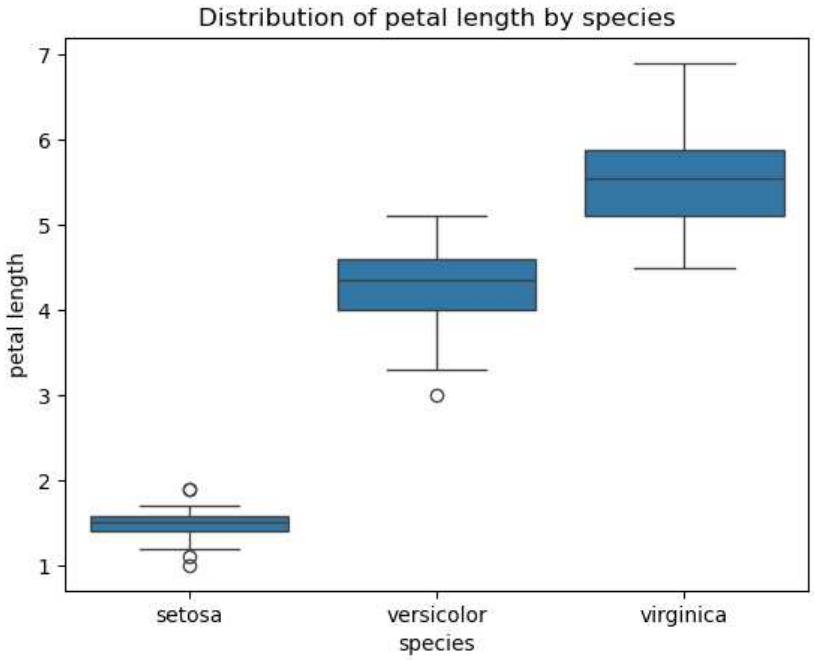
<기초 통계 / ML 과제 보고서>

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 검정한다.

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

```
setosa: p-value = 0.0548
versicolor: p-value = 0.1585
virginica: p-value = 0.1098
```

- 등분산성 검정: p-value < 0.05 (세 그룹 간 분산은 유의하게 다름) 로 나타났으나, 이후 분석에서는 등분산성 가정을 수용하고 ANOVA를 수행

```
setosa variance: 0.030159183673469397
versicolor variance: 0.22081632653061237
virginica variance: 0.304587755102041
```

- 가설 수립: 귀무가설(H0)-3개 Species 간의 평균 Petal Length에 차이가 없다, 대립 가설(H1)-적어도 한 Species의 평균 Petal Length는 나머지와 다르다.
- ANOVA 결과: F=1180.16, p<0.001 로 귀무가설 기각

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

- 사후검정: Tukey HSD로 모든 쌍 간 유의미한 차이를 확인

Multiple Comparison of Means - Tukey HSD, FWER=0.50

group1	group2	meandiff	p-adj	lower	upper	reject
setosa	versicolor	2.798	0.0	2.7011	2.8949	True
setosa	virginica	4.09	0.0	3.9931	4.1869	True
versicolor	virginica	1.292	0.0	1.1951	1.3889	True

3) 결론
- 각 Species의 Petal Length는 “setosa < versicolor < virginica” 순으로 길다.
- 세 Species 간 Petal Length는 통계적으로 유의한 차이가 존재한다.

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 개발한다.

2) 주요 처리 과정
- 데이터 로드 및 기본 탐색: 사기거래(Class-1)의 비율이 0.1% 이하로 매우 적음을 확인

```
Class
0    0.998273
1    0.001727
Name: proportion, dtype: float64
```

- 샘플링: Class 0을 10,000건 무작위 추출

```
Class
0    0.953107
1    0.046893
Name: proportion, dtype: float64
```

- 전처리: Amount 변수를 표준화한 후 Amount_Scaled로 대체
- SMOTE 적용: 학습 데이터셋에서 Class 1을 오버샘플링

-> 필요한 이유: SMOTE는 소수 클래스(사기거래)의 synthetic 데이터를 생성해 다수 클래스와의 불균형을 안화하여 모델이 소수 클래스(사기거래)를 더 잘 학습할 수 있게 해줌

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

- 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

```
=== Classification Report ===
              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


Class 0 Recall: 0.9975, F1-score: 0.9960
Class 1 Recall: 0.8878, F1-score: 0.9158


Class 1 PR-AUC: 0.9537
```

3. 결론

- 목표 성능 (Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90)을 모두 충족하였다.
- 모델이 실제 사기 거래를 높은 비율로 탐지하면서도 False Positive를 최소화하여 전체적으로 우수한 분류 성능을 가졌다.
- PR-AUC가 0.95 이상으로 나온 점으로 봤을 때, 모델이 불균형 데이터 상황에서도 안정적으로 Precision-Recall trade-off를 처리했음을 보여주었다.
- 해당 모델은 실제 금융 사기 거래 탐지 시스템에 적용 가능할 만큼 신뢰성 있는 탐지 성능을 가질 수 있다.