표준화, 정규화

표준화(Standardization): 데이터의 피처 각각이 평균이 0, 분산이 1인 가우시안 정규분포를 가진 값으로 변환하는 작업을 표준화라고 함.

$$x_{i\_new} = \frac{x_i - \mathrm{mean}(x)}{stdev(x)}$$

실제 구현시에는 사이킷런의 StandardScaler를 사용해 표준화를 진행하는것이 일반적임

정규화(Normalization): 서로 다른 피처들의 크기를 통일하기 위해 크기를 변화해주는 것.

$$x_{i\_new} = \frac{x_i - \min(x)}{max(x) - \min(x)}$$

실제 구현시에는 사이킷런에서 제공하는 MinMaxScaler는 음수 값이 없으면 0 ~ 1의 값으로, 음수 값이 있으면 -1 ~ 1의 값으로 변환해준다.

표준화, 정규화

```
import pandas as pd

df = pd.read_csv('./kaggle_diabetes.csv', sep=',')

df.describe()
```

```
# 표준화 (Standardization)

from sklearn.preprocessing import StandardScaler

std_cols = ['Pregnancies', 'Glucose', 'BloodPressure','SkinThickness',
            'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age' ]

scaler = StandardScaler()

df_std = scaler.fit_transform(df[std_cols])

print(type(df_std))

df_std = pd.DataFrame(df_std, columns=std_cols)

df_std['Outcome'] = df['Outcome'].values

df_std.describe()
```

표준화, 정규화

```
# 정규화 (Normalization)

from sklearn.preprocessing import MinMaxScaler

norm_cols = ['Pregnancies', 'Glucose', 'BloodPressure','SkinThickness',
             'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age' ]

scaler = MinMaxScaler()

df_norm = scaler.fit_transform(df[norm_cols])

print(type(df_norm))

df_norm = pd.DataFrame(df_norm, columns=std_cols)

df_norm['Outcome'] = df['Outcome'].values

df_norm.describe()
```