20220529

Participants	
Created By	
Created	@2022년 5월 22일 오후 8:51
Last Edited Time	@2022년 5월 24일 오후 9:13
Host	

📊통계분석
✂️전처리
🛠️모델링

📊통계분석

1. 다음은 어느 기업에서 부품의 지름이 실제로 납품서에 적혀있는 값과 동일한지 검사하기 위해 19개의 부품을 임의로 추출하여 지름을 측정한 결과로 얻은 값입니다.

12.5	12.6	12.5	12.4	12.6	12.8	12.3	12.6

1-1. 위 표본의 모평균에 대한 95% 신뢰구간을 구하세요.

1-2. 부품의 납품서에는 부품의 표준편차가 0.2cm라고 적혀있습니다. 이때 모평균에 대한 95% 신뢰구간을 구하세요.

`Hint1` n=9이므로 소표본에 대한 경우입니다.

`Hint2` 표본평균의 분포 알아보기

2. 다음은 3가지 음식에 포함된 콜레스테롤 함량을 4개의 다른 실험실에서 측정한 결과입니다.

	치킨	피자	아이스크림
실험실A	3.4	2.6	2.8
실험실B	3.0	2.7	3.1
실험실C	3.3	3.0	3.4
실험실D	3.5	3.1	3.7

2-1. 음식에 포함된 콜레스테롤 함량이 음식의 종류와 어느 실험실에서 실험이 이루어졌는지에 따라 다르다고 할 수 있는지 유의수준 0.05 하에서 검정하세요. 절차는 아래와 같습니다.

1) 가설설정
2) 검정통계량 값 혹은 유의확률 도출
3) 가설 기각 여부 결정
4) 해석

`Hint` 이원분산분석

✂️전처리

3. **첨부된 'smoke' 데이터는 2018년도 성인의 건강검진 데이터입니다.**

3-1. 수축기혈압과 이완기혈압기 값의 차이로 새로운 컬럼(`혈압차`)을 생성하고, 연령대 코드별 각 그룹 중 `혈압차`의 분산이 5번째로 큰 연령대 코드를 구하세요.

3-2. 허리둘레를 신장으로 나눈값으로 새로운 컬럼(`WHtR`)을 생성하고, 아래의 표에 따라 비만인 남성과 여성의 비율을 구하세요.

- Waist-to-Height Ratio(WHtR)= 허리둘레(in/cm) ÷ 신장(in/cm)

WHtR	성인 여성	성인 남성
심각한 저체중	≤ 0.34	≤ 0.34
저체중	0.35 – 0.41	0.35 – 0.42
건강함	0.42 – 0.48	0.43 – 0.52
과체중	0.49 – 0.53	0.53 – 0.57
심각한 과체중	0.54 – 0.57	0.58 – 0.62
비만	≥ 0.58	≥ 0.63

ref. https://www.mdapp.co/waist-to-height-ratio-whtr-calculator-433/

🛠️모델링

4. **전처리가 완료된 'smoke' 데이터를 활용하여 흡연상태를 예측하는 모델을 만드세요.**
   - `연속형 피처` min-max scaling
   - `범주형 피처` 'fastDummies' 패키지 활용한 원핫인코딩
   - `타겟` 흡연상태 (1:흡연, 0:비흡연)
   - `시드 넘버` 2022
   - `트레인-테스트 비율` 7:3

4-1. svm, xgboost, randomforest 3개 알고리즘의 공통점을 쓰고, 예측 분석에 적합한 알고리즘인지 설명하세요.

4-2. 위의 3가지 방법으로 모두 모델링 해보세요. 그리고 가장 적합한 알고리즘 선택하고 이유를 설명하세요. 절차는 아래와 같습니다.

1) 한계점

2) 보완 가능한 부분

3) 현업에서 사용시 주의할 점