기초 통계&ML 과제 REPORT

Iris

1. 데이터 확인 및 기술통계량 분석

- Iris 데이터셋은 꽃받침 길이(sepal_length), 꽃받침 너비(sepal_width), 꽃잎 길이(petal_length), 꽃잎 너비(petal_width), 그리고 종(Species: setosa, versicolor, virginica)로 구성된 데이터로, 총 150개의 샘플을 포함함.
- petal_length를 기준으로 종(Species)별 평균, 표준편차, 사분위수를 산출한 결과, 각 종은 50개의 샘플로 구성되어 있으며 평균은 다음과 같이 증가함:
  - Setosa < Versicolor < Virginica
- Setosa는 평균이 작고 분산도 작으며, Virginica는 평균과 분산이 모두 큼.

2. Boxplot 분석

- Setosa: Petal Length가 가장 짧음, 분포 범위가 좁음, 이상치가 일부 존재
- Versicolor: 분포 범위가 비교적 넓음, 이상치 일부 존재
- Virginica: Peral Length가 가장 길, 분포 범위가 넓음, 이상치 없음

- 세 종의 IQR과 중앙값을 보았을 때 종 간 Petal Length에 명확한 차이가 있는 것으로 보임.

3. 정규성 검정

- 가설 설정
  - 귀무가설: 각 그룹의 Petal Length는 정규성을 따른다.
  - 대립가설: 각 그룹의 Petal Length는 정규성을 따르지 않는다.
- 정규성 검정 결과
  - p-value는 모두 0.05 이상으로, 귀무가설을 기각할 수 없음.
  - 따라서 각 그룹은 정규성을 따른다고 볼 수 있음.

4. 등분산 검정

- 가설 설정

    - 귀무가설: 3개 Species의 Petal Length 분산은 서로 같다. (등분산)
    - 대립가설: 적어도 하나의 그룹은 분산이 다르다. (이분산)
- 등분산 검정 결과
    - p-value가 0.05보다 작기 때문에 귀무가설을 기각함. 즉, 등분산 가정을 할 수 없음
    - 하지만, 명세에 따라 등분산 가정 하에 분석을 진행함.

5. ANOVA

- 가설 설정
    - 귀무가설: 세 종 간의 평균 petal_length는 같다.
    - 대립가설: 적어도 한 종의 평균 petal_length는 다르다.
- 검정 결과
    - F = 1180.16, p<0.001로 귀무가설을 기각함. 즉, 세 종 간의 평균 petal_legnth가 같다고 가정할 수 없음.

6. 사후검정

- 어떤 종 간의 평균 petal_length가 같지 않은지 판단하기 위해 사후 검정 진행.
- 검정 결과
    - 모든 쌍(setosa-veriscolor, setosa-virginica, veriscolor-virginica)의 p-adj 값이 유의수준(0.05)보다 낮음.
    - 즉, 세 쌍 모두 평균 petal_length의 차이가 유의하게 있음.

7. 결과 요약

- boxplot
    - 세 종의 petal_length의 분포에 명확한 차이를 볼 수 있음
    - Virginica > Versinolor > Setosa 순으로 petal_length가 큼.
- anova
    - p-value가 유의 수준보다 낮기 때문에 귀무가설을 기각하여 세 종의 평균 petal_length가 같다고 할 수 없음.
- tukey HSD
    - 모든 종 간 평균 petal_length의 길이에는 유의한 차이가 있다고 할 수 있음.

- 최종 결론:
    - 세 종 간 petal_legnth는 유의한 차이가 있으며 Virginica의 평균 petal_length가 가장 크며, Setosa의 petal_length가 가장 작음.

신용카드

1. 데이터 로드 및 탐색

- 28개 변환 특성과 Time, Amount, Class 포함
- Class=1 비율 약 0.17% → 극심한 불균형

2. 샘플링

- Class=1(사기) 전체 유지, Class=0(정상) 10,000건 무작위 추출
- 샘플링 후 비율: Class 1 약 4.7%

3. 전처리

- amount 표준화 후 삭제
- x, y 분리
- Train/Test = 8:2로 분할

4. SMOTE 적용

- 학습 데이터에 대해 Class 1을 오버샘플링하여 Class 0과 동일 수로 맞춤

5. 모델 학습 결과 (Ramdom Forest)

- Recall: 0.89
- F1-score: 0.92
- PR-AUC: 0.9537

6. 결론

- 기준 성능(Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90)을 모두 충족
- SMOTE 적용이 성능 개선에 효과적이었음