식품영양학과 논문 데이터셋 준비

2021-07-15

컴퓨터과학과 황승현

논문 설명

- 식품영양학과 김혜림 박사님
- 사람의 나이, 영양, 식이 등 생활 패턴과 특정 질환의 상관 관계 조사
  - 식이 패턴과 고혈압의 상관 관계를 집중 연구
- 입력한 여러 조건 중, 결과에 가장 많은 영향을 주는 조건 찾기
- 질병 모델을 만들고, 새로운 사람이 들어갔을 때 질병이 걸릴 확률 알아내기

해야 할 일

- 데이터 정규화
  - 초기 데이터 10000여개
  - 추적 데이터 4개년은 조사 항목이 달라 보류
- 모델 선정
  - Classification
- 모델 튜닝
- 그 외
  - 고혈압 모델 외에 질환마다 다른 모델 만들기

생활 패턴

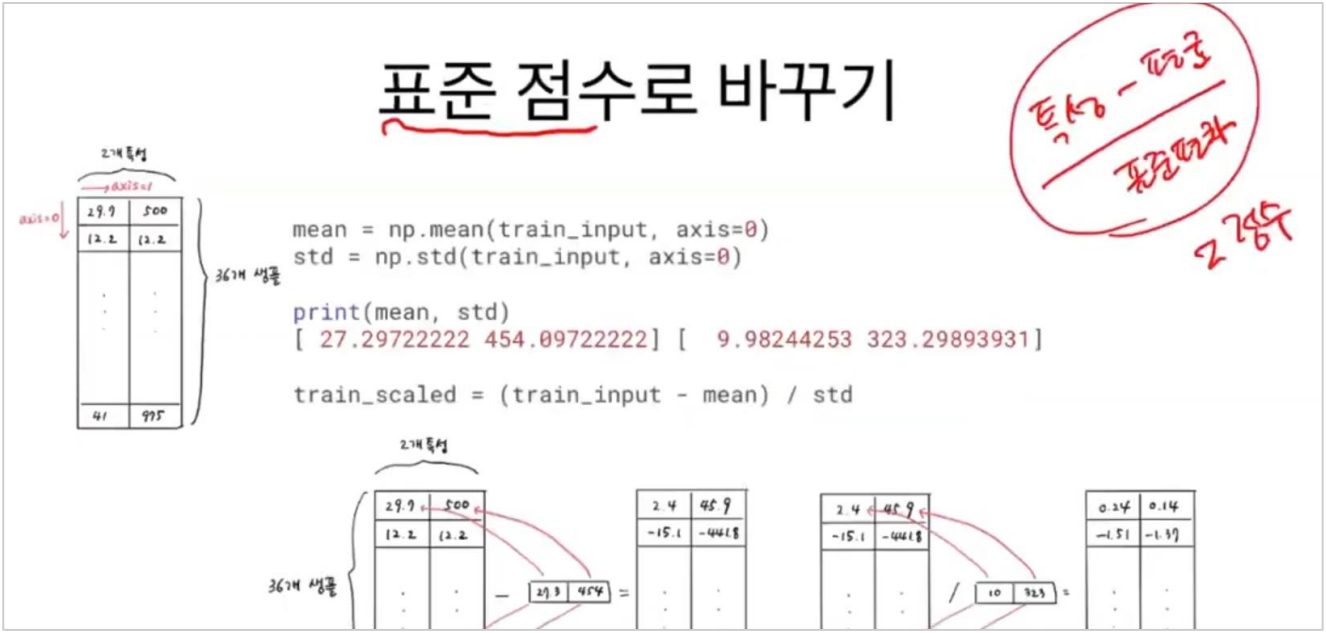
- 독립 변수
- 성별, 연령, 직업 종류, 교육 수준 등 기본 정보
- 음주, 흡연 등 생활습관
- 신체 활동 시간, 수면 시간 등 생활패턴
- 영양소 등 식품섭취빈도

- 요인 분석을 위해 재가공한 데이터
  - 육류, 어패류, 과일류, 면류, 빵류 등 전반적인 식이 패턴

고혈압

- 종속 변수
- 고혈압 여부 판단 3가지 경우 중 1개 이상일때
  - 누운 자세 – 2회 sys 측정 평균 140이상일 경우
  - 누운 자세 – 2회 dia 측정평균 90이상일 경우
  - 혈압약 현재 지속여부 '예'일 경우

데이터 정규화 하는 방법



- CSV 파일을 불러와 NumPy array로 저장
- 데이터의 평균을 구해 특성으로 빼고 표준편차로 나눔
  - Z정수, 가우시안 분포
- sklearn.preprocessing.StandardScaler

이번주 한 일

- 2021-07-09
  - KOGES 데이터 원본 받음 (약 300MB)
- 2021-07-13
  - 데이터 정규화 사전 조사
- 2021-07-14
  - 수정한 데이터 받음

다음주 할 일

- 식품영양학과 김혜림 박사님과 미팅
  - 데이터 설명
  - 필요한 변수가 무엇인지 회의
- 인간대상 및 인체 유래물 연구 관련 연구자 교육 수강
  - 질병관리청
  - KOGES 데이터 다룰 때 필요함
- 데이터 전처리 시작
  - 결측 값 제거
  - 데이터 정규화