로지스틱 회귀를 이용한 질병 발생 예측(뇌졸증)

로지스틱 회귀 : 선형 회귀 방식을 분류에 적용한 알고리즘으로

주로 이진 분류에 사용된다

필요한 모듈을 불러온후 분석에 필요한 데이터를 불러온다

In [88]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.linear_model import LogisticRegression
import warnings
warnings.filterwarnings(action='ignore')
df = pd.read_csv("val3_1.csv")
df.head(5)
```

Out[88]:

	고혈압	이상지질혈증	뇌졸증	심근경색증	협심증	관절염	골관절염	류마티스관절염	골다공증	폐결핵	...	b11	b12	b13	b14	b15	b16	b17	b18	
0	1	1	1	0	0	0	0	0	0	0	...	0	0	6	1	0	162.4	56.0	82.2	21
1	0	0	0	0	0	0	0	0	0	0	...	0	0	4	3	0	167.7	76.4	98.3	27
2	1	0	0	0	0	1	1	0	0	0	...	0	0	2	6	0	157.7	53.2	80.7	21
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	1	151.0	53.5	79.7	23
4	0	0	0	0	0	0	0	0	0	0	...	0	0	4	1	0	146.1	57.6	93.0	26

5 rows × 46 columns

고혈압, 뇌졸증등 질병 진단 여부에 대한 변수들과

환자 특성에 대한 변수(b1~b20)이 있다.

독립변수와 종속변수 즉, X와 Y를 분류한다

이번 분석에서는 '뇌졸증'에 대해서 분석하겠다

In [89]:

```
df_Y = df['뇌졸증']
df_X = df.drop('뇌졸증',axis=1)
```

로지스틱 회귀분석에서는 데이터 전처리가 필요하다.

StandardScaler( )로 평균이 0, 분산 1로 데이터 분포를 변환한다

학습 데이터를 80%, 테스트 데이터를 20%로 하였다

In [90]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

scaler = StandardScaler()
data_scaled = scaler.fit_transform(df_X)
X_train , X_test, Y_train , Y_test = train_test_split(data_scaled, df_Y, test_size=0.2, random_s
tate=0)
print('학습 데이터 갯수 :',len(X_train))
print('테스트 데이터 갯수 :',len(X_test))
```

```
학습 데이터 갯수 : 6393
테스트 데이터 갯수 : 1599
```

로지스틱 회귀를 이용하여 학습 및 예측 수행

그 후 예측 모델의 정확도 측정

In [97]:

```
from sklearn.metrics import accuracy_score, roc_auc_score
import statsmodels.api as sm

lr_clf = LogisticRegression()
lr_clf.fit(X_train, Y_train)
lr_preds = lr_clf.predict(X_test)
print('정확도: {:0.3f}'.format(accuracy_score(Y_test, lr_preds)))
```

```
정확도: 0.981
```

정확도가 98.1%로 높은 수치가 나왔다

로지스틱 회귀분석의 각각의 독립 변수들은 종속변수에 끼치는 영향인 상관계수를 가지고 있다.

위 분석에서 상관계수가 높은 순으로 7개만 알아보았다

In [98]:

```
df2 = pd.DataFrame(lr_clf.coef_)
df2.columns=[df_X]
df2.index=['상관계수']
df3 = df2.T.sort_values(by=['상관계수'], axis=0,ascending=False)
df3.head(7)
```

Out[98]:

	상관계수
b2	0.742444
고혈압	0.569321
b18	0.490675
골관절염	0.385269
b5	0.321552
골다공증	0.159554
b1	0.139451

가상의 데이터를 생성 후 위 예측모델에 적용하였다.

In [99]:

```
df_New = pd.read_csv("New.csv")
df_New = df_New.drop('뇌졸증',axis=1)
df_New
```

Out[99]:

	고혈압	이상지질혈증	심근경색증	협심증	관절염	골관절염	류마티스관절염	골다공증	폐결핵	천식	...	b11	b12	b13	b14	b15	b16	b17	b18	
0	1	0	0	0	0	1	0	0	0	0	...	0	0	5	2	0	172.5	44.7	82.2	21
1	0	0	0	0	0	0	0	0	0	0	...	0	0	5	2	0	180.9	92.5	98.3	27
2	0	0	0	0	0	0	0	0	0	0	...	0	0	3	4	0	157.7	49.9	80.7	21

3 rows × 45 columns

In [100]:

```
scaler.transform(df_New)
New_predict = lr_clf.predict_proba(df_New)
print(New_predict)
```

```
[[1.26936047e-02 9.87306395e-01]
 [8.25724651e-06 9.99991743e-01]
 [1.14567091e-03 9.98854329e-01]]
```

앞의 숫자가 결과 값이 0이 나올 확률,

뒷의 숫자가 결과 값이 1이 나올 확률이다

위의 데이터에서는 모든 환자가 99%이상의 확률로 결과 값이 1이 나올 수 있다.

즉, 위의 환자들은 뇌졸증의 걸릴 확률이 아주 크다고 할 수 있다.