12주차 과제

Classifier

담당교수님: 윤장혁 교수님

과목명: Data Analytics

이름: 박민성

학번: 201611145

전공: 산업공학과

제출일: 2020.06.05

Object

- 다양한 binary classifier를 학습하여 결과를 확인
  - ✓ 여러 classifier모델을 테스트해보고 결과를 함께 첨부 (2개 이상)
  - ✓ 최종적으로 선택한 모델은 test accuracy 기준 55% 이상을 만족

Classifier Model 1 - Logistic Regression Model

C: > Users > minisong > Desktop > 12주차 2.py > ...

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
df=pd.read_csv('./data_week12.csv')
#sns.countplot(x='Class variable (0 or 1).', data=df)
#plt.show()
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
Y=df['Class variable (0 or 1).']
X=df[['Number of times pregnant.','Plasma glucose concentration a 2 hours
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3)
log_clf = LogisticRegression()
log_clf.fit(X_train,Y_train)
print(log_clf.score(X_test, Y_test)*100,"%")
```

```
76.62337662337663 %
```

전체 자료 중 70%를 학습시키고 30%로 테스트를 진행한 결과 정확도는 약 76.6%가 나왔다.

Classifier Model 2 – Naïve Bayes Model

> Users > minisong > Desktop > 12주차 1.py > ...

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
df=pd.read_csv('./data_week12.csv')
Y=df['Class variable (0 or 1).']
X=df[['Number of times pregnant.','Plasma glucose concentration a 2 hours
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3)
model = GaussianNB()
model.fit(X_train,Y_train)
accuracy = model.score(X_test,Y_test)*100
print(accuracy,"%")
```

```
77.05627705627705 %
```

마찬가지로 전체 자료 중 70%를 학습시키고 30%로 테스트를 진행한 결과 정확도는 약 77.1%가 나왔다.

Final Selection Model – Naïve Bayes Model

두 모델 다 test accuracy 가 55% 이상으로 조건에 부합하지만 최종적으로는 정확도가 조금 더 높은 Naïve Bayes Model을 선정하기로 하였다.