Seoul National University
College of Engineering
Department of Naval Architecture and Ocean Engineering
1, Gwanak-ro, Gwanak-gu, Seoul 151-744, Korea

Fall 2020

머신러닝

PA # 4

Instructor name	김태완 교수님
Student name	이용준
Department	조선해양공학과
Student ID	2015-19595
Submission date	2020.11.20

Contents

1. Problem Definition

Types	Column	Explanation
INPUT	Age	환자의 나이
	Sex	환자의 성별
	BP	환자의 혈압(blood pressure)
	Cholesterol	환자의 콜레스테롤 수치
	Na_to_K	혈액 내 나트륨 대 칼륨 비율
OUTPUT	Drug	처방 약의 종류

1. 주어진 input data와 output의 관련성을 데이터 시각화 및 분석하기
2. KNN 모델을 이용하여 주어진 input에 따른 output (처방 약) 결과 출력하기
3. SVM 모델을 이용하여 주어진 input에 따른 output (처방 약) 결과 출력하기
4. Decision Tree 모델을 이용하여 주어진 input에 따른 output (처방 약) 결과 출력하기

2. Problem Analysis

1. 데이터 분석 및 시각화를 위해 pandas 라이브러리를 활용하여 데이터를 선택적으로 정리하고 분류한다. 그리고 이를 각 input 항목 별로 plot한다. Plot의 형태는 막대그래프로 한다. 연속적 데이터인 Age, Na_to_K는 구간을 설정하여 분류해주고 개수를 세는 방식으로 분석한다.

2. sklearn 라이브러리를 활용하여 KNN, SVM, Decision Tree 모델을 만들고 학습하여 테스트를 진행한다. 테스트 정확도를 출력한다.

3. Code Explanation

프로젝트에 사용된 함수에 관한 설명이다. 문제 별 사용한 함수는 동일하다.

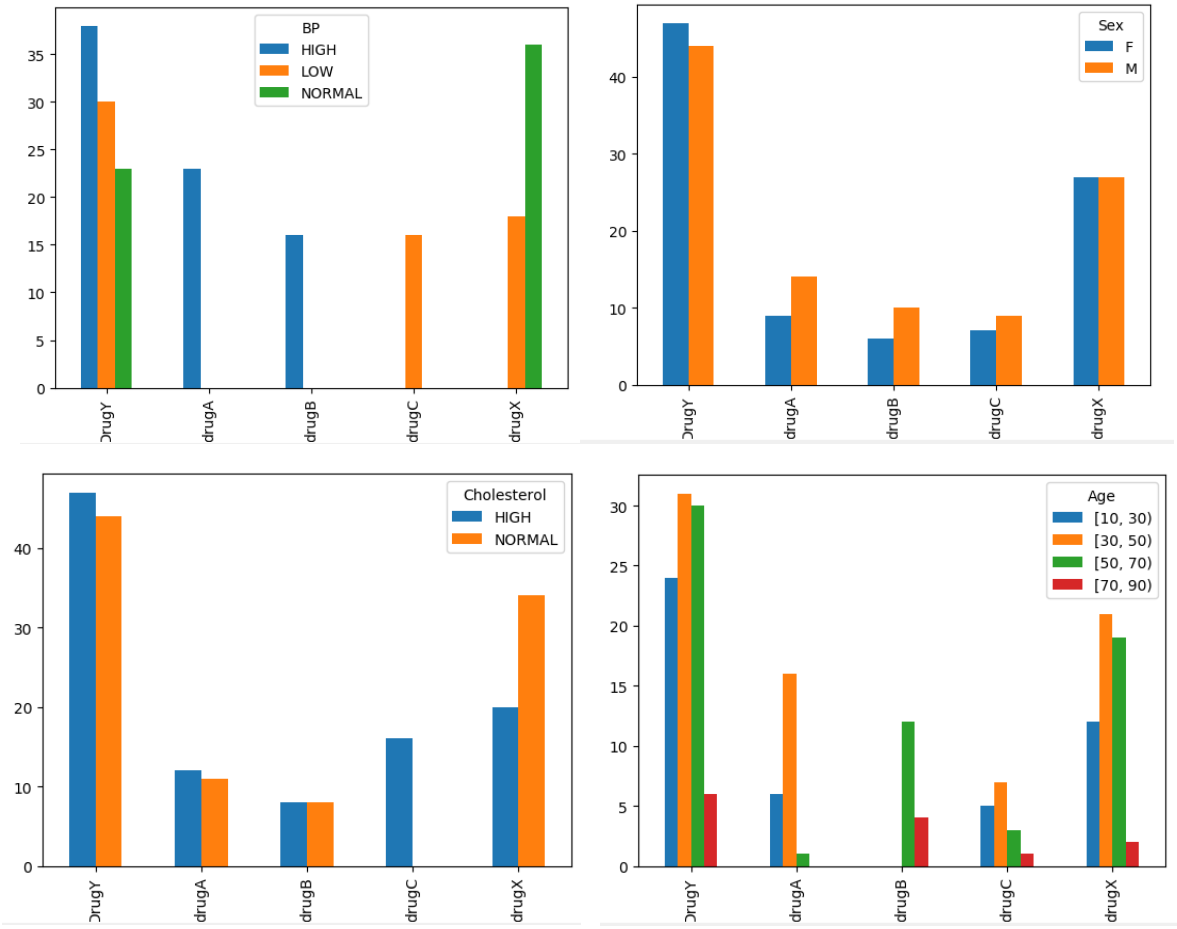
normalized(x)		
Parameter	Type	Explanation
x	array	정규화 할 input array
data	dataframe	csv파일의 내용을 담고 있는 dataframe

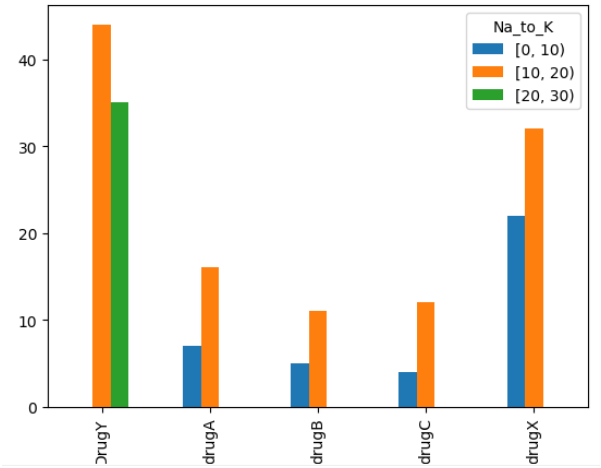
preprocessing(data)		
Parameter	Type	Explanation
data	dataframe	csv파일의 내용을 담고 있는 dataframe

4. Result & Conclusion

4.1 Result

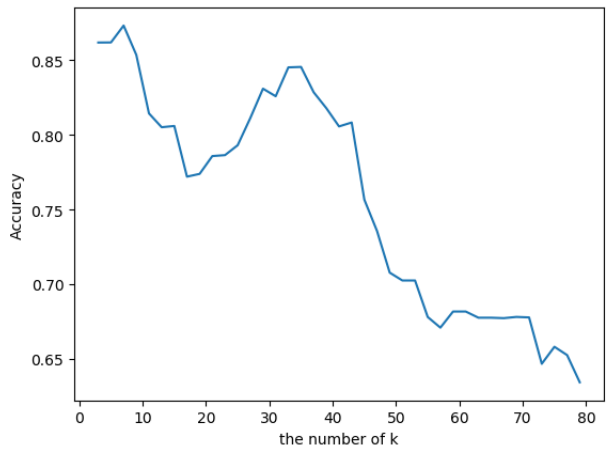
1번 문제 : 각 input data별로 데이터를 분석하기 위하여 시각화를 해봤다. 다음과 같다.





분석 결과 Sex 데이터는 남녀의 차이가 없어 처방 약을 결정하는 데에 무의미하다고 볼 수 있었다. Cholesterol 데이터도 거의 비슷했지만 차이가 있는 곳이 있었고, 다른 데이터들도 차이를 보였다.

2번 문제 : KNN 모델을 만들고 학습을 해봤다. Input data로는 차이가 뚜렷했던 BP, Na_to_K, Age 3개로만 주었다. 차이가 거의 없던 Sex와 Cholesterol은 제외시켰다. 실제로 정확도 비교 결과 이 두 데이터를 제외했을 때 정확도가 가장 높았다. 최적의 k를 찾기 위해 교차검증을 해보았다.



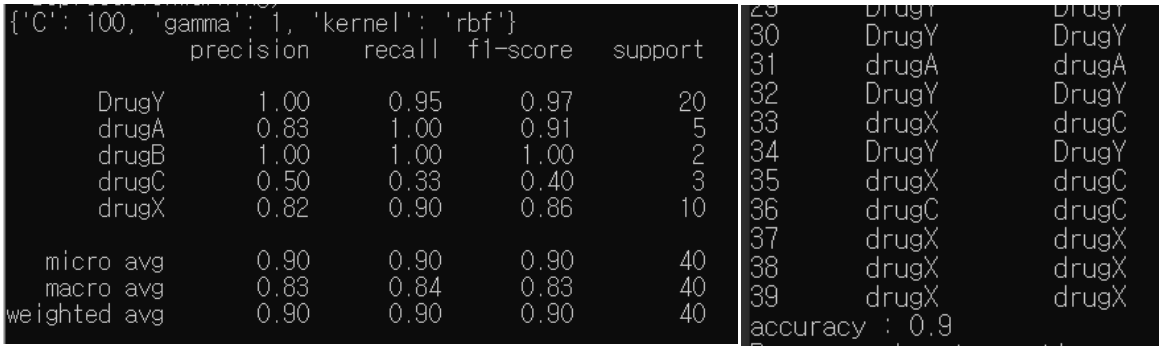

```
The best number of k : 7
   prediction ground_truth
0       drugX         drugX
1       drugB         drugB
2       drugX         drugX
3       DrugY         DrugY
4       DrugY         DrugY
5       DrugY         DrugY
6       DrugY         DrugY
7       drugX         drugX
8       DrugY         DrugY
9       drugA         drugA
10      drugX         drugX
11      drugC         drugX
12      DrugY         DrugY
13      DrugY         DrugY
14      drugA         drugA
15      drugB         DrugY
16      drugA         drugA
17      DrugY         DrugY
18      DrugY         DrugY
19      DrugY         DrugY
20      DrugY         DrugY
21      drugX         drugX
22      drugC         drugX
23      DrugY         DrugY
24      DrugY         DrugY
25      DrugY         DrugY
26      drugB         drugB
27      drugA         drugA
28      DrugY         DrugY
29      DrugY         DrugY
30      drugB         DrugY
31      drugA         drugA
32      drugX         DrugY
33      drugX         drugC
34      DrugY         DrugY
35      drugC         drugC
36      drugC         drugC
37      drugX         drugX
38      drugX         drugX
39      drugX         drugX
accuracy : 0.85
```

가시화 결과 최적 k는 7로 나타났으나 그 부근의 k를 모두 테스트해보고 결과를 정리했다.

K	Test Accuracy
3	0.825
5	0.9
7	0.85
9	0.825
11	0.825

비교 결과, 테스트 데이터에 대한 정확도가 가장 높은 것은 k=5였다.

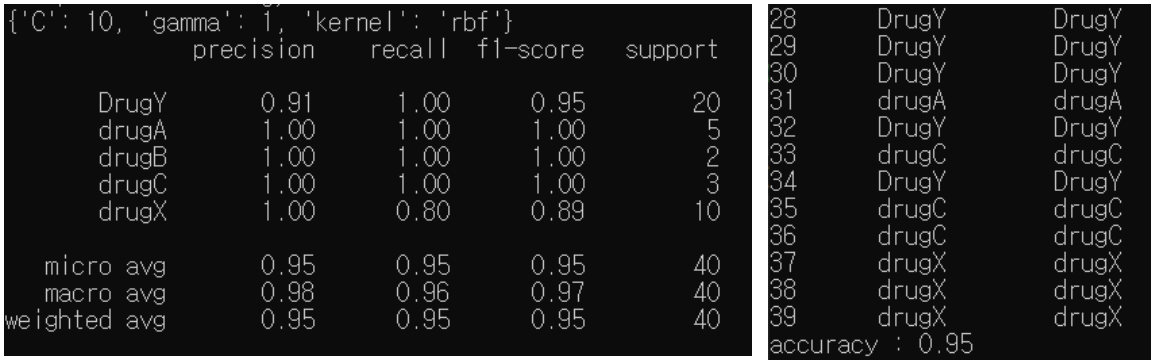
3번 문제 : SVM 모델을 만들고 학습시켰다. Input 데이터는 동일하게 3가지 항목(BP, Na_to_K, Age)만 사용하였다. sklearn의 gridsearch를 활용하여 최적의 파라미터를 찾고 학습한 뒤, 테스트 데이터로 테스트를 해보았다. 테스트 결과 정확도는 0.9였다.



```
{'C': 100, 'gamma': 1, 'kernel': 'rbf'}
              precision    recall  f1-score   support

       DrugY       1.00      0.95      0.97        20
       drugA       0.83      1.00      0.91         5
       drugB       1.00      1.00      1.00         2
       drugC       0.50      0.33      0.40         3
       drugX       0.82      0.90      0.86        10

   micro avg       0.90      0.90      0.90        40
   macro avg       0.83      0.84      0.83        40
weighted avg       0.90      0.90      0.90        40
```

```
30      DrugY      DrugY
31      drugA      drugA
32      DrugY      DrugY
33      drugX      drugC
34      DrugY      DrugY
35      drugX      drugC
36      drugC      drugC
37      drugX      drugX
38      drugX      drugX
39      drugX      drugX
accuracy : 0.9
```

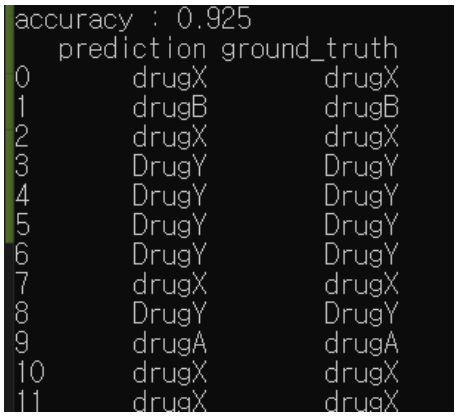
정확도를 높이기 위하여 input 항목을 4개로 늘렸다. (age, bp, na_to_k, cholesterol) 그리고 다시 학습을 시켜보았더니 이번에는 0.95의 정확도를 보였다.



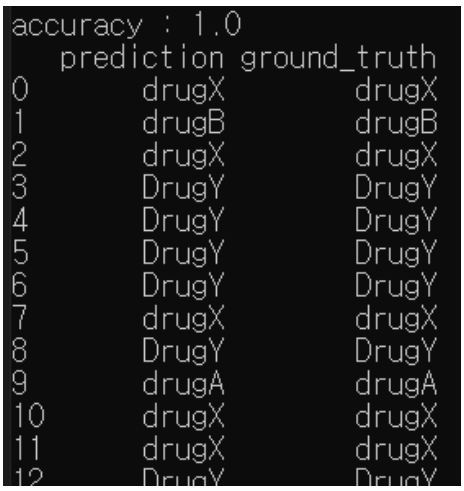
```
{'C': 10, 'gamma': 1, 'kernel': 'rbf'}
              precision    recall  f1-score   support

       DrugY       0.91      1.00      0.95        20
       drugA       1.00      1.00      1.00         5
       drugB       1.00      1.00      1.00         2
       drugC       1.00      1.00      1.00         3
       drugX       1.00      0.80      0.89        10

   micro avg       0.95      0.95      0.95        40
   macro avg       0.98      0.96      0.97        40
weighted avg       0.95      0.95      0.95        40
```

```
28      DrugY      DrugY
29      DrugY      DrugY
30      DrugY      DrugY
31      drugA      drugA
32      DrugY      DrugY
33      drugC      drugC
34      DrugY      DrugY
35      drugC      drugC
36      drugC      drugC
37      drugX      drugX
38      drugX      drugX
39      drugX      drugX
accuracy : 0.95
```

4번 문제 : Decision Tree 모델로 데이터를 학습시켰다. Input 데이터는 동일하게 3가지 항목(BP, Na_to_K, Age) 만 사용하였다. sklearn의 tree 라이브러리를 활용하여 최적의 파라미터를 찾고 학습한 뒤, 테스트 데이터로 테스트를 해보았다. 테스트 결과 정확도는 0.925였다.



정확도를 높이기 위하여 input 항목을 4개로 늘렸다. (age, bp, na_to_k, cholesterol) 그리고 다시 학습을 시켜보았더니 이번에는 1.0의 정확도를 보였다.



4.2 Conclusion

1. 학습 전, 데이터 시각화 분석을 통해 무의미한 데이터인 sex를 제외시킬 수 있었다.

2. KNN 모델은 input 데이터로 3개(age, bp, na_to_k)를 사용했을 때가 테스트 정확도(0.9)가 가장 높았다.

3. SVM 모델과 Decision Tree 모델은 input 데이터로 4개(age, bp, na_to_k, cholesterol)를 사용했을 때가 테스트 정확도(각각 0.95, 1.0)가 가장 높았다.

4. 이번 과제를 통하여 sklearn을 처음 사용해보고, 여러가지 라이브러리를 사용하여 다양한 classification 모델을 만들어보고 검증해볼 수 있었다.