9/28 수업대체 과제 Decision Tree

학번	32171373	이름	노병우

제공된 PimaIndiansDiabetes dataset을 가지고 Classification 모델을 개발 하고자 한다.

(마지막의 diabetes 컬럼이 class label 임)

다음의 문제를 해결하기 위한 Python 코드를 작성하고 실행 결과를 캡쳐하여 보이시오

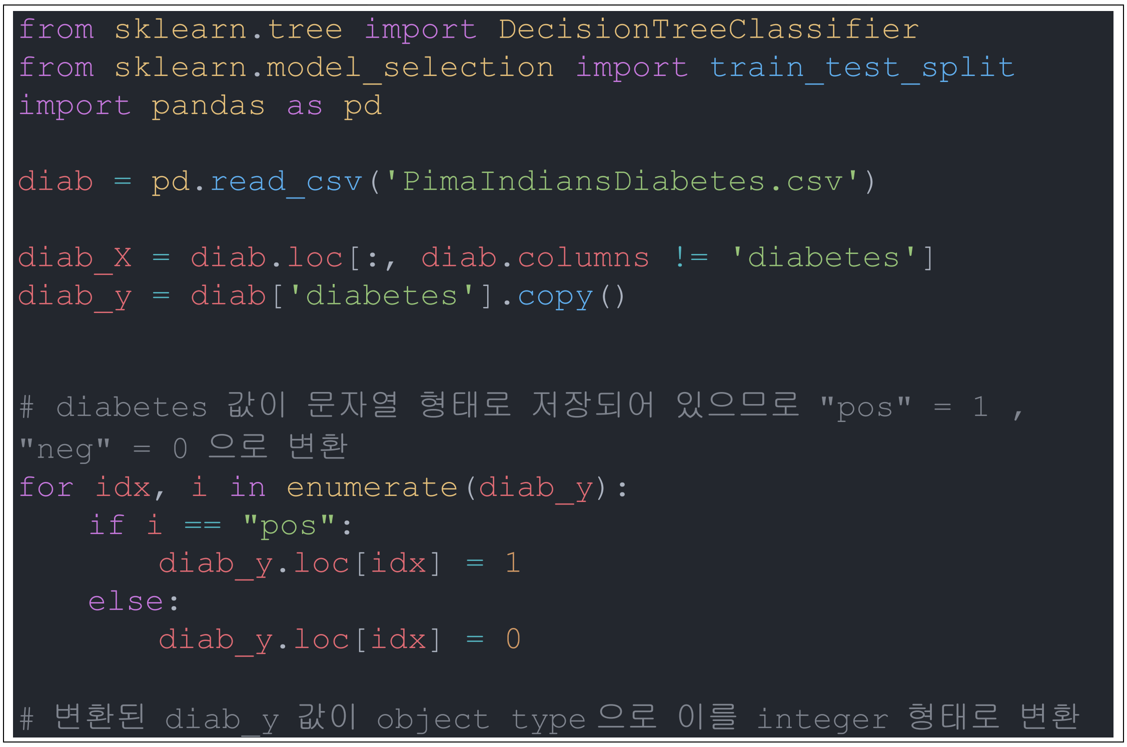
(실행 결과가 길 경우에는 마지막 10줄 정도만 캡쳐)

Q1 (2점) scikit-learn에서 제공하는 DecisionTree 알고리즘을 이용하여 모델을 생성하고

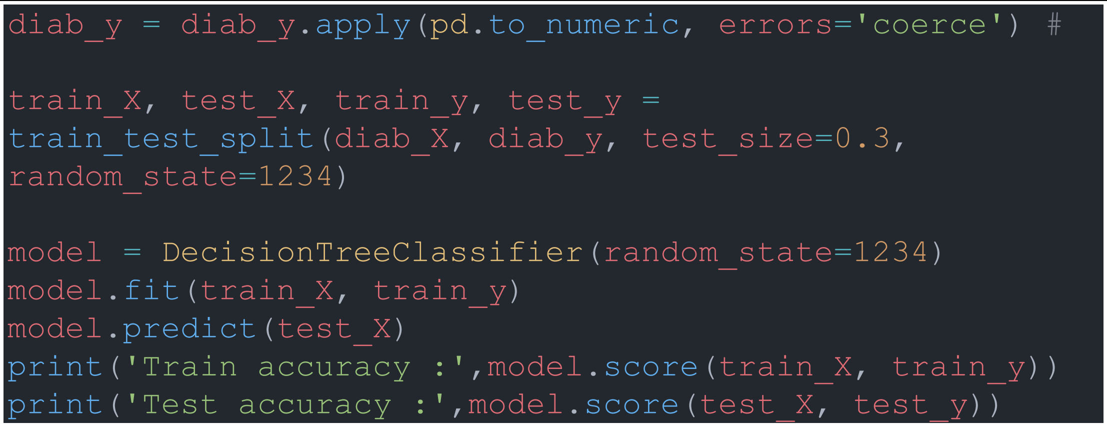
training accuracy 와 test accuracy를 보이시오

* 강의 ppt 21~23의 코드를 활용

Source code :

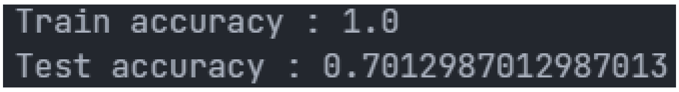


```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
import pandas as pd

diab = pd.read_csv('PimaIndiansDiabetes.csv')

diab_X = diab.loc[:, diab.columns != 'diabetes']
diab_y = diab['diabetes'].copy()


# diabetes 값이 문자열 형태로 저장되어 있으므로 "pos" = 1 ,
"neg" = 0 으로 변환
for idx, i in enumerate(diab_y):
    if i == "pos":
        diab_y.loc[idx] = 1
    else:
        diab_y.loc[idx] = 0

# 변환된 diab_y 값이 object type 으로 이를 integer 형태로 변환
```



```
diab_y = diab_y.apply(pd.to_numeric, errors='coerce') #

train_X, test_X, train_y, test_y =
train_test_split(diab_X, diab_y, test_size=0.3,
random_state=1234)

model = DecisionTreeClassifier(random_state=1234)
model.fit(train_X, train_y)
model.predict(test_X)
print('Train accuracy :',model.score(train_X, train_y))
print('Test accuracy :',model.score(test_X, test_y))
```

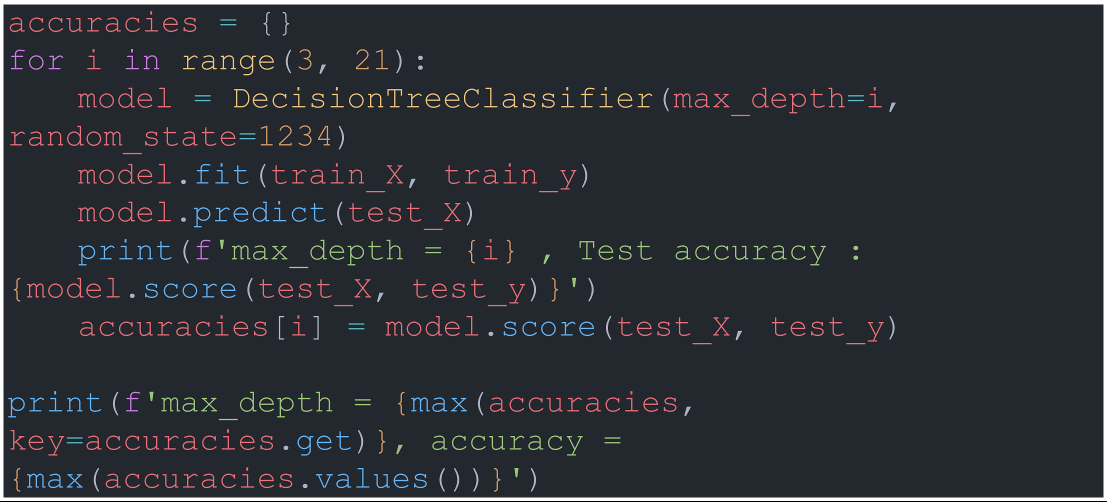
실행화면 캡처:



```
Train accuracy : 1.0
Test accuracy : 0.7012987012987013
```

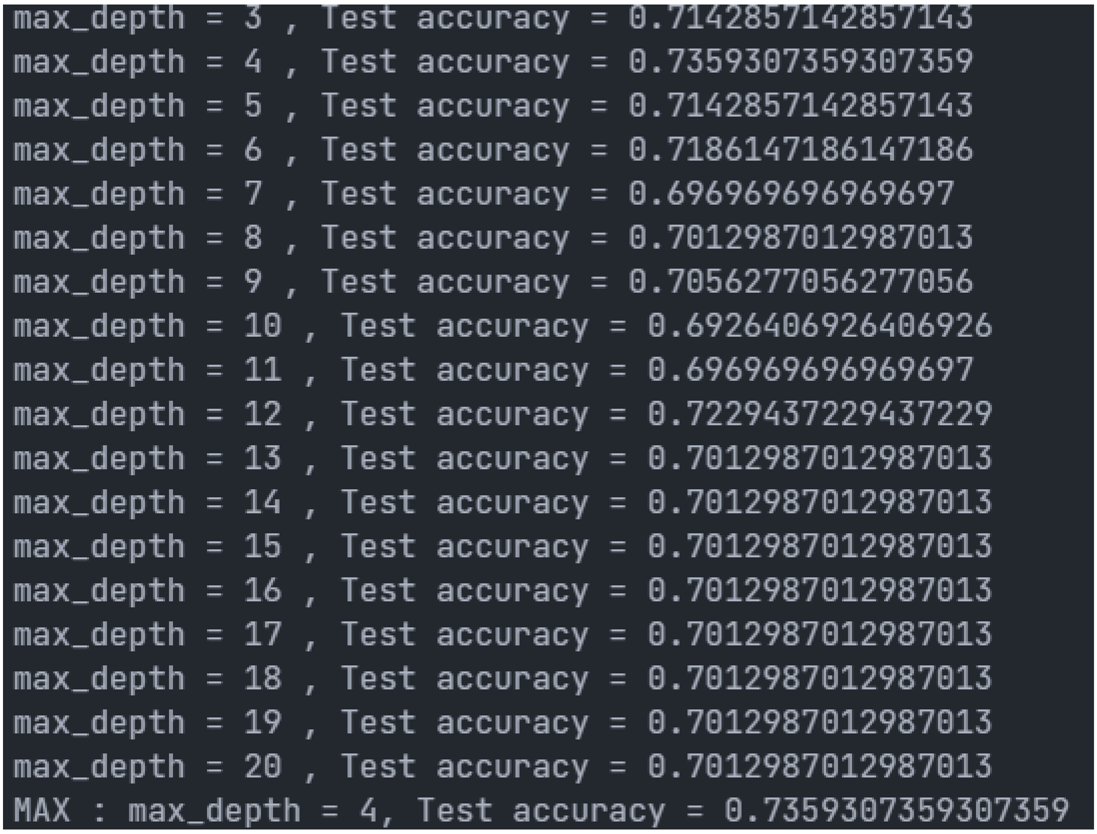
Q2 (4점) Q1 모델에서 DecisionTreeClassifier() 의 파라메터중 하나인 **max_depth**를 3~20까지 차례로 변화시킬 때, training/test accuracy 의 변화를 보이시오. test accuracy를 가장 크게하는 max_depth의 값과 이때의 test accuracy를 출력하시오.

Source code :



```
accuracies = {}
for i in range(3, 21):
    model = DecisionTreeClassifier(max_depth=i,
random_state=1234)
    model.fit(train_X, train_y)
    model.predict(test_X)
    print(f'max_depth = {i} , Test accuracy :
{model.score(test_X, test_y)}')
    accuracies[i] = model.score(test_X, test_y)

print(f'max_depth = {max(accuracies,
key=accuracies.get)}, accuracy =
{max(accuracies.values())}')
```

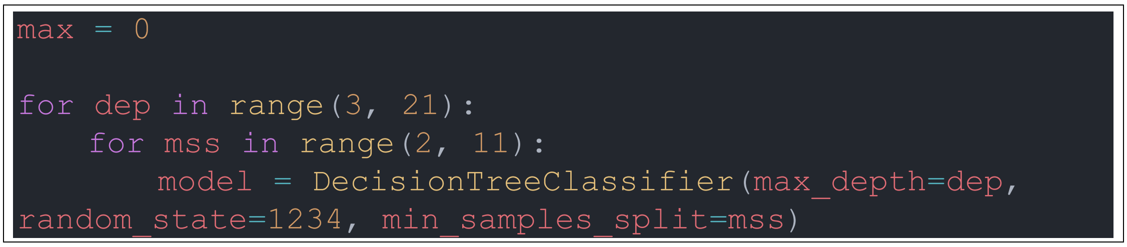
실행화면 캡처:



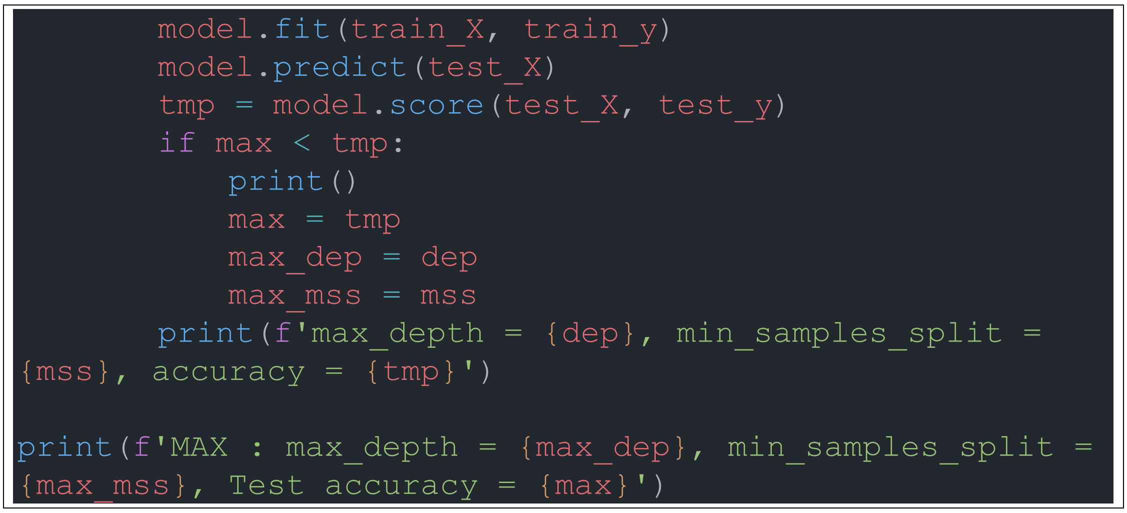
```
max_depth = 3 , Test accuracy = 0.7142857142857143
max_depth = 4 , Test accuracy = 0.7359307359307359
max_depth = 5 , Test accuracy = 0.7142857142857143
max_depth = 6 , Test accuracy = 0.7186147186147186
max_depth = 7 , Test accuracy = 0.696969696969697
max_depth = 8 , Test accuracy = 0.7012987012987013
max_depth = 9 , Test accuracy = 0.7056277056277056
max_depth = 10 , Test accuracy = 0.6926406926406926
max_depth = 11 , Test accuracy = 0.696969696969697
max_depth = 12 , Test accuracy = 0.7229437229437229
max_depth = 13 , Test accuracy = 0.7012987012987013
max_depth = 14 , Test accuracy = 0.7012987012987013
max_depth = 15 , Test accuracy = 0.7012987012987013
max_depth = 16 , Test accuracy = 0.7012987012987013
max_depth = 17 , Test accuracy = 0.7012987012987013
max_depth = 18 , Test accuracy = 0.7012987012987013
max_depth = 19 , Test accuracy = 0.7012987012987013
max_depth = 20 , Test accuracy = 0.7012987012987013
MAX : max_depth = 4, Test accuracy = 0.7359307359307359
```

Q2 (4점) Q1 모델에서 DecisionTreeClassifier() 의 파라메터중 하나인 **max_depth**를 3~20까지 차례로 변화시키고 **min_samples_split** 은 2~10까지 변화시킬 때 training/test accuracy 의 변화를 보이시오. test accuracy 의 값을 가장 크게 하는 두 파라메터의 조합과 이때의 test accuracy를 출력하시오.

Source code :

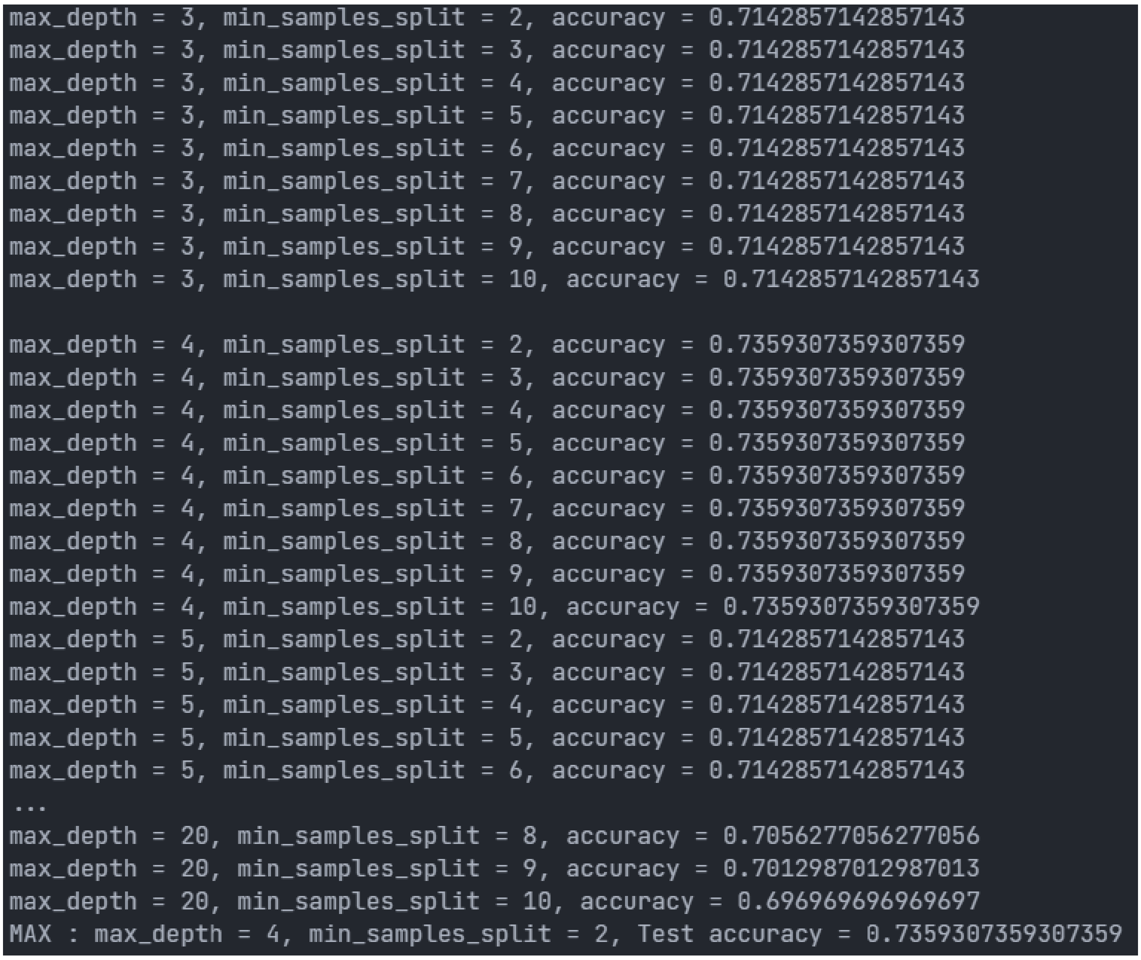


```
max = 0

for dep in range(3, 21):
    for mss in range(2, 11):
        model = DecisionTreeClassifier(max_depth=dep,
random_state=1234, min_samples_split=mss)
```



```
        model.fit(train_X, train_y)
        model.predict(test_X)
        tmp = model.score(test_X, test_y)
        if max < tmp:
            print()
            max = tmp
            max_dep = dep
            max_mss = mss
        print(f'max_depth = {dep}, min_samples_split = {mss}, accuracy = {tmp}')

print(f'MAX : max_depth = {max_dep}, min_samples_split = {max_mss}, Test accuracy = {max}')
```

실행화면 캡처:



```
max_depth = 3, min_samples_split = 2, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 3, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 4, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 5, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 6, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 7, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 8, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 9, accuracy = 0.7142857142857143
max_depth = 3, min_samples_split = 10, accuracy = 0.7142857142857143

max_depth = 4, min_samples_split = 2, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 3, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 4, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 5, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 6, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 7, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 8, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 9, accuracy = 0.7359307359307359
max_depth = 4, min_samples_split = 10, accuracy = 0.7359307359307359
max_depth = 5, min_samples_split = 2, accuracy = 0.7142857142857143
max_depth = 5, min_samples_split = 3, accuracy = 0.7142857142857143
max_depth = 5, min_samples_split = 4, accuracy = 0.7142857142857143
max_depth = 5, min_samples_split = 5, accuracy = 0.7142857142857143
max_depth = 5, min_samples_split = 6, accuracy = 0.7142857142857143
...
max_depth = 20, min_samples_split = 8, accuracy = 0.7056277056277056
max_depth = 20, min_samples_split = 9, accuracy = 0.7012987012987013
max_depth = 20, min_samples_split = 10, accuracy = 0.696969696969697
MAX : max_depth = 4, min_samples_split = 2, Test accuracy = 0.7359307359307359
```