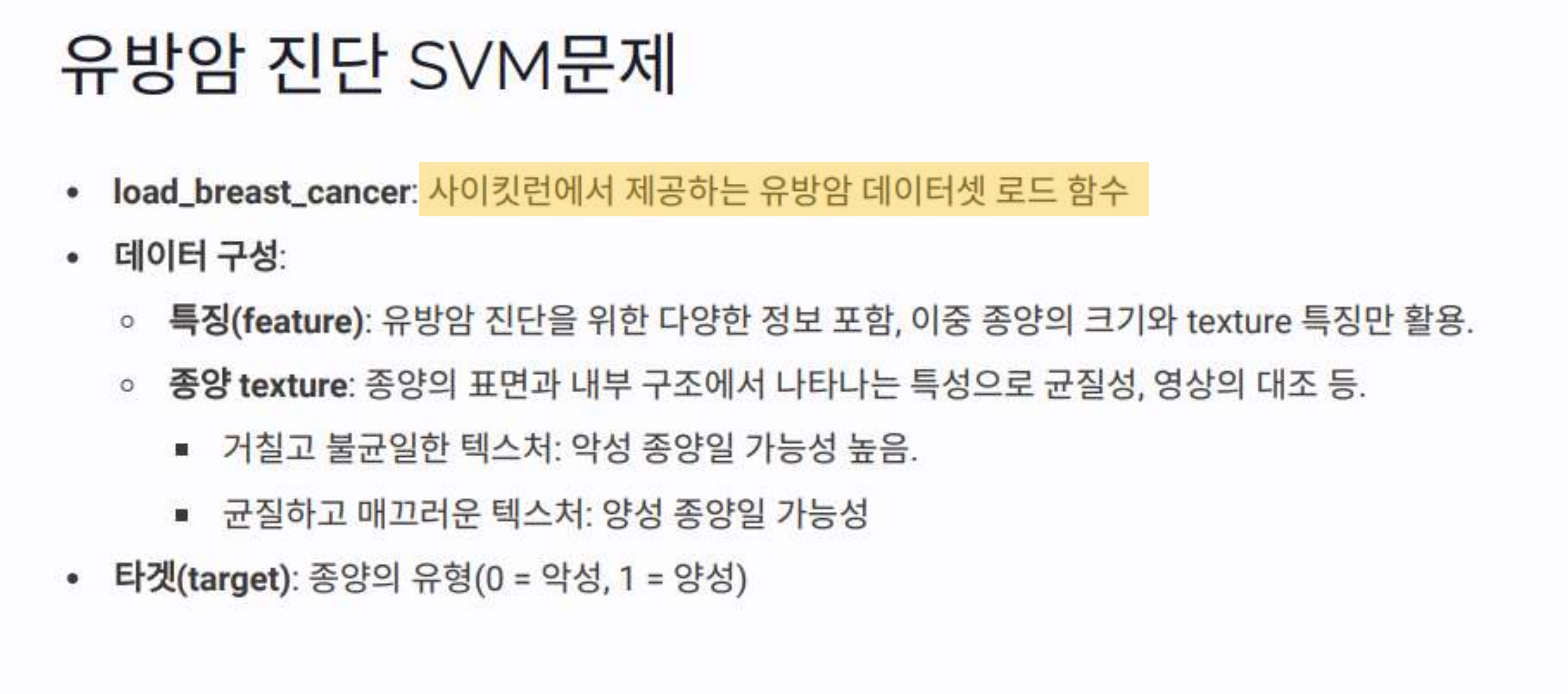
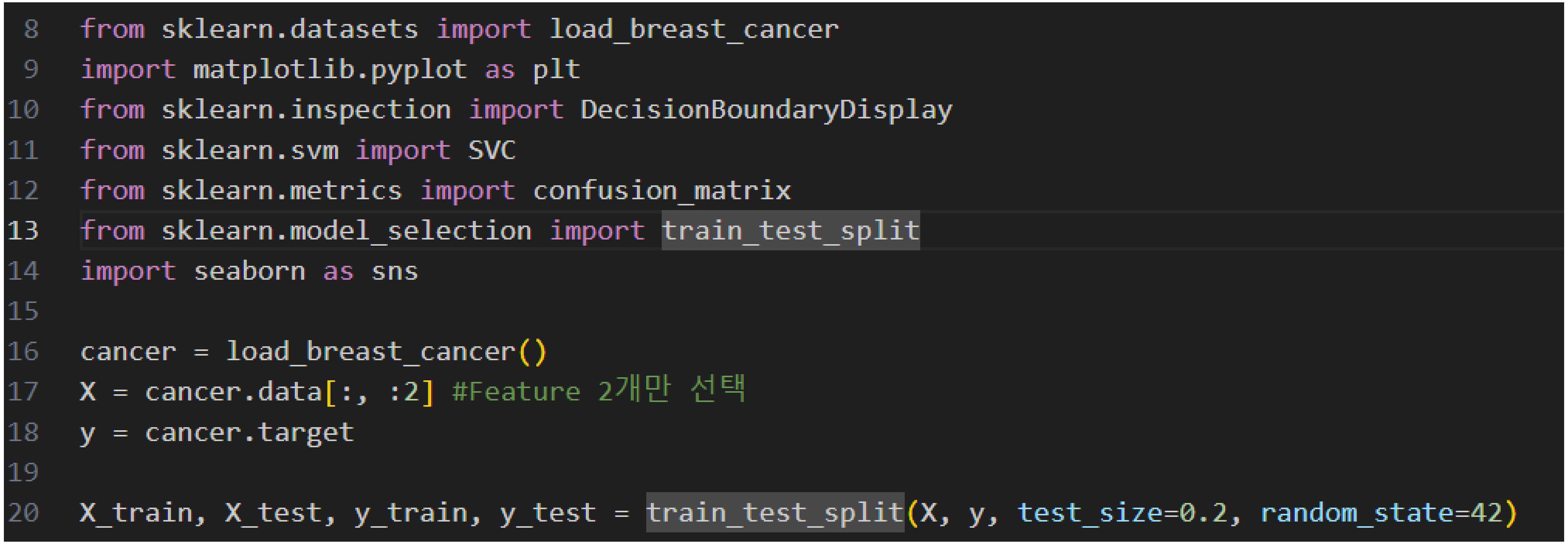
1. 유방암 진단 SVM 문제



유방암 진단 SVM문제

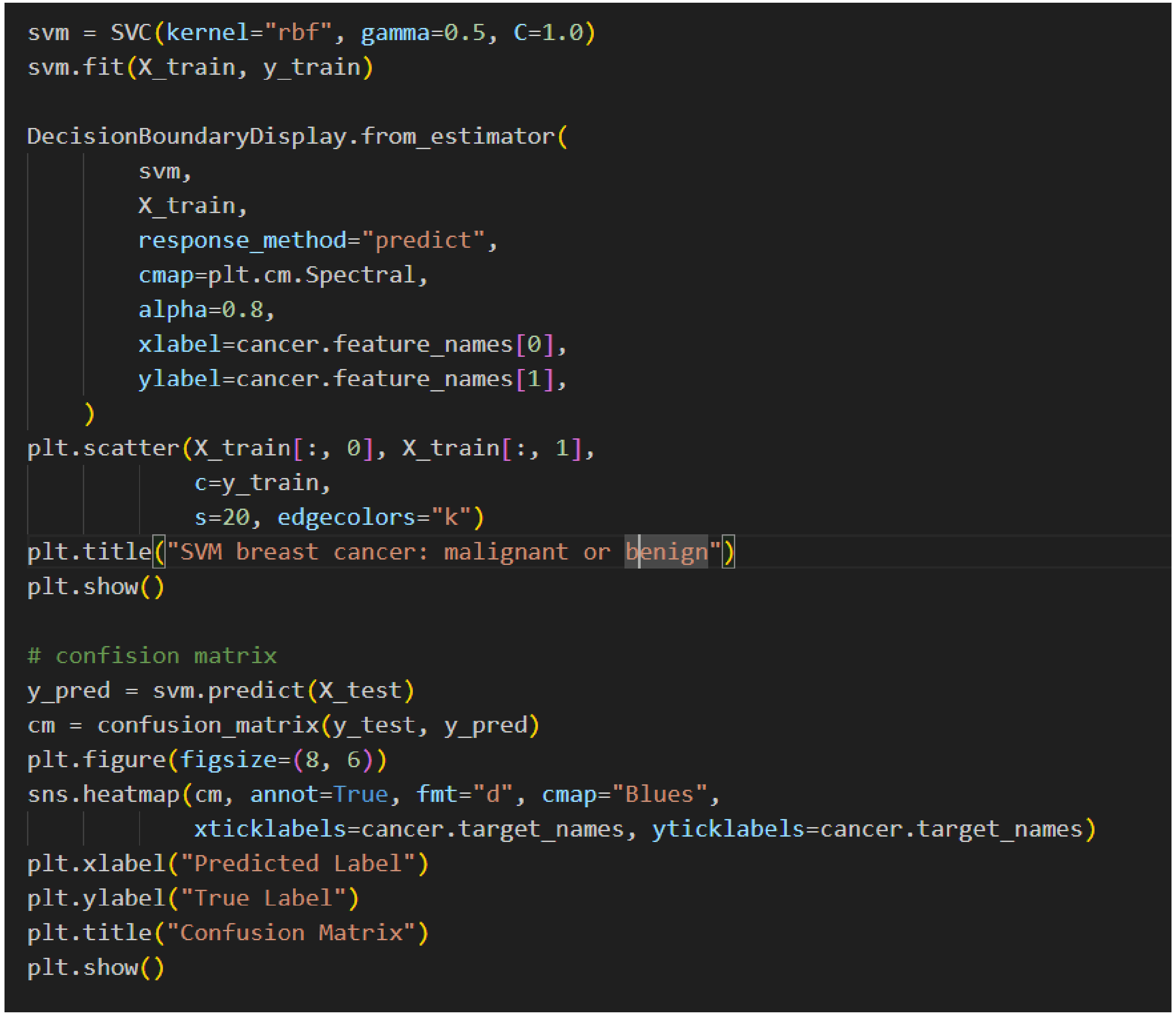
- **load_breast_cancer**: 사이킷런에서 제공하는 유방암 데이터셋 로드 함수
- **데이터 구성**:
    - **특징(feature)**: 유방암 진단을 위한 다양한 정보 포함, 이중 종양의 크기와 texture 특징만 활용.
    - **종양 texture**: 종양의 표면과 내부 구조에서 나타나는 특성으로 균질성, 영상의 대조 등.
        - 거칠고 불균일한 텍스처: 악성 종양일 가능성 높음.
        - 균질하고 매끄러운 텍스처: 양성 종양일 가능성
- **타겟(target)**: 종양의 유형(0 = 악성, 1 = 양성)

유방암 입력 데이터



```python
from sklearn.datasets import load_breast_cancer
import matplotlib.pyplot as plt
from sklearn.inspection import DecisionBoundaryDisplay
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
import seaborn as sns

cancer = load_breast_cancer()
X = cancer.data[:, :2] #Feature 2개만 선택
y = cancer.target

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

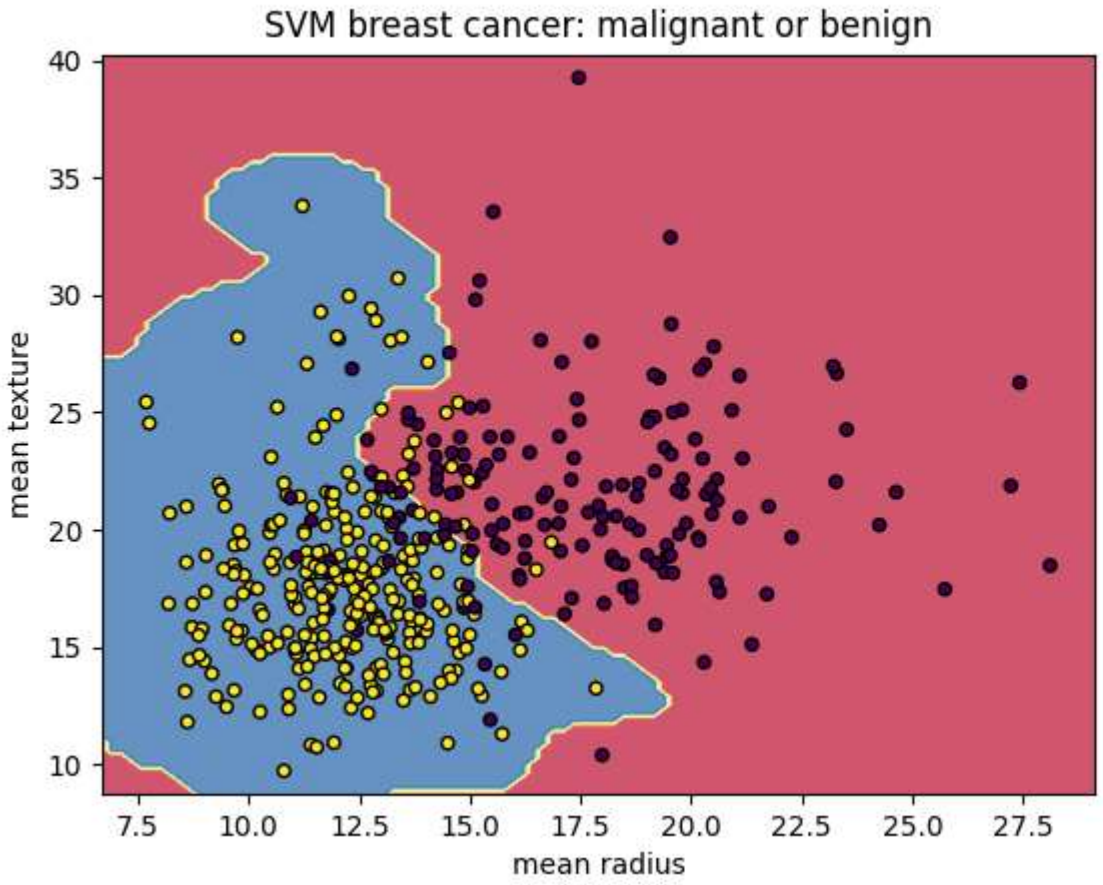
유방암 SVM모델 학습 및 시각화 코드

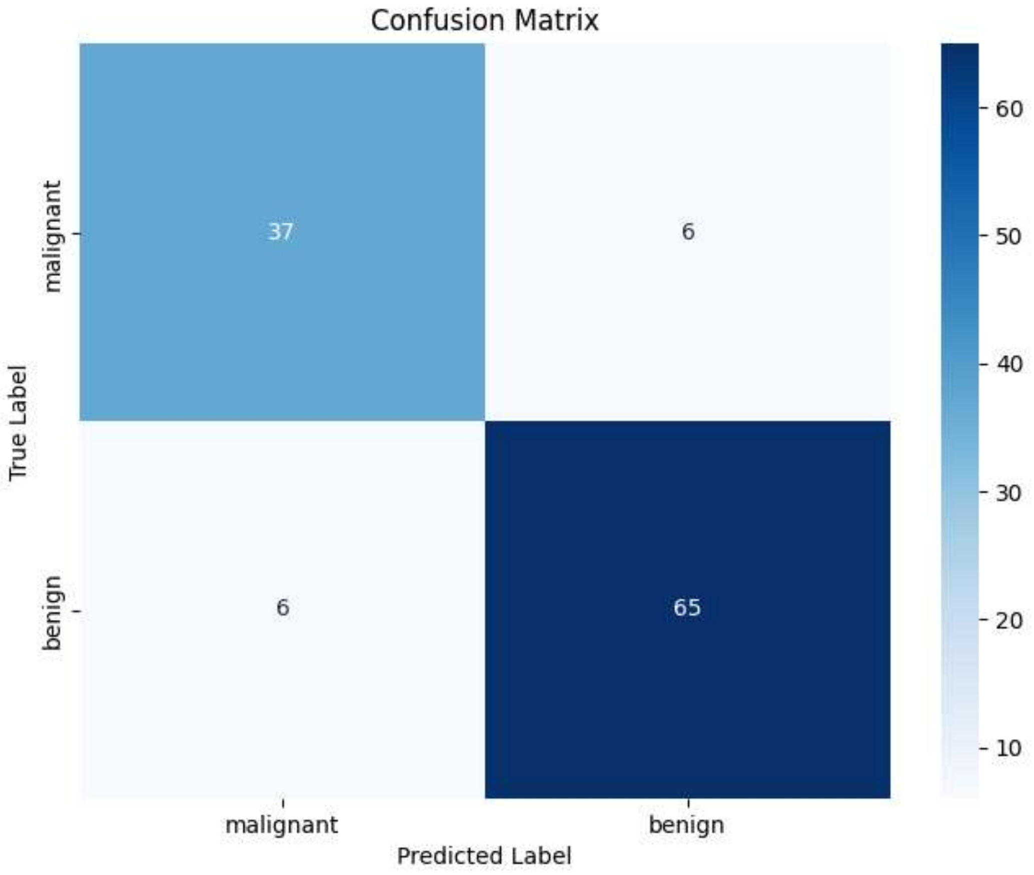


```python
svm = SVC(kernel="rbf", gamma=0.5, C=1.0)
svm.fit(X_train, y_train)

DecisionBoundaryDisplay.from_estimator(
        svm,
        X_train,
        response_method="predict",
        cmap=plt.cm.Spectral,
        alpha=0.8,
        xlabel=cancer.feature_names[0],
        ylabel=cancer.feature_names[1],
    )
plt.scatter(X_train[:, 0], X_train[:, 1],
            c=y_train,
            s=20, edgecolors="k")
plt.title("SVM breast cancer: malignant or benign")
plt.show()

# confision matrix
y_pred = svm.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
            xticklabels=cancer.target_names, yticklabels=cancer.target_names)
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.title("Confusion Matrix")
plt.show()
```

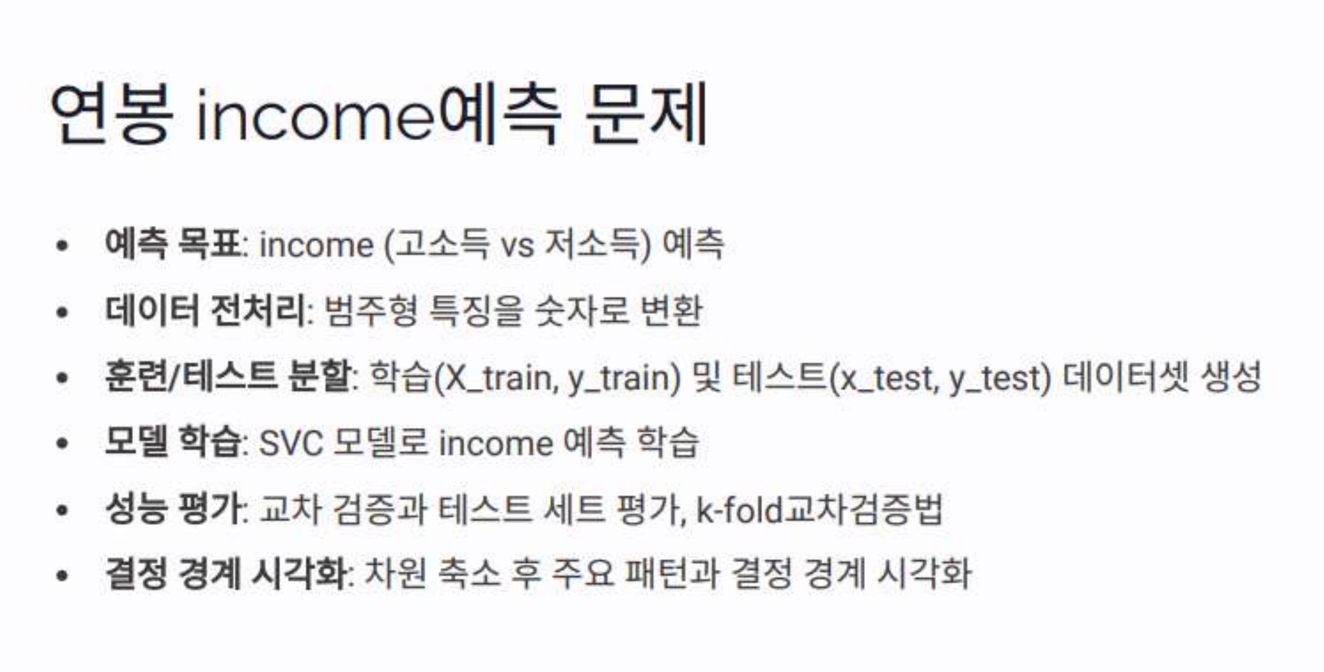
.

유방암 SVM 결과 시각화


SVM breast cancer: malignant or benign

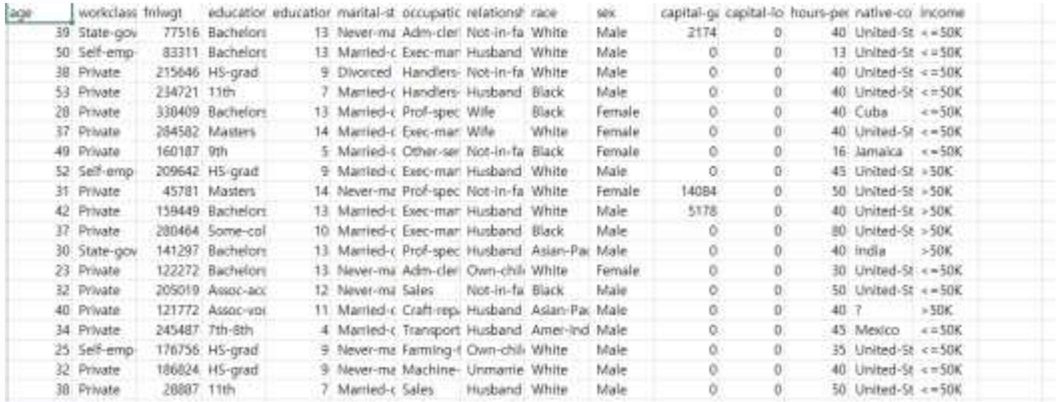

Confusion Matrix

2. 연봉 income 예측 문제



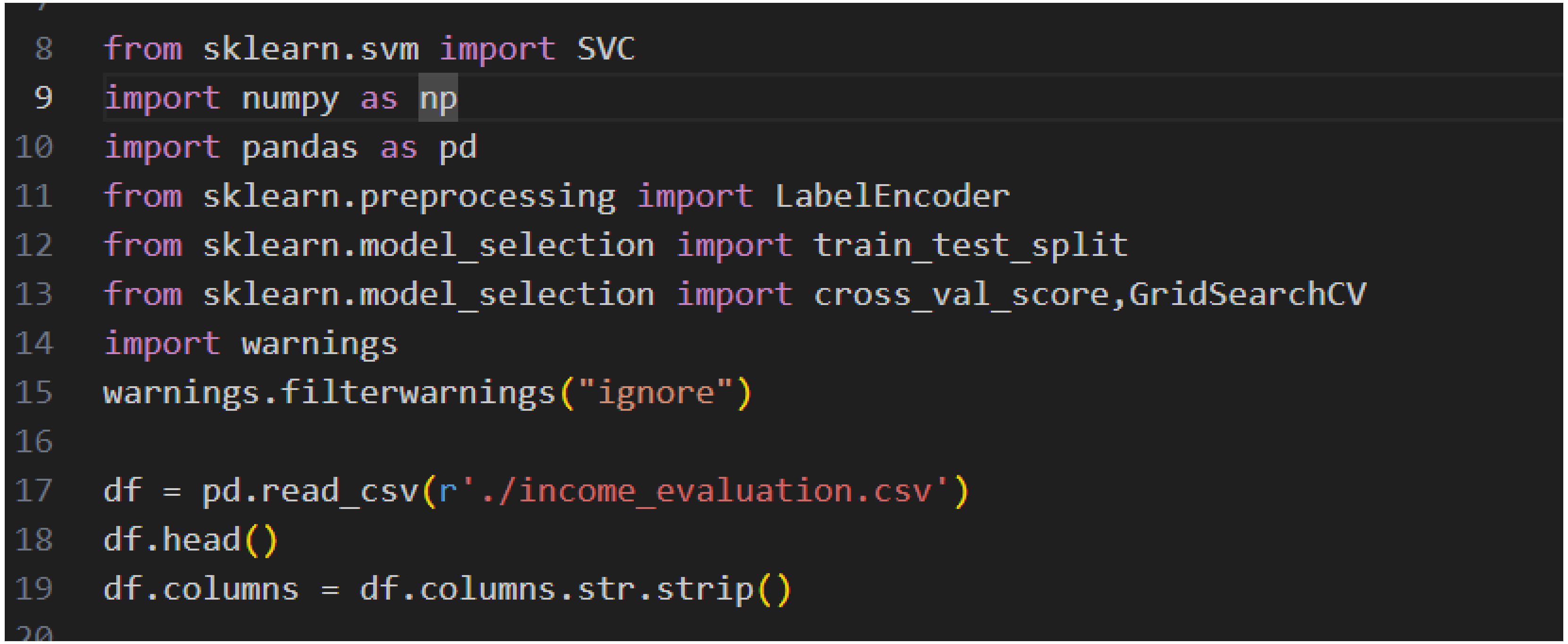
연봉 income예측 문제

- **예측 목표**: income (고소득 vs 저소득) 예측
- **데이터 전처리**: 범주형 특징을 숫자로 변환
- **훈련/테스트 분할**: 학습(X_train, y_train) 및 테스트(x_test, y_test) 데이터셋 생성
- **모델 학습**: SVC 모델로 income 예측 학습
- **성능 평가**: 교차 검증과 테스트 세트 평가, k-fold교차검증법
- **결정 경계 시각화**: 차원 축소 후 주요 패턴과 결정 경계 시각화



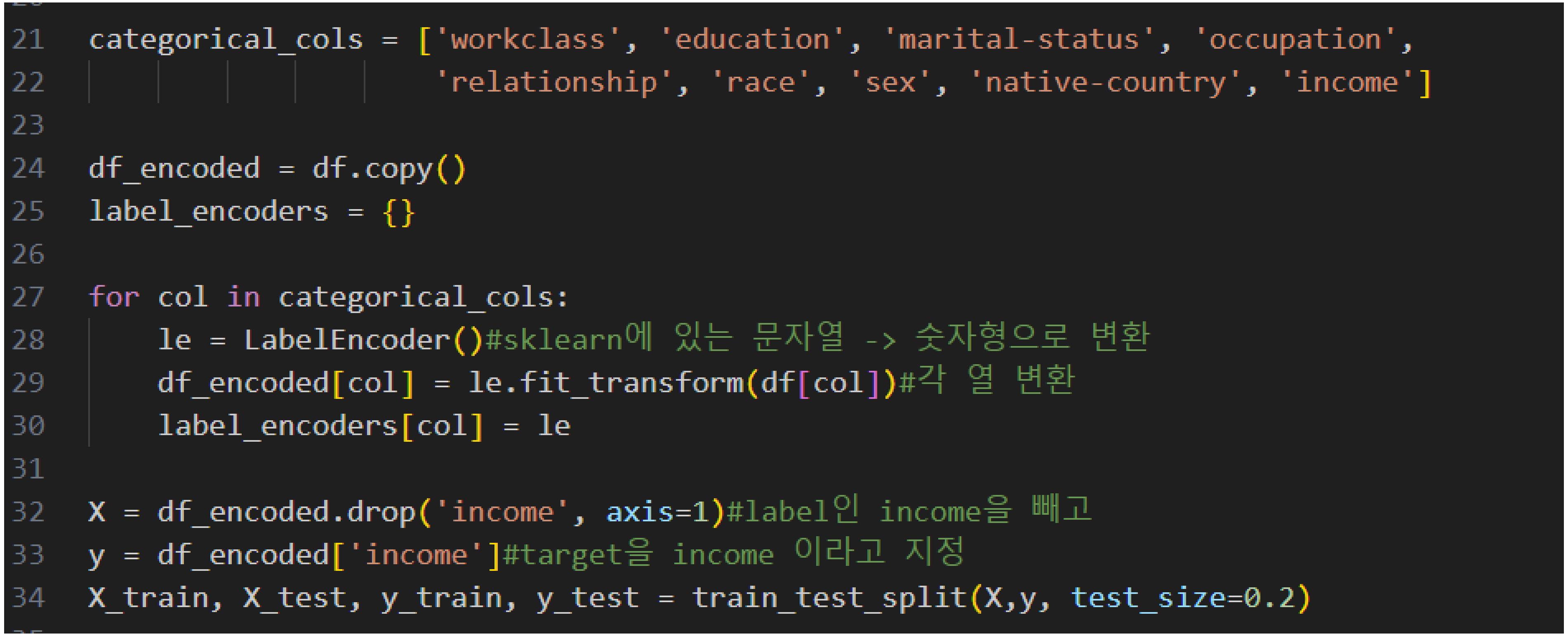
age	workclass	fnlwgt	educatior	educatior	marital-st	occupatic	relationsh	race	sex	capital-ga	capital-lo	hours-per	native-co	income
39	State-gov	77516	Bachelors	13	Never-ma	Adm-cler	Not-in-fa	White	Male	2174	0	40	United-St	<=50K
50	Self-emp-	83311	Bachelors	13	Married-c	Exec-mar	Husband	White	Male	0	0	13	United-St	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-	Not-in-fa	White	Male	0	0	40	United-St	<=50K
53	Private	234721	11th	7	Married-c	Handlers-	Husband	Black	Male	0	0	40	United-St	<=50K
28	Private	338409	Bachelors	13	Married-c	Prof-spec	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-c	Exec-mar	Wife	White	Female	0	0	40	United-St	<=50K
49	Private	160187	9th	5	Married-s	Other-ser	Not-in-fa	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp	209642	HS-grad	9	Married-c	Exec-mar	Husband	White	Male	0	0	45	United-St	>50K
31	Private	45781	Masters	14	Never-ma	Prof-spec	Not-in-fa	White	Female	14084	0	50	United-St	>50K
42	Private	159449	Bachelors	13	Married-c	Exec-mar	Husband	White	Male	5178	0	40	United-St	>50K
37	Private	280464	Some-col	10	Married-c	Exec-mar	Husband	Black	Male	0	0	80	United-St	>50K
30	State-gov	141297	Bachelors	13	Married-c	Prof-spec	Husband	Asian-Pac	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-ma	Adm-cler	Own-chil	White	Female	0	0	30	United-St	<=50K
32	Private	205019	Assoc-acc	12	Never-ma	Sales	Not-in-fa	Black	Male	0	0	50	United-St	<=50K
40	Private	121772	Assoc-voc	11	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-c	Transport	Husband	Amer-Ind	Male	0	0	45	Mexico	<=50K
25	Self-emp	176756	HS-grad	9	Never-ma	Farming-f	Own-chil	White	Male	0	0	35	United-St	<=50K
32	Private	186824	HS-grad	9	Never-ma	Machine-	Unmarrie	White	Male	0	0	40	United-St	<=50K
38	Private	28887	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K

SVM 구현을 위한 입력데이터



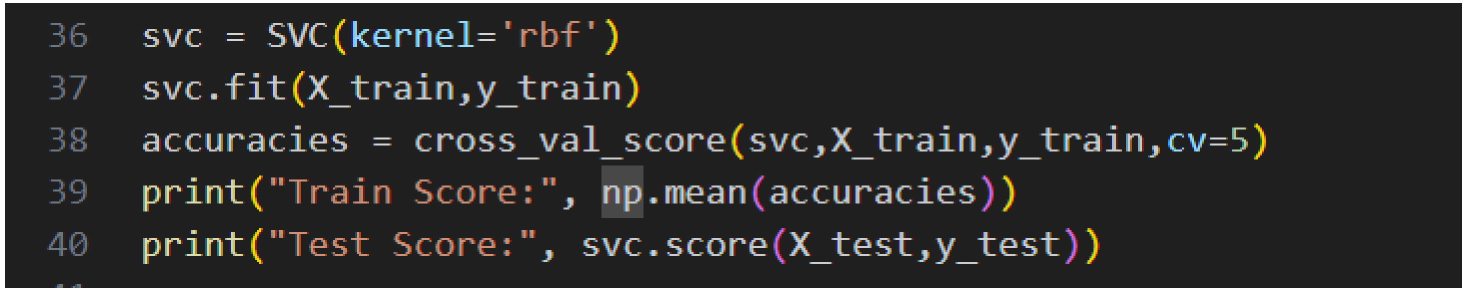
```python
from sklearn.svm import SVC
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score,GridSearchCV
import warnings
warnings.filterwarnings("ignore")

df = pd.read_csv(r'./income_evaluation.csv')
df.head()
df.columns = df.columns.str.strip()
```

데이터 전처리



```python
categorical_cols = ['workclass', 'education', 'marital-status', 'occupation',
                    'relationship', 'race', 'sex', 'native-country', 'income']

df_encoded = df.copy()
label_encoders = {}

for col in categorical_cols:
    le = LabelEncoder()#sklearn에 있는 문자열 -> 숫자형으로 변환
    df_encoded[col] = le.fit_transform(df[col])#각 열 변환
    label_encoders[col] = le

X = df_encoded.drop('income', axis=1)#label인 income을 빼고
y = df_encoded['income']#target을 income 이라고 지정
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2)
```

SVM모델 학습 코드



```python
svc = SVC(kernel='rbf')
svc.fit(X_train,y_train)
accuracies = cross_val_score(svc,X_train,y_train,cv=5)
print("Train Score:", np.mean(accuracies))
print("Test Score:", svc.score(X_test,y_test))
```