1. 기초 통계 분석

A. 실습 목표

i. Iris 데이터셋을 활용하여 세 품종(setosa, versicolor, virginica) 간에 꽃잎 길이(petal Length)의 평균 차이가 통계적으로 유의미한지 검정한다. 이를 위해 기술 통계 분석, 시각화, 정규성 및 등분산성 검정, ANOVA, 사후 검정(Tukey HSD)을 수행한다.

B. 분석 내용

i. 데이터셋 구조 확인

1. Iris 데이터셋은 총 150개의 샘플로 구성되어 있으며, 각 샘플은 Sepal Length, Sepal Width, Petal Length, Petal Width, Species 정보를 포함한다.

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

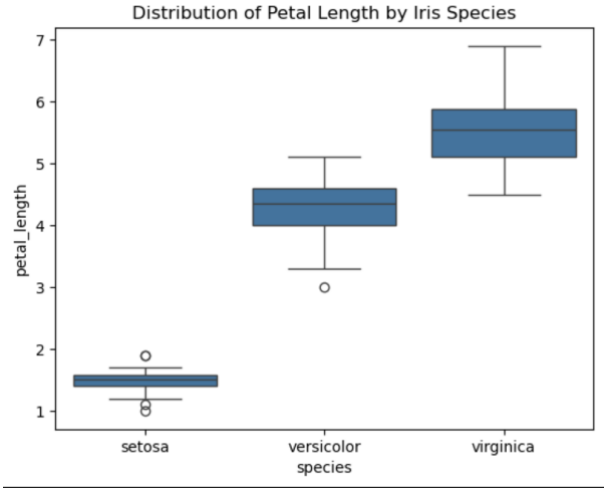
ii. 기술 통계량 산출

1. 각 종 별 petal length의 평균, 표준편차, 최소/최댓값 등을 확인했다.

species	count	mean	std	min	25%	50%	75%	max
setosa	50.0	1.462	0.173664	1.0	1.4	1.50	1.575	1.9
versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1
virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9

iii. 시각화

1. Boxplot을 통해 세 품종의 petal length 분포를 시각화 했다.
2. 결과



3. 해석
   A. Petal length의 평균은 virginica가 가장 크고, setosa가 가장 작다.
      i. Setosa : 중앙값이 가장 낮고, 분산이 작아 꽃잎 길이가 짧고 일정하게 분포함을 보여준다.
      ii. Virginica: 중앙값이 가장 높고, 분산이 커 꽃잎 길이가 길며 개체 간 차이가 큰 경향을 보인다.

iv. 정규성 검정

1. 가설 수립
   A. H0: 해당 종의 petal length는 정규분포를 따른다.
   B. H1: 해당 종의 Petal length 데이터는 정규분포를 따르지 않는다.
2. 검증 과정
   A. Species 별로 Shapiro-wilk 검정을 실시했다.
3. 결과
   A. Setosa: p-value = 0.0548
   B. Versicolor: p-value = 0.1585
   C. Virginica: p-value = 0.1098
4. 해석
   A. 모든 그룹에서 p-value가 0.05보다 크므로, 세 그룹 모두에서 정규성을 만족한다고 판단할 수 있다.

v. 등분산성 검정

1. 가설
   A. H0: 세 그룹의 분산이 모두 같다.
   B. H1: 적어도 하나의 그룹은 분산이 다르다.
2. 검증 과정

A. 세 그룹 간의 등분산성을 검정하기 위해 Levene 검정을 실시했다.

3. 결과

A. P-value = 0.00000003

4. 해석

A. P-value가 유의수준 0.05보다 작으므로, 귀무가설을 기각하고 등분산성이 만족되지 않는다고 판단하였다.

vi. One-way ANOVA

1. 가설

A. H0: 3개 species 간의 petal length 평균은 모두 같다.

B. H1: 적어도 한 species의 petal length 평균은 나머지와 다르다.

2. 결과

	sum_sq	df	F	PR(>F)
species	437.1028	2.0	1180.161182	2.856777e-91
Residual	27.2226	147.0	NaN	NaN

3. 해석

A. P-value가 유의수준 0.05보다 작기 때문에 귀무가설을 기각한다. 종에 따라 petal length 평균에 통계적으로 유의미한 차이가 있다고 판단된다.

vii. 사후검정

1. 검증 과정

A. ANOVA 분석 결과가 유의하였기에, 사후 검정으로 Tukey HSD를 수행하였다.

2. 결과

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
setosa	versicolor	2.798	0.0	2.5942	3.0018	True
setosa	virginica	4.09	0.0	3.8862	4.2938	True
versicolor	virginica	1.292	0.0	1.0882	1.4958	True

3. 해석

A. 모든 그룹 쌍에서 p-value가 0.05보다 작고, 95% 신뢰구간이 0을 포함하지 않았다. 따라서 모든 품종 간 평균의 차이가 통계적으로 유의미함을 확인하였다.

C. 요약

i. Iris 데이터셋을 활용하여 세 종 간의 petal length 평균 차이가 존재하는지 통계적으로 검정했다.

1. Boxplot 시각화를 통해 virginica 품종이 가장 긴 petal length를, setosa가 가장 짧은 petal length 가짐을 확인하였다.
2. One-way ANOVA 분석을 통해 세 품종 간 평균 petal length의 차이가 통계적으로 유의미함을 확인하였다. ($p<0.05$)
3. Tukey의 사후검정에서 모든 품종 쌍 간에 유의한 평균 차이가 존재함을 확인하였다.

- 따라서 세 품종의 평균 petal length는 통계적으로 유의미한 차이가 있으며, setosa < versicolor < virginica 순으로 petal length가 유의미하게 길다.

2. 기초 머신러닝 분석

A. 실습 목표

i. 신용카드 거래 데이터에서 사기 거래(Class=1)를 식별할 수 있는 분류 모델을 구축하고, 평가 지표를 통해 모델의 성능을 검증한다.

B. 실습 과정

i. 데이터셋 구조 확인

1. 284,807건의 거래가 포함되어 있으며, class 분포를 확인한 결과, 정상 거래가 99.8% 이상을 차지하고, 사기 거래는 약 0.17%로 매우 희소하게 나타났다.

ii. 샘플링

1. 과정
   A. Class 불균형 문제를 완화하기 위해 정상 거래 중 10,000건을 무작위로 추출하고, 사기 거래는 전체를 유지하여 새로운 학습용 데이터셋을 구성하였다.
2. 결과
   A. 샘플링된 데이터셋에서 사기 거래의 비율이 약 4.7%로 증가하였으며, 이는 원본 데이터에 비해 Class 1이 상대적으로 보강된 것을 의미한

다.

iii. 데이터 전처리

1. 과정
   A. Amount 변수를 StandardScaler를 이용하여 표준화하였다.
   B. X(Class 제외 열) , y(Class 열) 로 데이터프레임을 분리하였다.

iv. 학습 데이터와 테스트 데이터 분할

1. 전체 데이터를 학습셋과 테스트셋으로 8:2 비율로 분할하였다.

v. SMOTE 적용

1. 현재 학습 사기 거래(Class=1)의 비율이 여전히 낮아, 모델이 해당 클래스의 패턴을 제대로 학습하지 못할 가능성이 존재하였다. 이에 따라 SMOTE를 적용하여 소수 클래스 샘플을 합성하고, 정상 거래(class=0)와 유사한 수준으로 클래스 분포를 조정하였다.
2. 결과
   A. SMOTE 적용 후, 학습 데이터에서 소수 클래스였던 사기 거래(Class=1)가 394건에서 7,999건으로 합성되어, 두 클래스의 비율이 동일해졌다.

vi. 모델 학습

1. Random Forest 모델을 선정하여 하용하였다.
2. 예측값 및 예측 확률

```
예측값: [0 0 0 0 0 0 1 0 0 0]
예측 확률 [0.01 0.   0.01 0.   0.   0.01 0.99 0.   0.05 0.  ]
```

   A. 실제 예측 결과를 확인한 결과, 사기 거래로 판단되는 거래에 대해 높은 확률을 부여한 경우가 많아 신뢰도 있는 분류가 이루어졌음을 확인하였다.
3. 평가 지표 확인
   A. Precision, recall, f1-score를 확인하였다.
   B. 결과

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      2001
           1       0.98      0.83      0.90        98

    accuracy                           0.99      2099
   macro avg       0.98      0.91      0.95      2099
weighted avg       0.99      0.99      0.99      2099


 PR-AUC:
0.9156887960343137
```

i. 모든 성능 지표가 제시된 기준 (Recall ≥ 0.80, F1-score ≥ 0.88, PR-AUC ≥ 0.90)을 만족하였다.

ii. 따라서 해당 모델이 정상 거래와 사기 거래를 모두 높은 정확도로 분류할 수 있는 신뢰도 높은 분류 모델임을 확인할 수 있었다.

C. 요약

i. 초기 데이터셋은 총 284,807의 거래로 구성되어 있었으며, 이 중 사기 거래는 약 0.17%에 불과해 클래스 불균형 문제가 존재했다.

ii. 이를 해결하기 위해 정상 거래 중 10,000건을 무작위 추출하고 사기 거래는 전부 유지하여 샘플링된 데이터셋을 구성하였다.

iii. SMOTE를 적용하여 소수 클래스(Class=1)의 비율을 균형 있게 맞췄다.

iv. 랜덤 포레스트 모델을 학습하고 테스트셋에 대해 예측을 수행했다.

v. 최종 모델이 제시된 기준을 모두 만족하여 정상 거래와 사기 거래를 높은 정확도로 분류할 수 있는 신뢰도 높은 분류 모델임을 확인했다.