기초 통계 / ML 과제

이지민

1. 기초 통계 과제 (Iris 데이터셋)

Species별 Petal Length의 평균, 표준편차, 최소값, 최대값, 사분위수

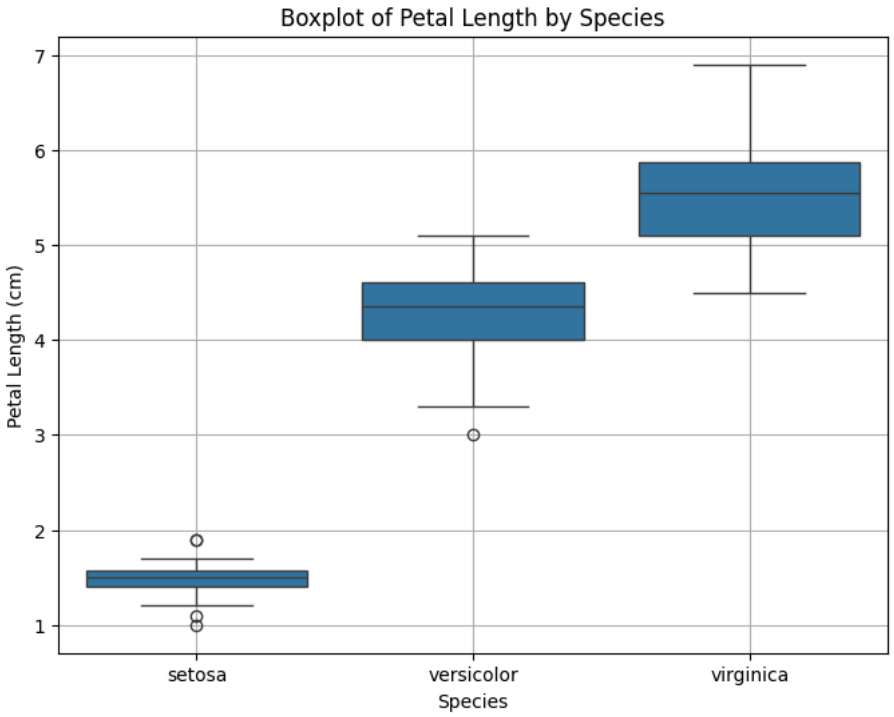
Species별 Petal Length의 평균, 표준편차, 최소값, 최대값, 사분위수 :

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

그룹별 데이터 개수 :

species	sepal_length	sepal_width	petal_length	petal_width
setosa	50	50	50	50
versicolor	50	50	50	50
virginica	50	50	50	50

boxplot



virginica 평균값 중앙값 높고 setosa 평균값 중앙값 낮다. 대체로 다른 위치의 분포를 보임

정규성 검정

```
Setosa     W-statistic: 0.9549767850318988, p-value: 0.0548114671955363
Versicolor W-statistic: 0.96600440254332, p-value: 0.15847783815657573
Virginica  W-statistic: 0.9621864428612802, p-value: 0.10977536903223506
```

모든 종에서 유의수준 0.05 에서 귀무가설 채택. 즉 정규성을 만족한다.

등분산성 검정

```
W-statistic: 19.480338801923573, p-value: 3.1287566394085344e-08
```

가설 :

귀무가설 (H_0): 세 그룹의 분산이 모두 같다 (등분산성 만족)
대립가설 (H_1): 세 그룹의 분산 중 적어도 하나 이상 다르다

유의수준 0.05에서 귀무가설 기각, 등분산성 만족하지 않음

ANOVA

```
ANOVA F-statistic: 1180.161182252981, p-value: 2.8567766109615584e-91
```

종 간 petal_length의 평균에는 유의미한 차이가 있다.

Tukey HSD

```
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
-------------------------------------------------------------
```

세종 사이에 유의미한 차이가 존재하고, 세 그룹 모두 유의수준 0.05에서 유의미한 차이를 보인다.

2. 기초 머신러닝 과제 (신용카드 사기 탐지)

4. 학습 데이터와 테스트 데이터 분할
   train_test_split을 사용해 학습셋:테스트셋 비율을 8:2로 나누고,
   stratify=y 옵션으로 클래스 비율 유지, 분할된 데이터의 Class 비율을 출력하시오.
   (random_state는 42로 설정)

```
학습셋 Class 분포: Class
0    7999
1     394
Name: count, dtype: int64

테스트셋 Class 분포: Class
0    2001
1      98
Name: count, dtype: int64
```

5. SMOTE 적용
   학습 데이터(X_train)에 SMOTE를 적용하여 소수 클래스(사기 거래)를
   오버샘플링하시오. (왜 SMOTE를 적용해야하는지까지 서술하시오.)
   SMOTE 적용 전후의 사기 거래 건수를 출력하시오.

```
SMOTE 적용 전:
 Class
0    7999
1     394
Name: count, dtype: int64

SMOTE 적용 후:
 Class
0    7999
1    7999
Name: count, dtype: int64
```

SMOTE를 적용해야하는 이유는 현재 클래스간의 불균형이 존재하기 때문이다. 사기 거래는 매우 적고, 편향된 모델로 학습되기 쉽다.

6. 선형모델 적합, 7. 최종성능 평가

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

           0     0.9945    0.9965    0.9955      2001
           1     0.9255    0.8878    0.9062        98

    accuracy                         0.9914      2099
   macro avg     0.9600    0.9421    0.9509      2099
weighted avg     0.9913    0.9914    0.9913      2099

PR-AUC (average_precision_score): 0.9546511280244321
```

조건을 만족한다.