

```
import warnings
```

```
# 불필요한 경고 출력을 방지합니다.  
warnings.filterwarnings('ignore')
```

```
import pandas as pd
```

✓ 실습에 주로 활용할 데이터 셋

[sklearn.datasets](#)

sklearn.dataset 에서 제공하는 다양한 샘플 데이터를 활용합니다.

✓ iris 데이터셋

꽃 종류 분류하기

[iris 데이터셋](#)

```
from sklearn.datasets import load_iris
```

```
# iris 데이터셋을 로드합니다.  
iris = load_iris()
```

- DESCR: 데이터셋의 정보를 보여줍니다.
- data: **feature data**.
- feature_names: **feature data**의 컬럼 이름
- target: **label data** (수치형)
- target_names: **label**의 이름 (문자형)

```
print(iris['DESCR'])
```

```
.. _iris_dataset:
```

```
Iris plants dataset
```

```
-----  
  
**Data Set Characteristics:**
```

```
:Number of Instances: 150 (50 in each of three classes)
:Number of Attributes: 4 numeric, predictive attributes and the class
:Attribute Information:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
  - class:
    - Iris-Setosa
    - Iris-Versicolour
    - Iris-Virginica
```

```
:Summary Statistics:
```

	Min	Max	Mean	SD	Class Correlation
sepal length:	4.3	7.9	5.84	0.83	0.7826
sepal width:	2.0	4.4	3.05	0.43	-0.4194
petal length:	1.0	6.9	3.76	1.76	0.9490 (high!)
petal width:	0.1	2.5	1.20	0.76	0.9565 (high!)

```
:Missing Attribute Values: None
:Class Distribution: 33.3% for each of 3 classes.
:Creator: R.A. Fisher
:Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
:Date: July, 1988
```

The famous Iris database, first used by Sir R.A. Fisher. The dataset is taken from Fisher's paper. Note that it's the same as in R, but not as in the UCI Machine Learning Repository, which has two wrong data points.

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

```
.. topic:: References
```

- Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.
- Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.

```
data = iris['data']
data[:5]
```

```
array([[5.1, 3.5, 1.4, 0.2],
       [4.9, 3. , 1.4, 0.2],
       [4.7, 3.2, 1.3, 0.2],
```

```
[4.6, 3.1, 1.5, 0.2],  
[5. , 3.6, 1.4, 0.2]])
```

```
feature_names = iris['feature_names']  
feature_names
```

```
['sepal length (cm)',  
 'sepal width (cm)',  
 'petal length (cm)',  
 'petal width (cm)']
```

- **sepal:** 꽃 받침
- **petal:** 꽃잎



```
target = iris['target']  
target[45:55]  
  
array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1])
```

```
iris['target_names']  
  
array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

✓ 데이터프레임 만들기

```
df_iris = pd.DataFrame(data, columns=feature_names)
```

```
df_iris.head()
```



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	
0	5.1	3.5	1.4	0.2	
1	4.9	3.0	1.4	0.2	
2	4.7	3.2	1.3	0.2	
3	4.6	3.1	1.5	0.2	
4	5.0	3.6	1.4	0.2	

```
df_iris['target'] = target
```

```
df_iris.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	
0	5.1	3.5	1.4	0.2	0	
1	4.9	3.0	1.4	0.2	0	
2	4.7	3.2	1.3	0.2	0	
3	4.6	3.1	1.5	0.2	0	
4	5.0	3.6	1.4	0.2	0	

df_iris

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	
0	5.1	3.5	1.4	0.2	0	
1	4.9	3.0	1.4	0.2	0	
2	4.7	3.2	1.3	0.2	0	
3	4.6	3.1	1.5	0.2	0	
4	5.0	3.6	1.4	0.2	0	
...	
145	6.7	3.0	5.2	2.3	2	
146	6.3	2.5	5.0	1.9	2	
147	6.5	3.0	5.2	2.0	2	
148	6.2	3.4	5.4	2.3	2	
149	5.9	3.0	5.1	1.8	2	

150 rows × 5 columns

시각화

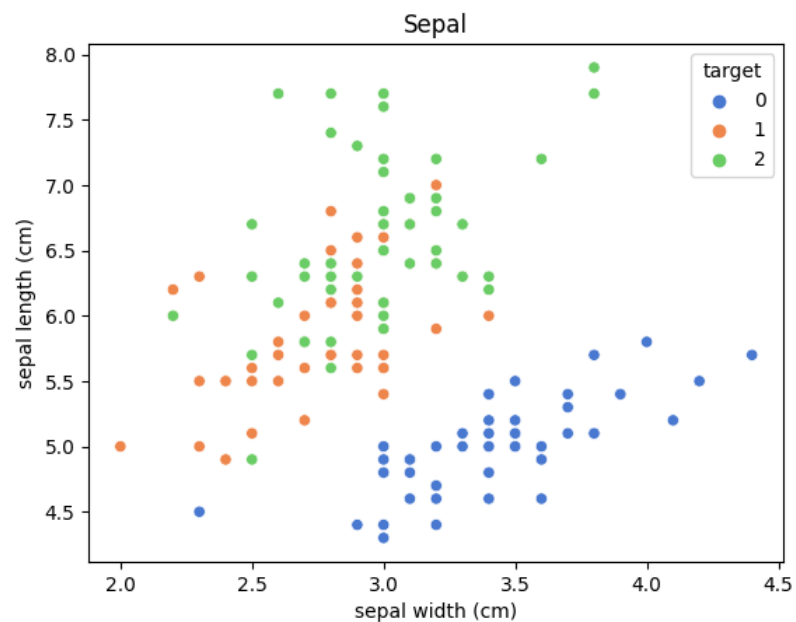
```
import matplotlib.pyplot as plt
import seaborn as sns
```

```
!pip install seaborn
```

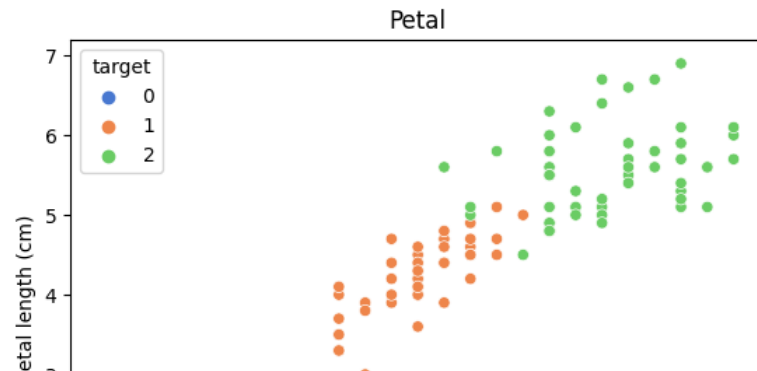
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (0.12.2)
 Requirement already satisfied: numpy!=1.24.0,>=1.17 in /usr/local/lib/python3.10/dist-packages (from seaborn) (1.23.5)
 Requirement already satisfied: pandas>=0.25 in /usr/local/lib/python3.10/dist-packages (from seaborn) (1.5.3)
 Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in /usr/local/lib/python3.10/dist-packages (from seaborn) (3.7.1)
 Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.2.0)
 Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (0.12.1)
 Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (4.44.3)

Requirement already satisfied: kiwisolver<=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.4.5)
Requirement already satisfied: packaging<=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (23.2)
Requirement already satisfied: pillow<=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (9.4.0)
Requirement already satisfied: pyparsing<=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (3.1.1)
Requirement already satisfied: python-dateutil<=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (2.8.2)
Requirement already satisfied: pytz<=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=0.25->seaborn) (2023.3.post1)
Requirement already satisfied: six<=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.1->seaborn) (1.16.0)

```
sns.scatterplot(x='sepal width (cm)', y='sepal length (cm)', hue='target', palette='muted', data=df_iris)
plt.title('Sepal')
plt.show()
```



```
sns.scatterplot(x='petal width (cm)', y='petal length (cm)', hue='target', palette='muted', data=df_iris)
plt.title('Petal')
plt.show()
```



```
from mpl_toolkits.mplot3d import Axes3D
from sklearn.decomposition import PCA

fig = plt.figure(figsize=(8, 6))
ax = Axes3D(fig, elev=-150, azimuth=110)
X_reduced = PCA(n_components=3).fit_transform(df_iris.drop('target', axis=1))
ax.scatter(X_reduced[:, 0], X_reduced[:, 1], X_reduced[:, 2], c=df_iris['target'],
           cmap=plt.cm.Set1, edgecolor='k', s=40)
ax.set_title("Iris 3D")
ax.set_xlabel("x")
ax.w_xaxis.set_ticklabels([])
ax.set_ylabel("y")
ax.w_yaxis.set_ticklabels([])
ax.set_zlabel("z")
ax.w_zaxis.set_ticklabels([])

plt.show()
```

☞ <Figure size 800x600 with 0 Axes>

```
from sklearn.model_selection import train_test_split

x_train, x_valid, y_train, y_valid = train_test_split(df_iris.drop('target', 1), df_iris['target'])

x_train.shape, y_train.shape

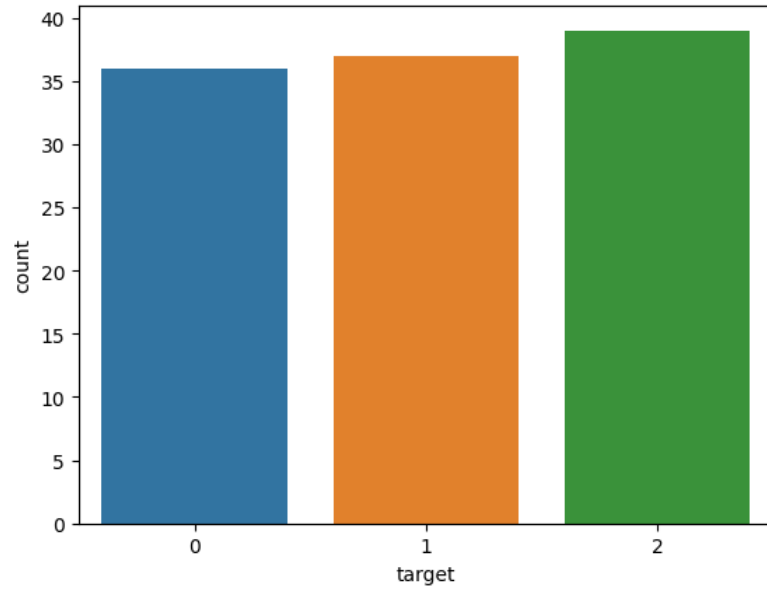
((112, 4), (112,))
```

```
x_valid.shape, y_valid.shape
```

```
((38, 4), (38,))
```

```
sns.countplot(x=y_train)
```

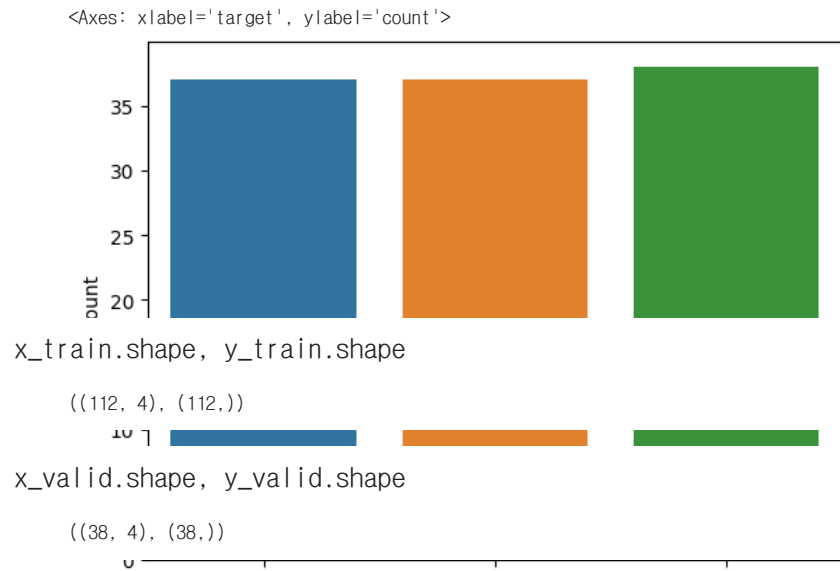
<Axes: xlabel='target', ylabel='count'>



stratify: label의 클래스의 분포를 균등하게 배분

```
x_train, x_valid, y_train, y_valid = train_test_split(df_iris.drop('target', 1), df_iris['target'], stratify=df_iris['target'])
```

```
sns.countplot(x=y_train)
```



Logistic Regression

[도큐먼트](#)

- 로지스틱 회귀(영어: logistic regression)는 영국의 통계학자인 D. R. Cox가 1958년에 제안한 확률 모델
- 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법

LogisticRegression, 서포트 벡터 머신 (SVM) 과 같은 알고리즘은 이진 분류만 가능합니다. (2개의 클래스 판별만 가능합니다.)

하지만, 3개 이상의 클래스에 대한 판별을 진행하는 경우, 다음과 같은 전략으로 판별하게 됩니다.

one-vs-rest (OvR): K 개의 클래스가 존재할 때, 1개의 클래스를 제외한 다른 클래스를 K개 만들어, 각각의 이진 분류에 대한 확률을 구하고, 총합을 통해 최종 클래스를 판별

one-vs-one (OvO): 4개의 계절을 구분하는 클래스가 존재한다고 가정했을 때, 0vs1, 0vs2, 0vs3, ..., 2vs3 까지 $N(N-1)/2$ 개의 분류기를 만들어 가장 많이 양성으로 선택된 클래스를 판별

대부분 **OvsR 전략을 선호합니다.**

```
from sklearn.linear_model import LogisticRegression
```

step 1: 모델 선언

```
model = LogisticRegression()
```


step 2: 모델 학습

```
model.fit(x_train, y_train)
```

```
▼ LogisticRegression  
LogisticRegression()
```

step 3: 예측

```
prediction = model.predict(x_valid)
```

```
prediction[:5]
```

```
array([2, 2, 1, 1, 1])
```

step 4: 평가

```
(prediction == y_valid).mean()
```

```
0.9473684210526315
```

▼ SGDClassifier

stochastic gradient descent (SGD): 확률적 경사 하강법

```
from IPython.display import Image
```

```
# 출처: https://machinelearningnotepad.wordpress.com/
```

```
Image('https://machinelearningnotepad.files.wordpress.com/2018/04/yk1mk.png', width=500)
```

$J(w)$ ↑ Initial Gradient

[sklearn 문서](#)

```
from sklearn.linear_model import SGDClassifier
```

step 1: 모델 선언

```
sgd = SGDClassifier(random_state=0)
```

step 2: 모델 학습

```
sgd.fit(x_train, y_train)
```

```
SGDClassifier  
SGDClassifier(random_state=0)
```

step 3: 예측

```
prediction = sgd.predict(x_valid)
```

```
(prediction == y_valid).mean()
```

```
1.0
```

✓ 하이퍼 파라미터 (hyper-parameter) 튜닝

[도큐먼트](#)

각 알고리즘 별, hyper-parameter의 종류가 다양합니다.

모두 다 외워서 할 수는 없습니다! 문서를 보고 적절한 가설을 세운 다음 적용하면서 검증해야 합니다.

(나중에는 이 또한 자동으로 할 수 있습니다)

- random_state: 하이퍼 파라미터 튜닝시, 고정할 것
- n_jobs=-1: CPU를 모두 사용 (학습속도가 빠름)

```
sgd = SGDClassifier(penalty='l1', random_state=0, n_jobs=-1)
```

```
sgd.fit(x_train, y_train)
```

```
SGDClassifier
SGDClassifier(n_jobs=-1, penalty='l1', random_state=0)
```

```
prediction = sgd.predict(x_valid)
```

```
(prediction == y_valid).mean()
```

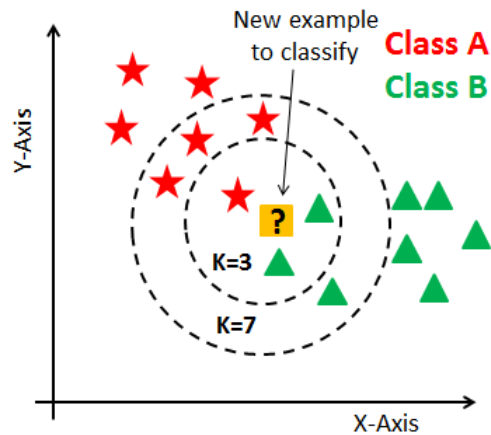
```
0.8157894736842105
```

▼ KNeighborsClassifier

최근접 이웃 알고리즘

출처: 데이터 캠프

```
Image('https://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1531424125/KNN_final_a1mr9.png')
```



```
from sklearn.neighbors import KNeighborsClassifier
```

```
knc = KNeighborsClassifier()
```

```
knc.fit(x_train, y_train)
```

```
▼ KNeighborsClassifier  
KNeighborsClassifier()
```

```
knc_pred = knc.predict(x_valid)
```

```
(knc_pred == y_valid).mean()
```

```
0.9736842105263158
```

```
knc = KNeighborsClassifier(n_neighbors=9)
```

```
knc.fit(x_train, y_train)
```

```
knc_pred = knc.predict(x_valid)
```

```
(knc_pred == y_valid).mean()
```

```
0.9473684210526315
```

✓ 서포트 벡터 머신 (SVC)

- 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만들.
- 경계로 표현되는 데이터들 중 가장 큰 폭을 가진 경계를 찾는 알고리즘.

```
Image('https://csstudy.files.wordpress.com/2011/03/screen-shot-2011-02-28-at-5-53-26-pm.png')
```

LogisticRegression과 같이 이진 분류만 가능합니다. (2개의 클래스 판별만 가능합니다.)

- OvsR 전략 사용

[도큐먼트](#)

```
from sklearn.svm import SVC

svc = SVC(random_state=0,)
svc.fit(x_train, y_train)
svc_pred = svc.predict(x_valid)
```

SVC

```
▼ SVC
SVC(random_state=0)
```

```
(svc_pred == y_valid).mean()
```

```
0.9473684210526315
```

```
svc_pred[:5]
```

```
array([2, 2, 1, 1, 1])
```

각 클래스 별 확률값을 return 해주는 `decision_function()`

```
svc.decision_function(x_valid)[:5]
```

```
array([[ -0.23806612,  1.00400122,  2.23773416],
       [-0.22231313,  0.85074607,  2.24591928],
       [-0.23155949,  2.21892513,  1.0885398 ],
       [-0.21943923,  2.22185472,  0.98012224],
       [-0.20671639,  2.237056  ,  0.84885505]])
```

▼ 의사 결정 나무 (Decision Tree)

스무고개처럼, 나무 가지치기를 통해 소그룹으로 나누어 판별하는 것

Image('https://upload.wikimedia.org/wikipedia/commons/thumb/a/a7/Classification_tree_on_iris_dataset.png/800px-Classification_tree_on_iris_dataset



도큐먼트

```
from sklearn.tree import DecisionTreeClassifier
```

```
dtc = DecisionTreeClassifier(random_state=0)
```

```
dtc.fit(x_train, y_train)
```

```
▼ DecisionTreeClassifier  
DecisionTreeClassifier(random_state=0)
```

```
dtc_pred = dtc.predict(x_valid)
```

```
(dtc_pred == y_valid).mean()
```

```
0.9473684210526315
```

✓ 오차 (Error)

✓ 정확도의 함정

```
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
import numpy as np
```

유방암 환자 데이터셋을 로드합니다.

target: 0: 악성종양, 1:양성종양

```
cancer = load_breast_cancer()
```

```
print(cancer['DESCR'])
```

This database is also available through the UW CS ftp server:

```
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/
```

.. topic:: References

- W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
- O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
- W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171

```
data = cancer['data']
target = cancer['target']
feature_names=cancer['feature_names']
```

데이터 프레임을 생성합니다.

```
df = pd.DataFrame(data=data, columns=feature_names)
df['target'] = cancer['target']
```

```
df.head()
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	wor perimet
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152

5 rows × 31 columns

```
pos = df.loc[df['target']==1]
neg = df.loc[df['target']==0]
```


pos

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	wc perime
19	13.540	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.047810	0.1885	0.05766	...	19.26	9
20	13.080	15.71	85.63	520.0	0.10750	0.12700	0.04568	0.031100	0.1967	0.06811	...	20.49	9
21	9.504	12.44	60.34	273.9	0.10240	0.06492	0.02956	0.020760	0.1815	0.06905	...	15.66	6
37	13.030	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.029230	0.1467	0.05863	...	22.81	8
46	8.196	16.84	51.71	201.9	0.08600	0.05943	0.01588	0.005917	0.1769	0.06503	...	21.96	5
...	
558	14.590	22.68	96.39	657.1	0.08473	0.13300	0.10290	0.037360	0.1454	0.06147	...	27.27	10
559	11.510	23.93	74.52	403.5	0.09261	0.10210	0.11120	0.041050	0.1388	0.06570	...	37.16	8
560	14.050	27.15	91.38	600.4	0.09929	0.11260	0.04462	0.043040	0.1537	0.06171	...	33.17	10
561	11.200	29.37	70.67	386.0	0.07449	0.03558	0.00000	0.000000	0.1060	0.05502	...	38.30	7
568	7.760	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.000000	0.1587	0.05884	...	30.37	5

357 rows × 31 columns

양성 환자 357개 + 악성 환자 5개

```
sample = pd.concat([pos, neg[:5]], sort=True)

x_train, x_test, y_train, y_test = train_test_split(sample.drop('target', 1), sample['target'], random_state=42)

모델을 정의하고, 학습합니다.

model = LogisticRegression()
model.fit(x_train, y_train)
pred = model.predict(x_test)

(pred == y_test).mean()

0.978021978021978

my_prediction = np.ones(shape=y_test.shape)
```

```
(my_prediction == y_test).mean()
```

```
0.989010989010989
```

정확도만 놓고 본다면, 제가 만든 무조건 음성 환자로 예측하는 분류기가 성능이 좋습니다

하지만, 의사가 과연 **무조건 음성 환자로 예측해서 예측율 98.9%로 말하는 의사**는 자질이 좋은 의사일까요?

정확도(accuracy)만 보고 분류기의 성능을 판별하는 것은 **위와 같은 오류**에 빠질 수 있습니다.

이를 보완하고 생겨난 지표들이 있습니다. 차차 알아보겠습니다.

▼ 오차 행렬 (confusion matrix)

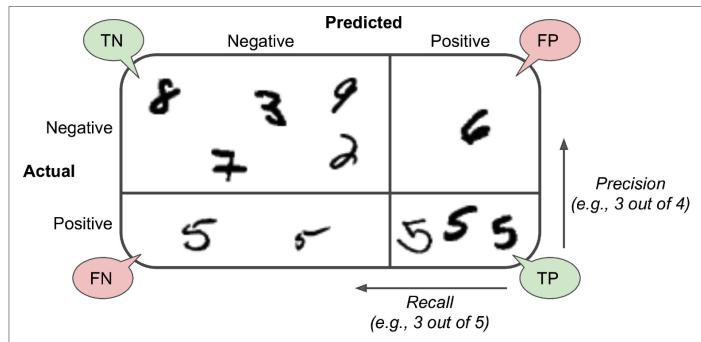
```
from sklearn.metrics import confusion_matrix
```

```
confusion_matrix(y_test, pred)
```

```
array([[ 1,  0],  
       [ 2, 88]])
```

```
sns.heatmap(confusion_matrix(y_test, pred), annot=True, cmap='Reds', )  
plt.xlabel('Predict')  
plt.ylabel('Actual')  
plt.show()
```

출처: <https://dojinkimm.github.io>
Image('https://dojinkimm.github.io/assets/imgs/ml/handson_3_1.png', width=500)



```
from sklearn.metrics import precision_score, recall_score
```

✓ 정밀도 (precision)

양성 예측 정확도

$TP / (TP + FP)$

```
precision_score(y_test, pred)
```

1.0

무조건 **양성**으로 판단하면 좋은 정밀도를 얻기 때문에 유용하지 않습니다.

✓ 재현율 (recall)

$TP / (TP + FN)$

정확하게 감지한 **양성 샘플의 비율**입니다.

민감도 (sensitivity) 혹은 True Positive Rate (TPR)이라고도 불리웁니다.

```
recall_score(y_test, pred)

0.9777777777777777

88/90

0.9777777777777777
```

▼ f1 score

정밀도와 재현율의 **조화 평균**을 나타내는 지표입니다.

$$2 * \frac{\text{정밀도} * \text{재현율}}{\text{정밀도} + \text{재현율}} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

```
from sklearn.metrics import f1_score

f1_score(y_test, pred)

0.9887640449438202
```