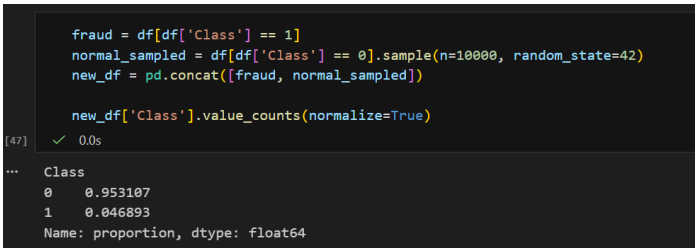
[5회차] ML 과제

iris 데이터셋을 활용해 다양한 통계 기법들을 실습해본 결과이다.

데이터 로드 및 기본 탐색

- creditcard.csv를 df로 불러와 정상 거래와 사기 거래 건수를 확인한 결과이다. 정상 거래는 284,315건, 사기 거래는 492건으로 확인된다.

샘플링



- 정상 거래를 10,000건만 무작위로 샘플링한 결과 이전에 0.1% 정도 되었던 사기 거래 건수 비율이 4%까지 상승하였다.

데이터 전처리 및 데이터 분할

```
from sklearn.preprocessing import StandardScaler
new_df['Amount_Scaled'] = StandardScaler().fit_transform(new_df[['Amount']])
new_df = new_df.drop('Amount', axis=1)

X = new_df.drop('Class', axis=1)
y = new_df['Class']
```

- Amount 변수를 StandardScaler를 이용해 표준화하였다.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, strati
fy=y)
```

```
print(f'학습 데이터 Class 비율: {y_train.value_counts(normalize=True)}')
print(f'테스트 데이터 Class 비율: {y_test.value_counts(normalize=True)}')
```

- y를 Class로 설정하고 train 데이터와 test 데이터를 8:2 비율로 나누었다.

SMOTE 적용

SMOTE를 사용하는 이유

- 신용카드 사기 탐지 데이터는 정상 거래에 비해 사기 거래 건수가 매우 적은 클래스 불균형 문제를 가지고 있다.
- SMOTE는 인접한 데이터들 사이의 값을 보간(Interpolation)하여 새로운 가상의 샘플을 생성한다. 이를 통해 오버피팅 위험을 줄이면서 모델이 사기 거래의 특징을 더 정교하게 학습할 수 있도록 돕는다.

모델 학습

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, average_precision_score

model = RandomForestClassifier(random_state=42)
model.fit(X_train_resampled, y_train_resampled)

y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:, 1]
```

- Resample 한 데이터를 RandomForestClassifier에 넣어서 모델을 학습시켰다.

최종 성능 평가

- Class 0(정상 거래)에 대해 Precision 0.99, Recall 1 , F1-score 1을 달성하여 과제 명세서에서 제시한 기준을 달성하였다.
- Class 0(사기 거래)에 대해 Precision 0.95, Recall 0.89 , F1-score 0.92을 달성하여 과제 명세서에서 제시한 기준을 달성하였다.
- PR-AUC Score는 0.9537로 0.90을 넘기는 데 성공하였다.