchandise
참여 기간	7 days(February 23, 2024~February 29, 2024)
참여자 수	1
등수	2038/3587
Kaggle Notebook	https://www.kaggle.com/code/jeosoyoung/obesity-risk?scriptVersionId=164855698

1. 대회 소개

◆ 데이터셋 소개

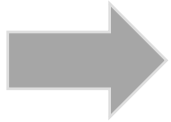
Data			description			Data			description		
Id			id		아이디	CAEC			Consumption of food between meals		식사 사이의 음식 섭취
Gender			Gender		성별	SMOKE			Smoke		흡연
Age			Age		나이	CH2O			Consumption of water daily		하루 물 소비
Height			Height is in meter		키(미터)	SCC			Calories consumption monitoring		열량소비량 모니터링
Weight			Weight is between 39 to 165		몸무게(39~165)	FAF			Physical activity frequency		신체활동 빈도
family_history_with_overweight			family history with overweight		과체중 가족력	TUE			Time using technology devices		전자기기 사용시간
FAVC			Frequent consumption of high calorie food		고칼로리 음식을 자주 소비하는가	CALC			Consumption of alcohol		알코올 소비
FCVC			Frequency of consumption of vegetables		야채 섭취 빈도	MTRANS			Transportation used		이동수단 사용
NCP			Number of main meals		주된 식사 횟수	NObesidad			(Target) Obesity		비만

1. 대회 소개

◆ 평가지표

Evaluation

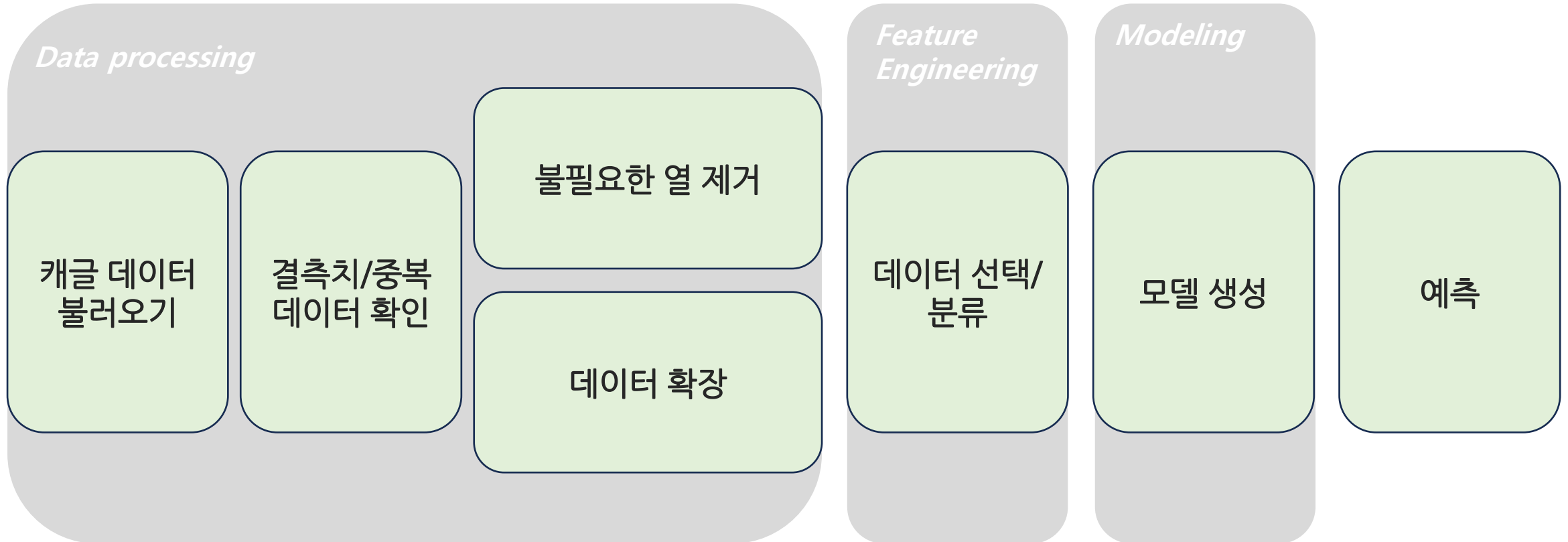
Submissions are evaluated using the accuracy score.



평가지표 : **정확도** 점수

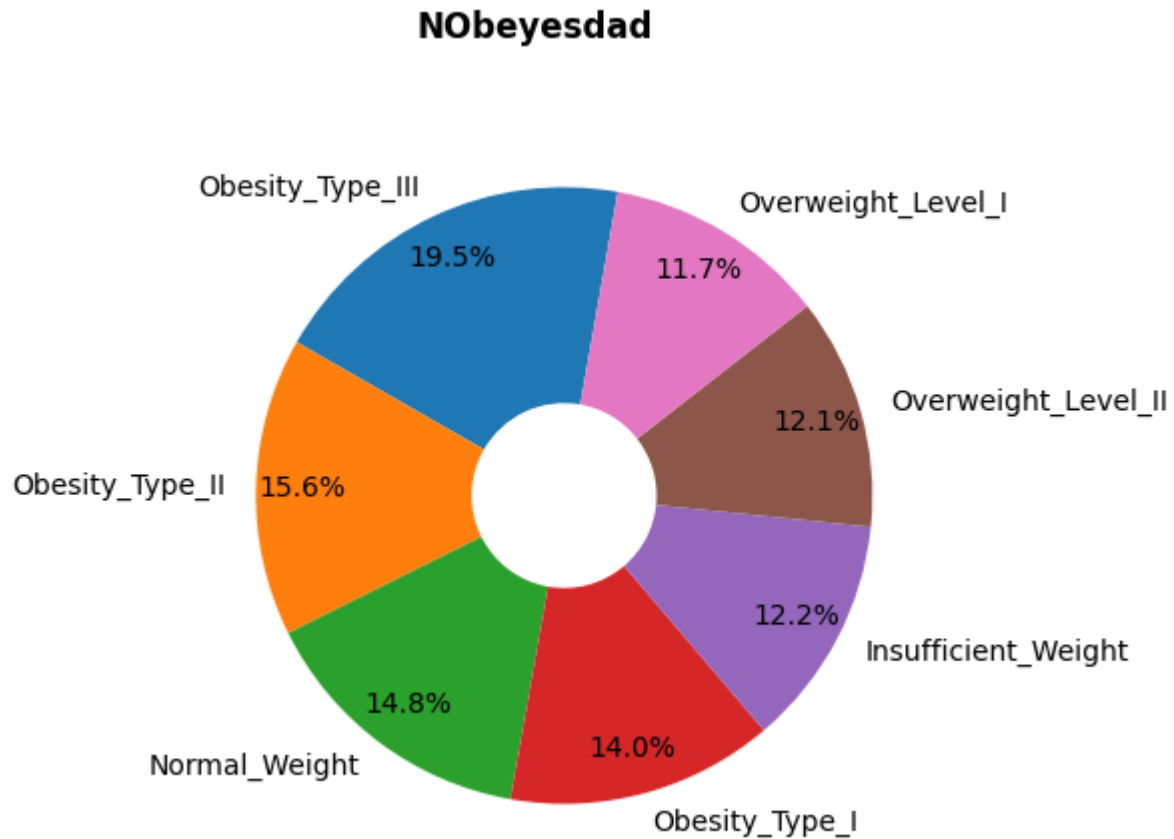
1. 대회 소개

◆ 모델링 프로세스



2. 탐색적 자료 분석

◆ 시각화



```
NObeyesdad
Obesity_Type_III    4046
Obesity_Type_II     3248
Normal_Weight       3082
Obesity_Type_I      2910
Insufficient_Weight 2523
Overweight_Level_II 2522
Overweight_Level_I  2427
Name: count, dtype: int64
```

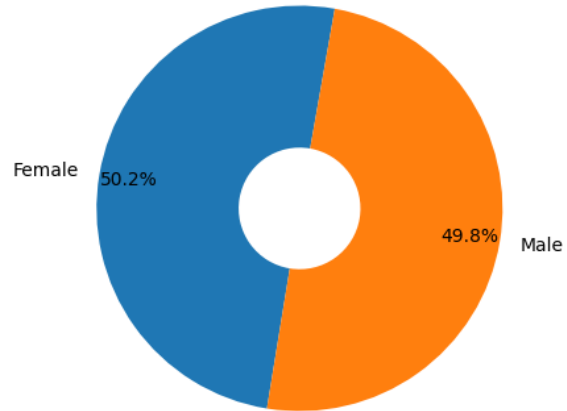
Nobeyesdad
(타겟 데이터)

-> 가장 많은 사람들이 분포된 유형:
Obesity_Type_III(19.5%)

2. 탐색적 자료 분석

◆ 시각화

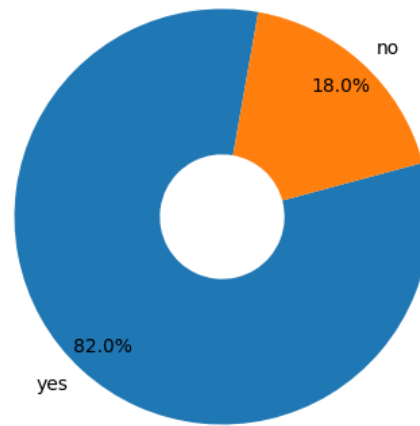
Gender



```
Gender
Female    10422
Male      10336
Name: count, dtype: int64
```

Gender
:남녀가 비슷한 분포를 보임

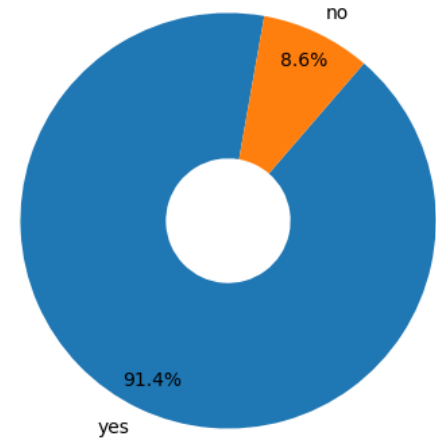
family_history_with_overweight



```
family_history_with_overweight
yes      17014
no        3744
Name: count, dtype: int64
```

Family_history_with_overweight
: 82.0%의 사람들이 과체중 가족력을 가짐

FAVC

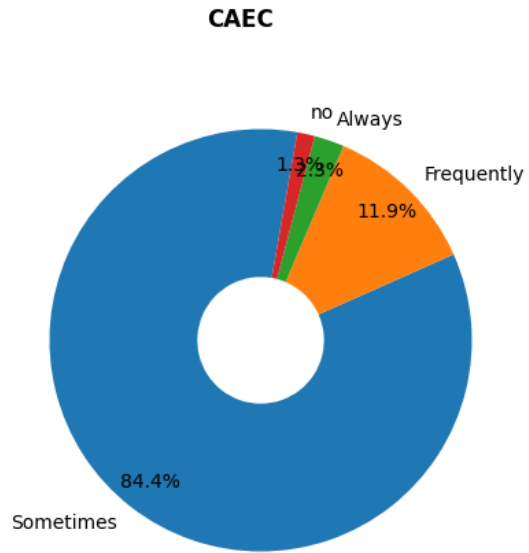


```
FAVC
yes    18982
no      1776
Name: count, dtype: int64
```

FAVC
: 91.4%의 사람들이 고열량 음식 자주 섭취

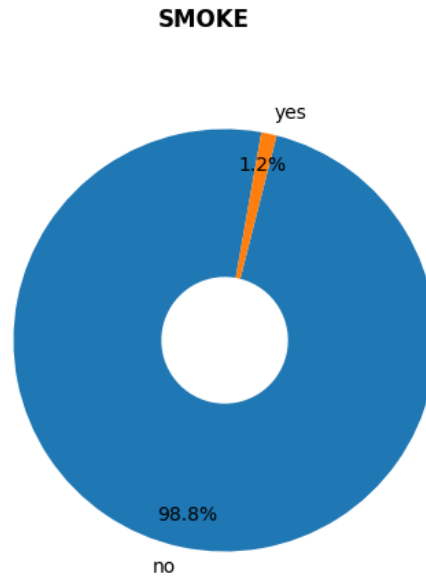
2. 탐색적 자료 분석

◆ 시각화



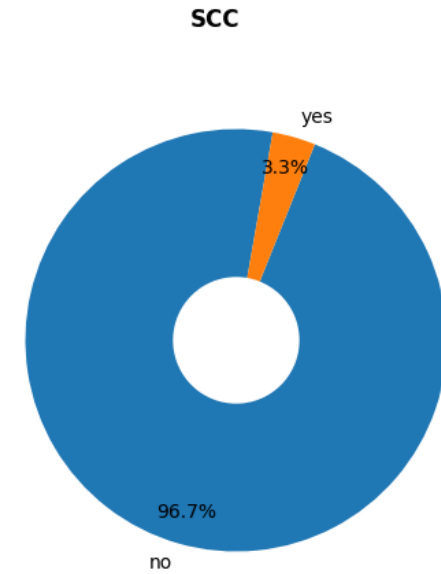
```
CAEC
Sometimes    17529
Frequently    2472
Always         478
no             279
Name: count, dtype: int64
```

CAEC
: 84.4%가 식사 사이 음식을 '가끔' 섭취
1.3%가 식사 사이 음식을 먹지 않음



```
SMOKE
no      20513
yes       245
Name: count, dtype: int64
```

SMOKE
: 98.8%가 비흡연

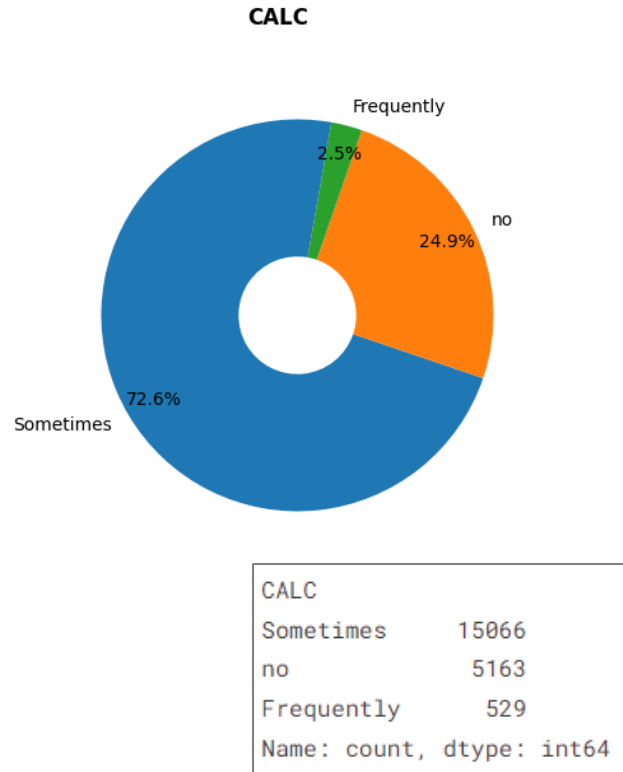


```
SCC
no      20071
yes       687
Name: count, dtype: int64
```

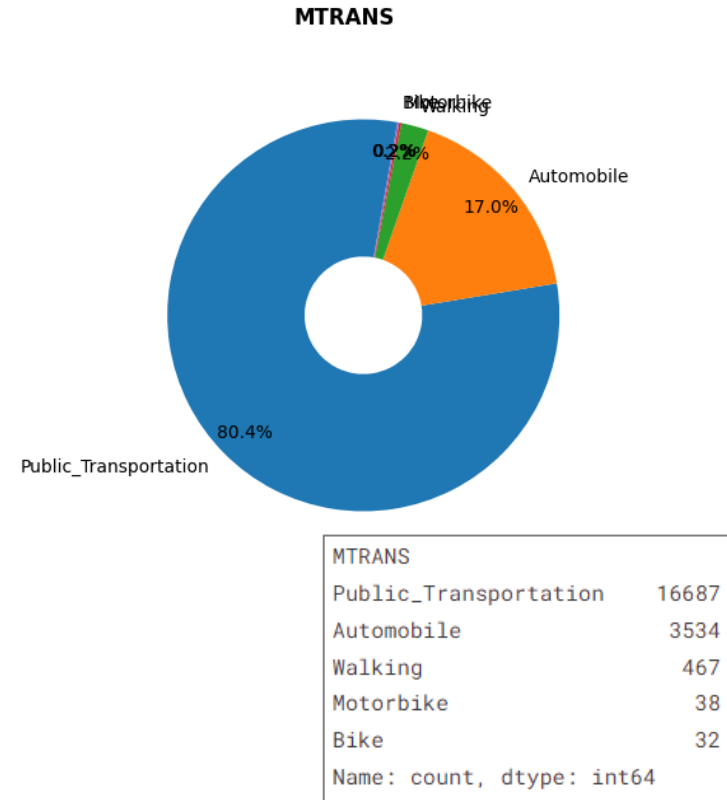
SCC
: 96.7%가 소비 칼로리를 신경쓰지 않음

2. 탐색적 자료 분석

◆ 시각화



CALC
: 72.6%가 술(알코올)을 '가끔' 마심
2.5%가 자주 마심



CALC
: 2.4%가 걷기/자전거타기로 이동
97.6%가 차량 통해 이동

3. 머신러닝 주요 알고리즘 소개

◆ LightGBM

특징

- **리프 중심 트리 분할(Leaf Wise) 방식 사용**
: 트리의 균형을 맞추지 않고 최대 손실 값을 가지는 리프 노드를 지속적으로 분할
> 예측 오류 손실을 최소화할 수 있음(더 나은 정확도)
- **학습 시간과 메모리 사용량이 적음**
- **카테고리형 피처의 자동 변환&최적 변환**
: 원-핫 인코딩 등과 같은 전처리 과정 없이 노드 분할 수행 가능
- **병렬 및 GPU 학습**
: 대용량 데이터에 대한 빠른 학습 가능

주요 파라미터

- **learning_rate**
: 0~1 사이 값 지정, 부스팅 스텝 반복적 수행 시 업데이트되는 학습률 값(기본값 0.1)
- **max_depth**
: 0보다 작은 값 지정 시 깊이 제한 없음. LightGBM은 다른 트리보다 큰 깊이를 가짐.(기본값 -1)
- **n_estimators(num_iterations)**
: 반복 수행하려는 트리의 개수를 지정. 클수록 예측 성능이 높아질 수 있지만 너무 크면 과적합으로 성능 저하될 가능성 있음(기본값 100)

3. 머신러닝 주요 알고리즘 소개

◆ Random Forest

특징

- **배깅의 대표적인 알고리즘**

: 배깅은 같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 알고리즘임.

- **결정 트리의 부트스트랩 샘플링**

: 각 트리는 원본 데이터셋에서 복원 추출된 부트스트랩 샘플을 사용해 학습되도록 함.(여러 개의 데이터 세트가 중첩되게 분리됨)

- **병렬 학습**

: 각 트리가 독립적으로 학습되므로 병렬 학습의 쉬운 구현 가능. 학습시간을 단축하는 데 도움이 됨.

주요 파라미터

- **n_estimators**

: 결정 트리의 개수 지정(기본값 10)

- **max_features**

: 최적의 분할을 위해 고려할 최대 피쳐 개수(기본값 auto=sqrt)

- **max_depth**

: 트리의 최대 깊이

- **min_samples_split**

: 노드를 분할하기 위한 최소한의 샘플 데이터 수. 과적합을 제어

- **min_samples_leaf**

: 분할이 될 경우 왼쪽과 오른쪽의 브랜치 노드에서 가져야 할 최소한의 샘플 데이터 수

4. 모델 평가

◆ 최종모델 선정 과정 (시나리오별 / 모델별)

모델	데이터 재가공	파생변수 유무	피처 엔지니어링	학습시간	정확도
LightGBM+RandomSearch(cv2)	Numerical Variables - Standardization (StandardScaler), Nominal Variables - OneHotEncoding	X	숫자/문자형 데이터 선택&분할	31.50 seconds	0.909
LightGBM+RandomSearch(cv5)				1 min, 53.45 seconds	0.910
RandomFores+RandomSearch				1 min, 2.27 seconds	0.887



◆ 최종모델

모델	학습시간	정확도	F1-Score	최종순위
LightGBM+RandomSearch(cv2)	31.50 seconds	0.909	0.90	2038/3587

최종 순위 : 2038/3587 (Top 57%)

5. 결론 및 리뷰

데이터 분석

응답 결과를 두 분류로 나눌 수 있는(Yes/No, 먹음/안먹음 등) 데이터
: 과체중 가족력, 고열량 음식 섭취, 식사 사이 음식 섭취, 흡연여부, 소비
칼로리 체크, 음주 빈도, 이동수단 등

- ➔ 8~90% 정도는 한 쪽으로 분포 집중
- ➔ 그럼에도 비만 단계는 각각 10~20% 정도로 골고루 분포
- ➔ 다양한 시각으로 상관관계를 파악할 필요가 있음

Feature Engineering

숫자형/문자형 데이터 선택 및 추출하는 작업만 진행

몸무게(Weight), 키(Height)를 통한 BMI 피쳐, 야채 섭취 빈도
(FCVC), NCP(주 식사 횟수)를 곱한 (Meal Habits) 등 다양한 피쳐
를 생성했다면 풍부한 분석이 가능했을 것 같음

모델 관련

LightGBM, RandomForest 모델 사용

- ➔ 2분 미만의 학습시간 소요(짧은 학습시간)
- ➔ 부족한 전처리 과정에도 불구하고 약 90%의 정확도 달성
- ➔ 타겟 데이터를 인코딩하지 않고도 분석 가능했음

5. 결론 및 리뷰

배운 점

- Kaggle을 통한 실전 데이터 경험
- 머신러닝 적용 단계 이해
- LightGBM, RandomForest 모델 특성 파악
- 데이터 전처리에 다양한 방법이 있음을 공부

부족한 점

- 데이터에 접근하는 방법 미숙
 - ➔ 전처리, 피처 엔지니어링 등을 정확하게 하지 못함
 - ➔ 데이터 이해도 부족으로 다양한 시각화를 하지 못함
- 머신러닝에 대한 공부 부족
 - ➔ 특성에 따라 적용할 수 있었던 것이나, 파라미터 조정 등을 제대로 하지 못함
 - ➔ 정확도를 높이지 못함