3-1 Report

기초 통계/ML

2022148072 정지윤

기초 통계 과제 (Iris 데이터셋)

1. 데이터 로드 및 구조 확인

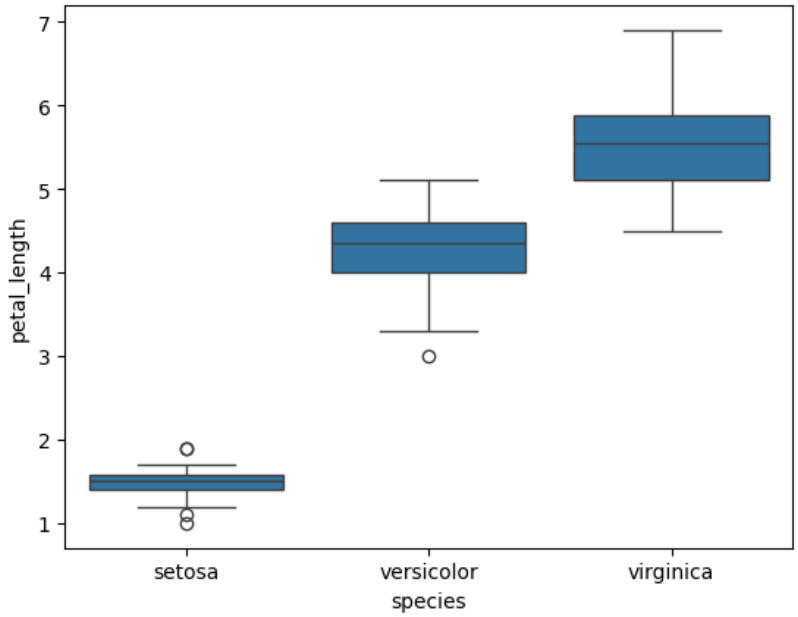
seaborn.load_dataset('iris')를 호출하여 데이터를 로드한다. head(), info() 등을 적용하여 데이터의 타입이 float64 이고, species 범주를 가짐을 확인할 수 있었다.

2. 기술통계량 산출

describe()로 확인한 petal_length 의 통계는 평균이 3.758, 표준편차가 1.765, 최소-최대값이 1.0-6.9 이다. 또한 value_counts()로 세 품종(setosa, versicolor, virginica) 별 관측치가 50 개씩 균등함을 확인하였다.

3. 시각화

seaborn 으로 그린 boxplot 은 다음과 같다.



세 품종 간 꽃잎 길이의 분포가 뚜렷하게 구분되는 것을 확인할 수 있다. setosa 는 중앙값이 약 1.5 로 가장 짧고 변동성이 작다. versicolor 의 중앙값은 약 4.3 으로 중간 수준의 변동성을 보인다. virginica 는 중앙값이 약 5.5 로 가장 긴 길이를 가지면서 변동성도 가장 크다. 세 그룹의 분포가 겹치지 않기 때문에 petal_length 는 분류 지표로서 적절하다고 판단된다.

4. 정규성 검정 (Shapiro-Wilk)

귀무가설(H0)은 '해당 품종의 petal_length 분포는 정규분포를 따른다.', 대립가설(H1)은 '해당 품종의 petal_length 분포는 정규분포를 따르지 않는다.'이다.

품종별로 shapiro()를 적용한 결과는 다음과 같다.

	setosa	versicolor	Virginica
p-value	0.0548	0.1585	0.1098

세 품종 모두 유의수준 0.05 를 넘는 p-value 를 보여 귀무가설이 기각되지 않았다. 즉, 정규성 가정을 크게 위배하지 않고 있기 때문에 이후 분산분석(ANOVA) 등을 적용할 수 있다.

5. 등분산성 검정 (Levene)

귀무가설(H0)은 '세 품종의 petal_length 분산은 동일하다.', 대립가설(H1)은 '적어도 하나의 품종 petal_length 분산은 다르다.'이다.

levene()를 적용한 결과 stat=19.4803, p-value=0.0 이 계산되었다. p-value 가 0.05 보다 작아 귀무가설을 기각한다. 즉, 세 품종 간 petal_length 의 분산은 통계적으로 유의하게 다르다.

6. 가설 수립

- 귀무가설(H0): 3 개 species 간 petal length 의 평균은 동일하다.
- 대립가설(H1): 적어도 하나의 species 의 petal length 평균은 다른 species 와 다르다.

7. ANOVA

ANOVA 결과 stat=1180.1612, p<0.001 로 유의수준 0.05 보다 작아 귀무가설을 기각한다. 이는 세 품종의 petal_length 평균이 모두 동일하다는 가정이 통계적으로 성립하지 않음을 의미한다. 따라서 품종별 petal_length 평균에는 유의미한 차이가 존재한다.

8. 사후검정 (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
-----------------------------------------------------
```

세 품종 간의 모든 쌍별 비교에서 p-adj 값이 0.0 이고, reject=True 인 것을 보아 각 비교가 통계적으로 유의미함을 확인할 수 있다.

9. 결론

세 그룹 간 petal_length 의 차이는 통계적으로 뚜렷하다. Boxplot 에서 각 그룹의 분포가 분리됨을 시각적으로 확인하였고, ANOVA 에서는 세 그룹 평균의 차이가 유의미함을 확인하였으며, Turkey HSD 에서는 모든 쌍별 비교에서 각 품종 간 평균 길이가 서로 다르다는 결론을 얻었다.

따라서 petal_length 는 virfinica 가 가장 길고, versicolor 가 중간, setosa 가 가장 짧다고 판단할 수 있다.

기초 머신러닝 과제 (신용카드 사기 탐지)

1. 데이터 로드 및 기본 탐색

```
Class
0    284315
1       492
Name: count, dtype: int64
```

crefitcard.csv 에는 결측치가 없으며, 사기(Class=1) 거래가 492 건으로 매우 희소하다.

descrive()로 확인한 수치 변수들은 평균이 0 에 가깝고 분산이 작다.

2. 샘플링

불균형을 완화하기 위해 사기 492 건을 모두 유지하고 정상 10,000 건을 무작위 추출하여 10,492 행의 분석용 데이터프레임을 생성하였다. 생성 후 Class 비율은 95:5 정도로 개선되긴 했으나 여전히 불균형이 크다.

```
Class
0    10000
1      492
Name: count, dtype: int64
```

3. 데이터 전처리

Amount 만 StandardScaler 로 표준화하여 Amount_scaled 로 교체하고 원본 열은 삭제한다.

4. 학습 데이터와 테스트 데이터 분할

train_test_split 으로 데이터셋을 분할하여 test 8399 개, test 2099 개를 얻었다.

```
Training set:
Class
0    7999
1     394
Name: count, dtype: int64

Test set:
Class
0    2001
1      98
Name: count, dtype: int64
```

5. SMOTE 적용

train dataset 에 SMOTE(k-NN 기반 합성 오버샘플링)을 적용한다. SMOTE 는 소수 클래스(Class=1)의 합성 샘플을 만들어 클래스 불균형을 완화하는 역할을 한다. 결과적으로 Class=1 샘플이 394->7999 로 증폭되어 양 클래스의 균형이 1:1 으로 맞춰졌다.

6. 모델 학습

랜덤포레스트를 선택하였다. 비선형적 패턴을 가지는 task 에서 결정경계의 유연성을 제공하고, 안정적인 예측 성능을 제공하기 때문이다. 학습 결과는 다음과 같다.

	precision	recall	f1-score	support
0	0.99	1.00	1.00	2001
1	0.95	0.89	0.92	98
accuracy			0.99	2099
macro avg	0.97	0.94	0.96	2099
weighted avg	0.99	0.99	0.99	2099

7. 최종 성능 평가

최종 성능은 전체 정확도 99%를 달성했으며, 정상거래(Class 0)는 Precision 0.99, Recall 1.00, F1 1.00 로 거의 완벽하게 분류하고 있다. 사기거래(Class 1)는 Precision 0.95, Recall 0.89, F1 0.92 를 달성하여 목표 기준을 만족하고 있다.