

Alchemist 3기 아이디어톤 모델 아웃라인 개요

박현서

선택한 데이터셋 :

<https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

주제 : 비만도 그룹에 따른 주 원인 파악 및 그에 따른 퍼스널 케어 방안

모델링 방법

1. 새 데이터가 들어왔을 때에 비만도 예측
 - 로지스틱 회귀
2. 각 그룹별 비만도에 가장 큰 영향을 미치는 변수 파악
 - 회귀계수 분석 및 상관관계 분석
3. 맞춤형 케어
 - 기존 데이터에 있는 비만도 그룹에서, 그 그룹별로 클러스터링 진행. 각 그룹에서 더 세분화된 그룹 생성. 세분화 된 그룹별 주요 원인 파악과 퍼스널 케어 방법 진단하기

알고리즘 선택 : 로지스틱 회귀, 클러스터링

전체 프로세스 (전체 단계에 대한 간단한 설명) (예) 데이터 전처리 → 회귀 → 시각화 → 하이퍼파라미터 튜닝

데이터 전처리

- 나이, 키, 체중 등 스케일이 맞지 않은 변수들이 많기 때문에 결측치나 이상치 처리 및 스케일링이 필수적.
- 범주형 변수가 있기에 one hot encoding 또한 필수적
- Target 인 비만도는 one hot encoding 이 아닌 순서가 정해진 숫자형으로 변환할 것. (ordinary encoding)

로지스틱 회귀

- 데이터셋에 있는 비만도를 target 으로 로지스틱 회귀 모델링

- 결과로 나타난 회귀계수 분석

상관관계 분석

- Target 에 가장 큰 영향을 주는 변수 찾기 위해 상관계수 이용

클러스터링 및 시각화

- 비만도에 따라 나뉜 그룹을 조건문 통해 각각 다른 변수에 저장
- 각 그룹에서 더 정밀한 클러스터링 진행
- 클러스터링 그룹 각 변수 시각화 (평균, 중앙값 이용해서 그룹 특징 파악)
- 시각화 결과에 맞춘 케어 방안 생각...