<기초 통계 / ML 과제 보고서>

1. Iris 데이터셋을 활용해 클래스별 변수 평균 차이를 검정

1) 분석 목적: Iris 데이터셋을 이용해 세 Species(setosa, versicolor, virginica) 간 petal length의 평균 차이가 통계적으로 유의한지를 검정하기 위함.

2) 분석 절차 및 결과
- 기술통계량 요약: virginica의 평균 petal length가 가장 길고, setosa의 평균 petal length가 가장 짧음
- Boxplot: setosa < versicolor < viriginica 순
- 정규성 검정: 모든 그룹에서 p-value > 0.05로 정규성 가정을 만족함
- 등분산성 검정: p-value < 0.05 (세 그룹 간 분산은 유의하게 다름) 로 나타났으나, 이후 분석에서는 등분산성 가정을 수용하고 ANOVA를 수행
- ANOVA 결과: F=1180.16, p<0.001 로 귀무가설 기각
- 사후검정: Tukey HSD로 모든 쌍 간 유의미한 차이를 확인

3) 결론
- 세 Species 간 꽃잎 길이에 통계적으로 유의한 차이가 존재하며, 길이 순서는 “setosa < versicolor < virginica” 순이다.

2. 실제 신용카드 사기 데이터셋을 활용해 클래스 불균형 상황에서 분류 모델을 학습

1) 학습 목표: 심각한 클래스 불균형을 갖는 신용카드 사기 거래 데이터에서 사기 거래 (Class=1)를 높은 재현율로 탐지하는 모델을 개발함

2) 주요 처리 과정
- 샘플링: Class 0을 10,000건 무작위 추출
- 전처리: Amount 변수를 표준화한 후 Amount_Scaled로 대체
- SMOTE 적용: 학습 데이터셋에서 Class 1을 오버샘플링
- 모델 학습: Random Forest 모델 사용
- 지표 확인 (Class 1): Recall 0.8878, F1-score 0.9158, PR-AUC 0.9537
- 지표 확인 (Class 0): Recall 0.9975, F1-score 0.9960

3. 결론
- 목표 성능 (Recall $\geq$ 0.80, F1 $\geq$ 0.88, PR-AUC $\geq$ 0.90)을 모두 충족하였으며, 해당 모델은 실제 금융 도메인에서도 신뢰성 있는 탐지 성능을 가질 수 있다.