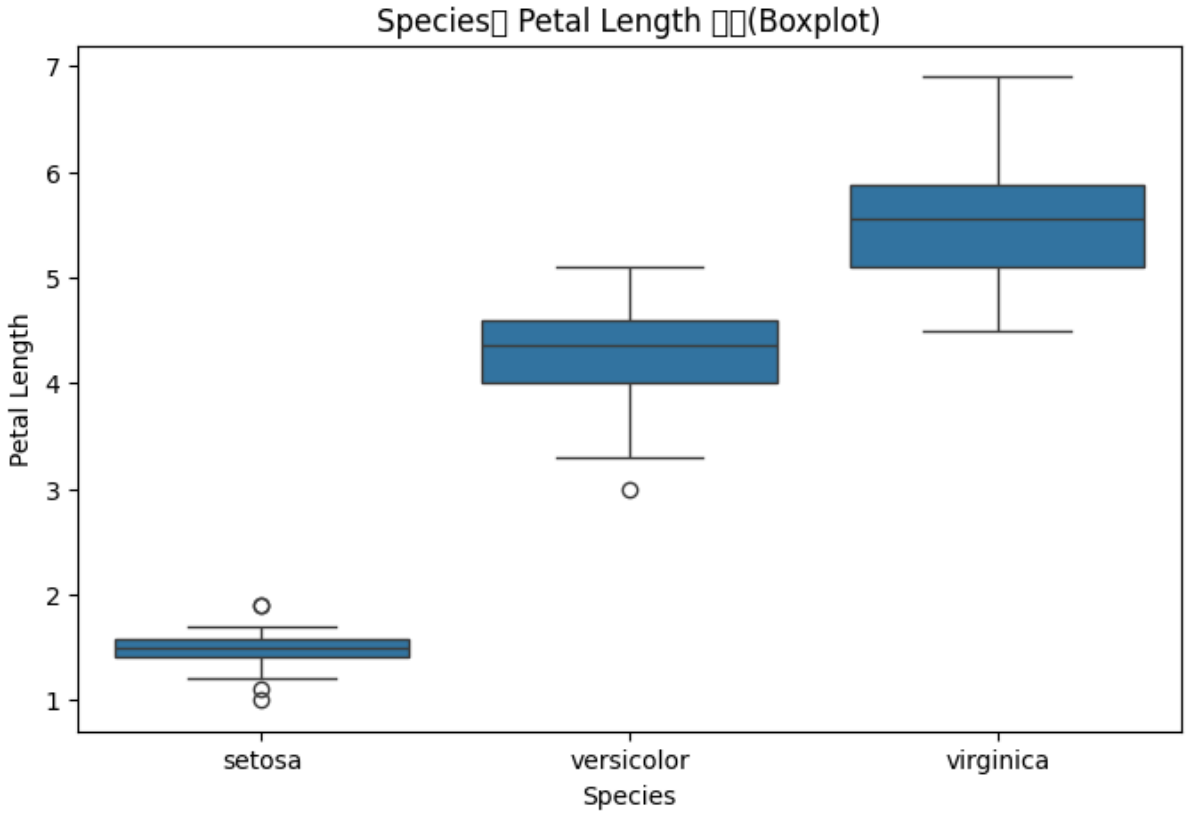
기초통계/ML 과제

기초통계



시각화결과, 이렇게 Virginia 그룹의 평균이 높은것을 확인할 수 있었음.

Shapiro-wilk 정규성 검정결과는 아래와 같음

setosa: p-value = 0.0548 → 정규성 가정 만족 (귀무가설 채택)
versicolor: p-value = 0.1585 → 정규성 가정 만족 (귀무가설 채택)
virginica: p-value = 0.1098 → 정규성 가정 만족 (귀무가설 채택)

Levene검정을 통한 등분산성 검정 결과는 아래와 같음

Levene 등분산성 검정 결과 (Species별 Petal Length)
귀무가설: 세 그룹의 분산은 모두 같다.
대립가설: 적어도 한 그룹의 분산이 다르다.
p-value = 0.0000 → 등분산성 가정 불만족 (귀무가설 기각)

따라서 다음과 같은 가설을 수립함

귀무가설(H_0):
3개 Species(품종) 간 Petal Length(꽃잎 길이)의 평균은 모두 같다.
대립가설(H_1):
적어도 한 개 이상의 Species(품종)에서 Petal Length(꽃잎 길이)의 평균이 다르다.

ANOVA 검정 결과

One-way ANOVA 결과 (Species별 Petal Length)
F값: 1180.1612
p-value: 0.0000
p-value < 0.05 이므로 귀무가설 기각 → 세 그룹의 평균이 적어도 하나는 다르다.

와 같은 결과가 나왔고, Tukey 사후검정 결과 setosa-versicolor, setosa-virginica, versicolor-virginica 사이에 유의미한 차이가 있다는 사실을 알 수 있었다.

결과적으로, Boxplot과 기술통계량을 보면 virginica 그룹의 Petal Length(꽃잎 길이) 평균이 가장 길고, 그 다음이 versicolor, setosa가 가장 짧았다.
ANOVA 분석 결과 세 그룹 간 Petal Length의 평균에 유의미한 차이가 있음이 확인되었고 (p-value < 0.05, 귀무가설 기각), Tukey HSD 사후검정 결과, 모든 그룹 쌍(setosa–versicolor, setosa–virginica, versicolor–virginica)에서 평균의 차이가 통계적으로 유의미한것도 확인되었다.

ML

SMOTE를 적용해야 하는 이유:
SMOTE(Synthetic Minority Over-sampling Technique)는
소수 클래스(여기서는 사기 거래)의 데이터를 인공적으로 생성하여
클래스 불균형 문제를 해결하는 기법인데, 데이터가 불균형하면, 모델이 다수 클래스(정상 거래)에 치우쳐
소수 클래스(사기 거래)를 잘 예측하지 못하는 문제가 발생하고 자꾸 False negative 응답을 하는 쓸모없는 예측기가 학습된다.
따라서 SMOTE를 적용하면 소수 클래스의 표본 수가 늘어나 모델이 소수 클래스도 잘 학습할 수 있게 도와준다.

모델로는 로지스틱 회귀 모델을 선택했는데, 이진분류에 적합한 모델이고 문제의 복잡성을 보았을때 충분히 성능을 내면서도 과적합되지 않을것같은 이유에서 본 모델을 선택했음.

결과적으로 요구받은 recall 0.8, F1 0.88, PR-AUC 0.9를 거의 만족하는 성능을 내는것을 확인할 수 있었음.