Multivariate Data Analysis Assignment #2

Rogistic Linear Regression

산업경영공학부 2020170831 민찬홍

[Q1] 데이터셋 선정 및 선정 이유

.-데이터셋: 당뇨병 데이터셋

-다운로드링크: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

```
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None
```

-선정 이유: 로지스틱 회귀 분석을 수행하기 위한 데이터셋의 기준 두 가지를 다음과 같이 잡았다. 첫 번째, 종속 변수가 명목형(categorical type)인 데이터 셋이다. 둘째, 종속 변수를 설명 변수의 선형 결합으로 잘 표현할 수 있는 데이터 셋이다.

위 당뇨병 데이터셋은 종속 변수인 당뇨병의 발병 유무(명목형 변수)가 연속형 변수가 아닌 이진형 변수(1/0)였고, 8개의 설명 변수들이 종속 변수와 선형 관계를 가질 것이라 생각하였기에 로지스틱 회귀 분석에 적합할 것이라 생각하여 위 데이터 셋을 선정하게 되었다.

[Q2] 데이터셋 설명

-종속 변수: Outcome- 당뇨병 유무(1=Yes, 0=No)

-설명 변수(8개)

Pregnancies- 임신 횟수

Glucose- 혈당

BloodPressure- 혈압

SkinThickness- 피부 두께

Insulin- 인슐린 레벨

BMI- 체질량 지수

Age- 나이

DiabetesPedrigreeFunction- 당뇨병 혈통 여부(가족력)

[Q2-1] 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

위 데이터에서 제공된 설명변수 중 종속변수인 당뇨병 유무와 높은 상관관계가 있을 것이라 생각되는 변수는 가족력, 체질량 지수, 인슐린 레벨, 혈당, 나이라고 생각한다.

모든 병의 경우 유전적 요인과 나이를 배제할 수 없다. 당장 의시가 병을 진단 시에도 가족력을 물어본다. 여기에 더해 인간은 나이를 먹을수록 노화가 진행되고, 자연스레 여러 질병의 발병률이 올라간다는 사실은 부정할 수 없다. 따라서 당뇨병 역시 가족력 및 나이와 강한 상관관계가 있을 것이라 생각한다.

체질량 지수(BMI)란 인간의 비만도를 나타내는 지수로 키와 몸무게만으로 측정한 지수이다. 간혹 체질량 지수(BMI)가 높더라도 근육이나 골밀도가 발달하여 체질량 지수(BMI)가 높은 사람도 있다. 그러나 인간의 신체는 생존을 위해 근육보다 지방을 저장하는 것이 더 쉽도록 진화하였기 때문에 대부분의 체 체질량 지수(BMI)가 높은 사람들은 체지방률이 높을 것이라 판단하였다. 체지방률이 높은 것은 모든 성인병의 근원이라고 이미 널리 알려진 사실이다. 뿐만 아니라 체질량 지수(BMI)가 높아지는 원인으로 칼로리가 높거나 당의 함량이 높은 음식의 섭취는 당뇨병에 직접적인 영향을 미치는 안 좋은 음식이 다라고 봐도 무방하다. 따라서 체질량 지수(BMI) 역시 당뇨병 유무와 강한 상관관계가 있을 것이라고 생각했다.

당뇨병이란 우리 몸의 혈당을 조절하는 물질인 인슐린이 모자라거나 제대로 일을 하지 않아 혈당이 상승하는 병을 말한다. 즉, 인슐린 레벨은 당뇨병과 직접적인 상관관계가 있을 수밖에 없다고 생각하였다.

인슐린과 마찬가지로 혈당 역시 당뇨병과 직접적인 상관관계가 있을 수밖에 없다고 생각한다. 왜냐하면 우리 몸은 항상성을 유지하려고 하는 성질이 있는데 이 항상성을 유지하려고 하는 힘이 깨져 혈당이 높아진 상태가 유지된 병이 당뇨병이기 때문이다.

[Q2-2] 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

임신 횟수, 피부 두께는 종속변수인 당뇨병 여부를 예상하는데 필요하지 않을 것이라 생각했다.

우선 피부 두께와 같은 외과적 요소들은 내과적 요소들로 인해 발병하는 병인 당뇨병에 어떠한 직접적인 영향을 미치지 못할 것이라고 생각했기 때문이다.

임신 횟수의 경우 체내 호르몬에 영향을 미쳐 당뇨병에 영향을 미친다고 생각할 수 있다. 실제로 임신성 당뇨병이란 것도 있으니 말이다. 하지만, 대부분의 경우 임신 중 강해진 인슐린 저항성에 맞서 췌장에서 더 많은 인슐린이 분비되고, 임신 후 다시 인슐린 저항성이 안정화되기에 임신을 많이 했다는 사실만으로 당뇨병에 직접적으로 영향을 주지 못한다고 생각하여 당뇨병 발병을 예측하는 데는 필요하지 않을 것이라 생각하였다.

[Q3] 개별 입력변수들에 대한 통계량 계산 및 분석

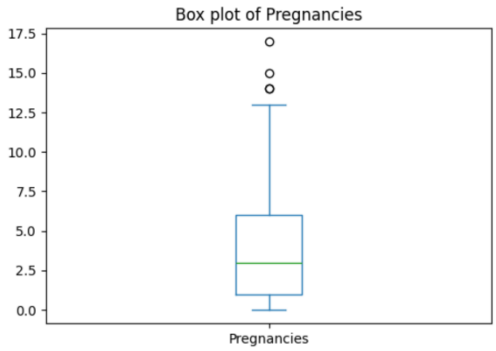
1) 개별 입력 변수의 통계량

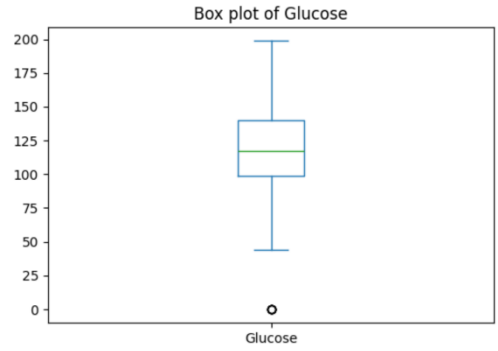
개별 변수들에 대한 mean, std dev, skewness, kurtosis는 다음과 같다.

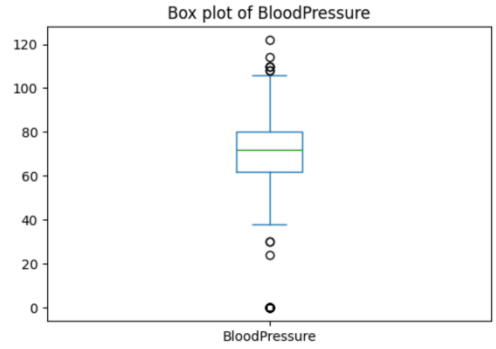
	Mean	Std Dev	Skewness	Kurtosis
Pregnancies	3.845052	3.369578	0.899912	0.150383
Glucose	120.894531	31.972618	0.173414	0.628813
BloodPressure	69.105469	19.355807	-1.840005	5.138691
SkinThickness	20.536458	15.952218	0.109159	-0.524494
Insulin	79.799479	115.244002	2.267810	7.159575
BMI	31.992578	7.884160	-0.428143	3.261257
DiabetesPedigreeFunction	0.471876	0.331329	1.916159	5.550792
Age	33.240885	11.760232	1.127389	0.631177
Outcome	0.348958	0.476951	0.633776	-1.598328

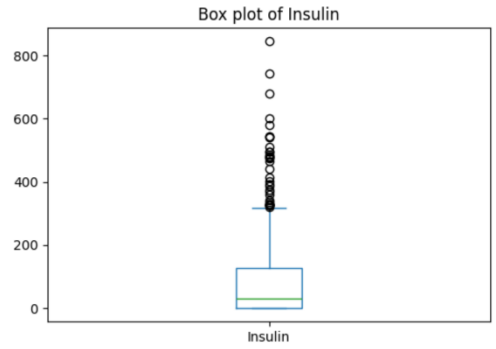
2) Box plot

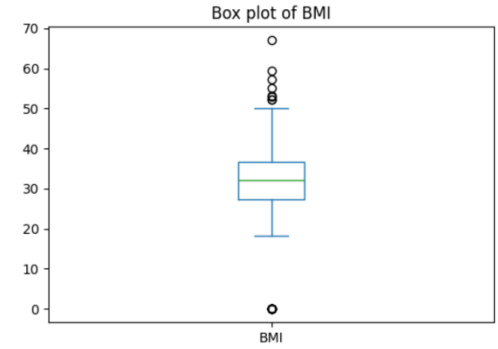
각 numerical 설명 변수들에 대한 box plot은 다음과 같다.

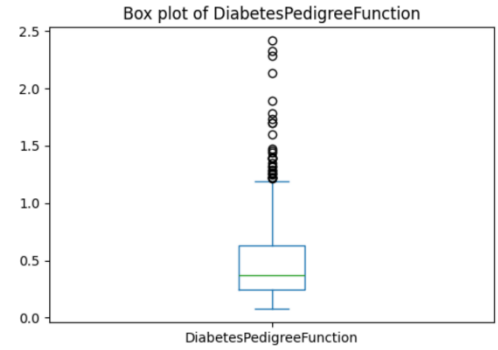


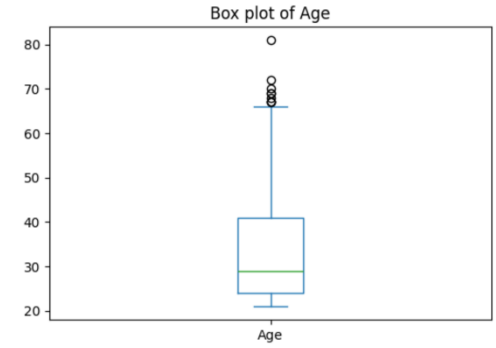












3) 정규분포를 가정할 수 있는 변수는 몇 개인가?

전체 변수들 중에서 정규분포를 따른다고 가정할 수 있는 변수는 Glucose(혈당), SkinThickness(피부 두께)라고 생각했다.

정규분포를 따르는 변수를 판단하기 위한 기준으로 skewness(왜도)와 kurtosis(첨도) 값을 사용하였다.

Skewness(왜도)란 데이터 분포의 비대칭성 정도를 나타내는 지표이다. 0에 가까울수록 좌우가 대칭이라 정규분포와 유사하고, 0보다 크거나 작을 경우 데이터의 분포가 각각 왼쪽, 오른쪽으로 치우친 형태를 나타낸다고 볼 수 있다.

Kurtosis(첨도)란 데이터 분포의 뾰족한 정도를 나타내는 지표이다. 3을 기준으로 3보다 작으면 완만하고, 3보다 크면 가파른 형태를 띈다고 볼 수 있다. 3에 가까울수록 데이터 분포는 정규 분포를 띈다고 할 수 있다.

Skewness(왜도), Kurtosis(첨도)를 구할 시 scipy 패키지를 사용하였다. 이 때, Kurtosis(첨도)의 경우 Fisher의 정의를 사용하여 0에 가까울수록 정규분포를 띈다고 할 수 있다.

먼저 Skewness(왜도)값이 0에 가까운 변수는 Glucose, SkinThickness, BMI가 있었다. 그 중 Kurtosis(첨도) 역시 0에 가까운 변수는 Glucose, SkinThickness이다.

따라서 Glucose, , SkinThickness 2개의 변수가 정규분포를 띈다고 가정할 수 있다.

[Q4] 변수에 대한 이상치 확인 및 제거

1) 이상치 정의

Box plot에서 box의 최하단 부분과 최상단 부분은 각각 Q1(25th percentile), Q3(75th percentile) 값을 의미한다. 그리고 Q1과 Q3 사이의 간격을 IQR((Interquartile Range)로 정의할 수 있다. 이를 바탕으로 다음 범위에 속하지 않는 데이터를 이상치로 간주하여 제거하였다.

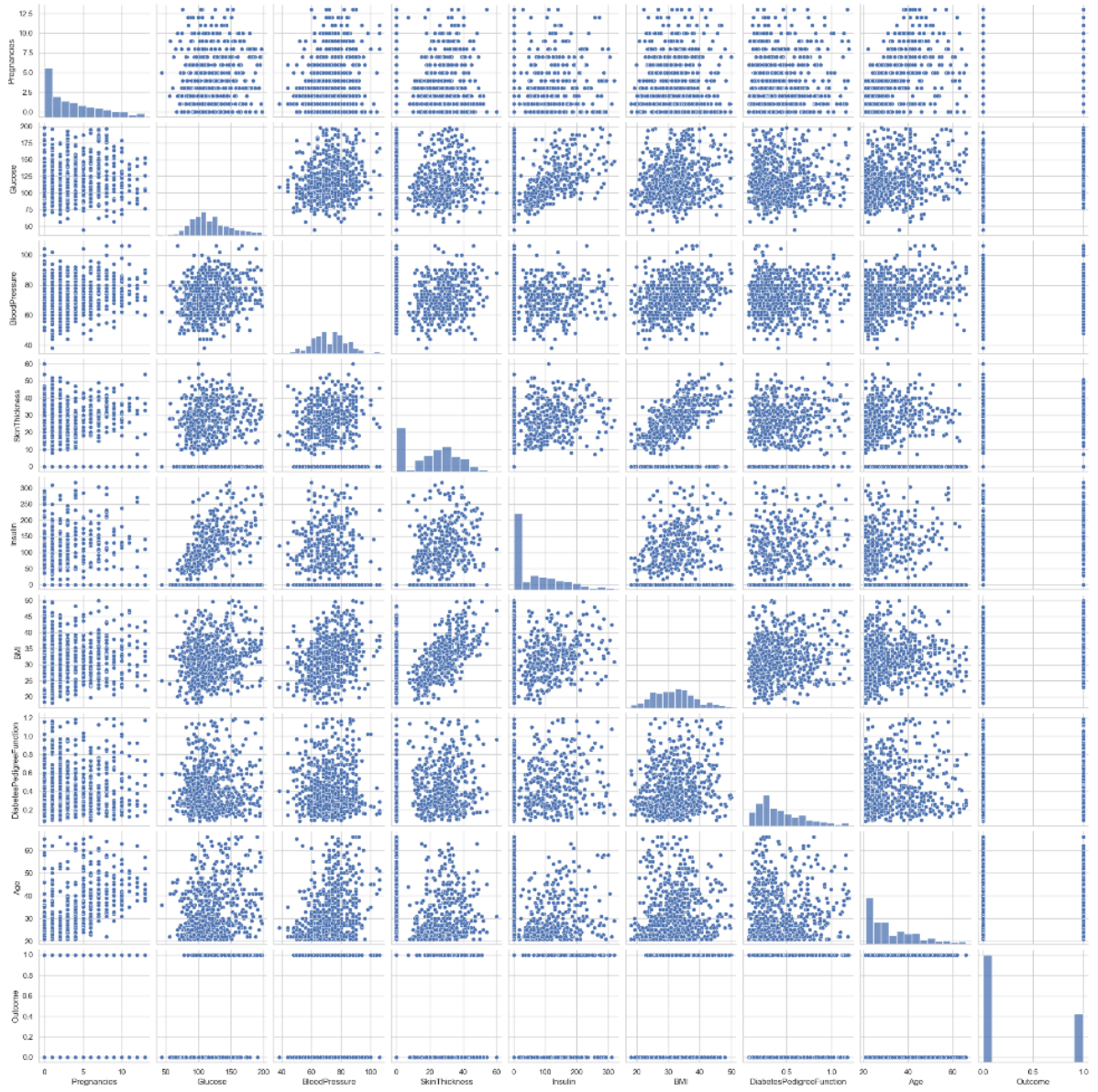
$$Q1 - 1.5 * IQR \leq value \leq Q3 + 1.5 * IQR$$

2) 이상치 데이터 제거

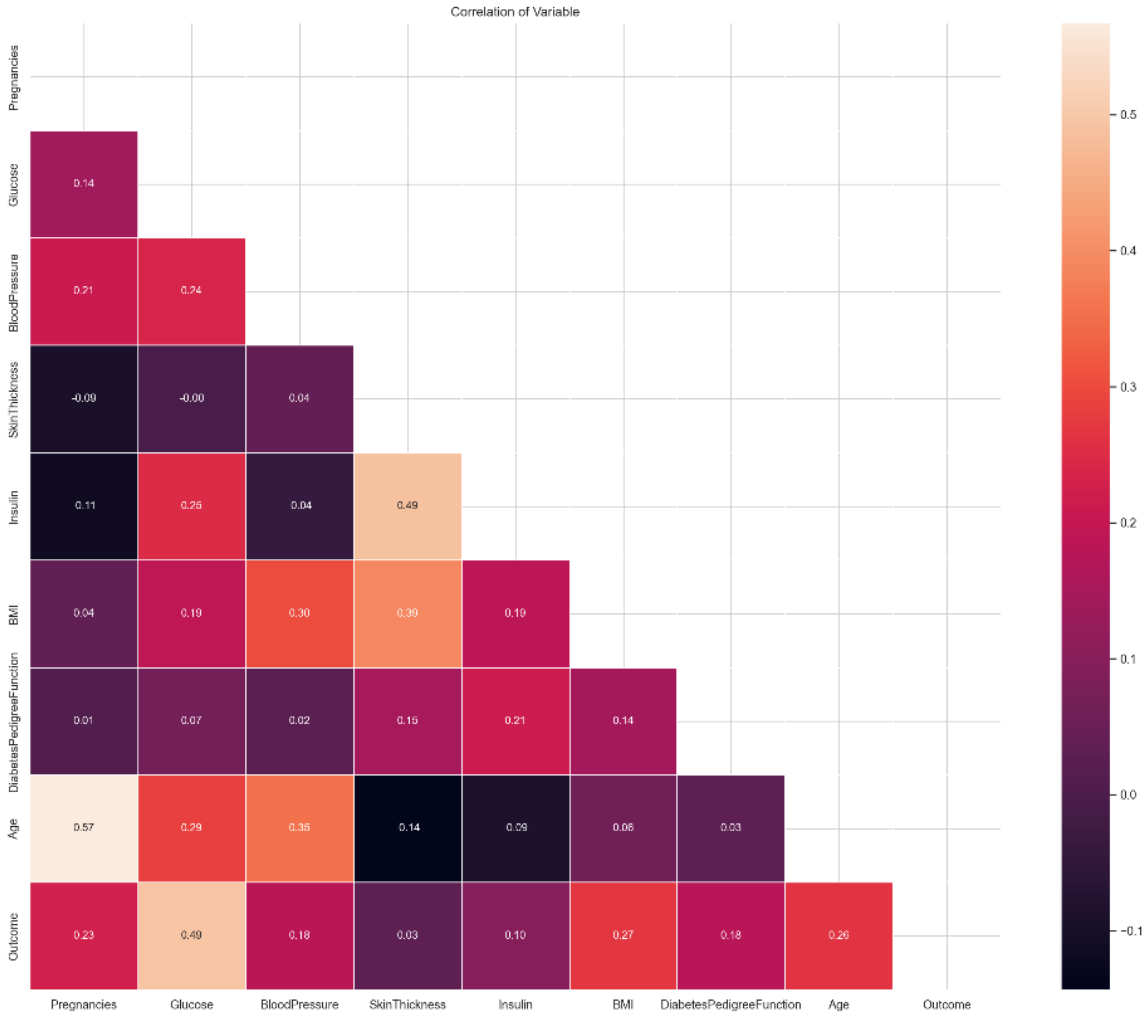
Box plot의 outlier 기준으로 1)의 이상치 정의를 사용하여 이상치를 제거한 결과 기존 768개의 데이터에서 636개의 데이터로 줄어든 것을 확인할 수 있었다.

[Q5] 상관성 분석

-Scatter plot



-Heat map



상관계수는 -1부터 1까지의 값을 가지며, 0에 가까울수록 두 변수의 상관관계가 약하고 상관계수의 절댓값이 1에 가까울수록 상관관계가 크다고 할 수 있다.

통상적으로 상관계수가 0.6이상이면 두 변수는 강한 상관관계가 있다고 한다. 그러나 위 데이터셋에서는 상관계수가 0.6이상인 변수가 없었다.

대신 'Age'와 'Preganancies' 변수 간 상관관계가 0.57로 가장 크게 나왔다. 나이가 많을수록 임신 경험 횟수가 많은 것은 상식적으로도 충분히 납득할 만한 결과이다.

이 'Age'와 'Preganancies' 조합을 사용하여 [Q7]을 진행하기 위해 변수의 개수를 감소시키도록 하겠다.

이 때, 제거할 변수를 선택할 기준으로 각각의 변수들을 하나씩 제거하고 Logistic Linear Regression model에 fit하였을 때 성능을 사용하도록 하겠다. 즉, 둘의 성능을 비교하여 최종적으로 제거할 변수를 선정하도록 하겠다.

-Age 제거한 경우

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0158	0.132	-7.710	0.000	-1.274	-0.758
Pregnancies	0.4682	0.127	3.676	0.000	0.219	0.718
Glucose	1.2929	0.162	7.995	0.000	0.976	1.610
BloodPressure	-0.1172	0.135	-0.867	0.386	-0.382	0.148
SkinThickness	-0.0717	0.148	-0.485	0.627	-0.361	0.218
Insulin	-0.2594	0.146	-1.774	0.076	-0.546	0.027
BMI	0.5259	0.149	3.533	0.000	0.234	0.818
DiabetesPedigreeFunction	0.4075	0.123	3.307	0.001	0.166	0.649

-Pregnancies 제거한 경우

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0291	0.133	-7.744	0.000	-1.290	-0.769
Glucose	1.2487	0.162	7.726	0.000	0.932	1.565
BloodPressure	-0.1877	0.141	-1.327	0.185	-0.465	0.090
SkinThickness	-0.0479	0.148	-0.324	0.746	-0.337	0.242
Insulin	-0.2559	0.146	-1.748	0.080	-0.543	0.031
BMI	0.5358	0.147	3.639	0.000	0.247	0.824
DiabetesPedigreeFunction	0.4020	0.122	3.290	0.001	0.163	0.641
Age	0.4829	0.136	3.547	0.000	0.216	0.750

Age, Pregnancies 변수를 제거하고 로지스틱 회귀 분석을 시행한 결과 두 결과 모두 동일하게 4개의 변수의 p-value값이 0.05보다 작아 유의함을 확인할 수 있었다.

따라서 Pregnancies(임신 횟수)가 [Q 2-2] 답변에 근거하여 당뇨병 발병 여부와 상관도가 떨어진다고 판단하여 Pregnancies(임신 횟수)를 제거하고 [Q7]을 수행하도록 하겠다.

[Q6] Logistic Regression 모델 학습

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습해 보시오. 이 때 70:30으로 구분하는 random seed를 저장하시오.

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.0359      0.134     -7.740      0.000      -1.298      -0.774
Pregnancies    0.3158      0.150      2.108      0.035       0.022       0.609
Glucose        1.2576      0.162      7.749      0.000       0.940       1.576
BloodPressure -0.1858      0.142     -1.312      0.190      -0.463       0.092
SkinThickness -0.0469      0.148     -0.317      0.751      -0.337       0.243
Insulin       -0.2475      0.147     -1.687      0.092      -0.535       0.040
BMI            0.5442      0.149      3.641      0.000       0.251       0.837
DiabetesPedigreeFunction  0.4128      0.124      3.340      0.001       0.171       0.655
Age            0.3000      0.161      1.864      0.062      -0.016       0.615
==============================================================================
```

1. 유의수준 0.05에서 유효한 변수의 수는 몇 개인지 확인하고 각 변수들이 본인의 상식 선에서 실제로 유효하다고 할 수 있는지 판단해 보시오.

유의수준 0.05에서 유효한 변수는 p-value의 값이 0.05보다 작은 변수이다. 이를 만족하는 변수는 Pregnancies, Glucose, BMI, DiabetesPedigreeFunction으로 4개이다.

이 때, 로지스틱 회귀 분석을 통해 구한 coef 값이 양수면 종속 변수와의 관계가 양의 선형관계를 음수면 음의 선형관계를 나타낸다고 할 수 있다. 이를 바탕으로 종속 변수인 당뇨병 발생 여부와의 관계를 서술하면,

➔ '임신 횟수가 많을수록, 혈당 수치가 높을수록, 체질량지수가 높을수록, 당뇨병 가족력이 있다면' 당뇨병의 발병 확률이 높다고 해석할 수 있다.

앞서 [Q 2-1]에서 언급한 바와 같이 혈당 수치가 높을수록, 체질량지수가 높을수록, 당뇨병 가족력이 높을수록 당뇨병의 발병 확률이 증가하는 것은 상식 선에서 유효하다고 할 수 있다.

[Q 2-2]에서 언급하였듯이 임신 횟수의 경우 임신 중 호르몬의 영향으로 인슐린 저항성이 커져 혈당을 낮추는 인슐린이 제 역할을 하지 못하지만, 그에 맞서 췌장에서 더 많은 인슐린이 분비되어 당뇨병 발병 여부와 큰 상관관계가 있지 않을 것이라 생각했다. 하지만, 임신 횟수가 증가함에 따라 인슐린 저항성이 커지는 상황이 반복되어 당뇨병 발병 확률이 증가할 수 있다는 사실 또한 충분히 납득 가능하기에 이 또한 실제로 유효한 결과라고 판단된다.

2. [Q2-2]에서 정성적으로 선택했던 변수들의 P-value를 확인하고 해당 변수가 모델링 측면에 서 실제로 유효하지 않는 것인지 확인해 보시오.

[Q2-2]에서 정성적으로 선택했던 변수들은 Pregnancies(임신 횟수)와 Skinthickness(피부 두께)이다. 임신 횟수의 경우 p-value 값이 0.035, 피부 두께의 경우 p-value 값이 0.751로 유의 수준 0.05를 기준으로 각각 유의함, 유의하지 않음의 결과가 도출되었다.

이를 통해 외과적 요소인 피부 두께의 경우 당뇨병 발병 여부와의 모델링 측면에서 유효하지 않은 변수임을 다시 확인할 수 있었다.

하지만, 임신 횟수의 경우 p-value 값이 0.05보다 작았기에 모델링 측면에서 유효한 변수임을 확인할 수 있었다. 이는 임신 횟수가 증가함에 따라 인슐린 저항성은 증가하지만, 이에 대응하여 췌장에서 나오는 인슐린의 양에 한계가 있는 등 다양한 원인 때문에 임신 횟수가 증가함에 따라 당뇨병 발병 확률이 증가될 수도 있다고 판단하였다.

3. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출하여 비교해보시오

학습된 모델의 클래스 레이블 확인을 통해 양성 클래스가 1(=당뇨병 발병)임을 확인할 수 있었다. 이를 바탕으로 학습 데이터, 테스트 데이터에 대한 정오행렬과 평가지표를 도출하면 다음과 같다.

-학습 데이터에 대한 confusion matrix

Confusion matrix		Predicted	
		0	1
Actual	0	278	28
	1	68	71

-테스트 데이터에 대한 confusion matrix

Confusion matrix		Predicted	
		0	1
Actual	0	115	18
	1	28	35

	TPR	Precision	TNR	ACC	BCR	F1
Train	0.5108	0.7172	0.9085	0.7843	0.6812	0.5967
Test	0.6034	0.6604	0.8647	0.7853	0.7223	0.6306

ACC의 경우 Train 데이터셋과 Test 데이터셋이 각각 0.7843, 0.7853(이후에도 동일한 순서로 서술)으로 거의 동일한 값이 나왔다. 이는 사람들이 약 78%정확도로 모델에 의해 적절하게 판별되었다고 해석할 수 있다.

그러나 데이터셋에 범주 불균형이 존재할 수 있기에 BCR, F1 지표도 고려해야한다. BCR의 경우에는 0.6812, 0.7223 값이 나왔다. 그리고 F1 지표의 경우에는 0.5967, 0.6306 값이 나왔다. BCR, F1 값이 ACC와 비교해서 떨어졌는데 이는 데이터셋에 범주 불균형이 존재함을 의미한다. 또한 F1 값을 제외하고 분류 모델이 나쁘지 않은 성능을 보인다고 볼 수 있다.

4. 학습 데이터와 테스트 데이터에 대한 AUROC를 산출하는 함수를 직접 작성하고 이를 사용하여 학습/테스트 데이터셋에 대한 AUROC를 비교해 보시오.

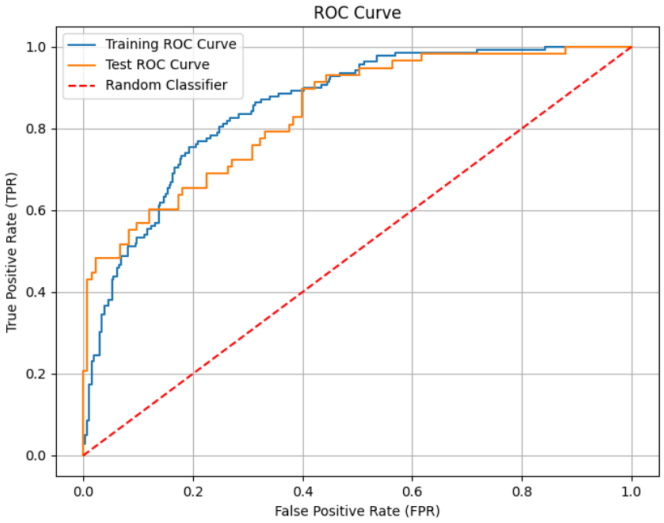
```
def calculate_auroc(y_true, y_score):
    sorted_index=np.argsort(y_score)[::-1]
    y_true=np.asarray(y_true)[sorted_index]
    n_pos=np.sum(y_true) #양성 클래스의 개수
    n_neg=len(y_true)-n_pos #음성 클래스의 개수
    tpr_list, fpr_list=[0], [0] #tpr, fpr 저장할 리스트
    tp, fp=0, 0
    for i in range(len(y_true)): #정렬된 요소에 대해 반복 수행
        if y_true[i] == 1:
            tp+=1
        else:
            fp+=1
        tpr_list.append(tp/n_pos)
        fpr_list.append(fp/n_neg)
    return np.trapz(tpr_list, fpr_list)  #TPR 및 FPR 사이의 면적을 계산하여 AUROC 값 반환
```

위 함수를 바탕으로 훈련 데이터와 테스트 데이터의 AUROC를 구하면 다음과 같다.

	AUROC
Train	0.8539
Test	0.8386

훈련 데이터와 테스트 데이트의 AUROC 값은 각각 0.8539, 0.8386으로 두 값 모두 준수하게 나왔음을 확인할 수 있다. 따라서 두 값 모두 낮게 나오거나 두 값의 차이가 크지 않기에 모델이 훈련세트에 과소적합 혹은 과대적합되지 않았다고 결론 내릴 수 있다. 따라서 테스트 데이터와 더불어 새로 추가되는 데이터에 대해서도 일반화를 잘할 수 있을 것이라 기대된다.

훈련 세트와 테스트 세트를 각각 하늘색, 주황색으로 표현하여 ROC Curve를 그리면 아래와 같다.



[Q7] 유의미한 변수 판단

-유의수준 0.05에서 유효한 변수의 수는 몇 개인지 확인하고 [Q6-1]의 결과와 비교하시오.

<Pregancies 변수를 제거한 뒤 LLR 결과>

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0291	0.133	-7.744	0.000	-1.290	-0.769
Glucose	1.2487	0.162	7.726	0.000	0.932	1.565
BloodPressure	-0.1877	0.141	-1.327	0.185	-0.465	0.090
SkinThickness	-0.0479	0.148	-0.324	0.746	-0.337	0.242
Insulin	-0.2559	0.146	-1.748	0.080	-0.543	0.031
BMI	0.5358	0.147	3.639	0.000	0.247	0.824
DiabetesPedigreeFunction	0.4020	0.122	3.290	0.001	0.163	0.641
Age	0.4829	0.136	3.547	0.000	0.216	0.750

유의수준 0.05에 대해 유의한 변수

➔ Glucose, BMI, DiabetesPedigreeFunction, Age

<전체 변수 LLR 결과>

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0359	0.134	-7.740	0.000	-1.298	-0.774
Pregnancies	0.3158	0.150	2.108	0.035	0.022	0.609
Glucose	1.2576	0.162	7.749	0.000	0.940	1.576
BloodPressure	-0.1858	0.142	-1.312	0.190	-0.463	0.092
SkinThickness	-0.0469	0.148	-0.317	0.751	-0.337	0.243
Insulin	-0.2475	0.147	-1.687	0.092	-0.535	0.040
BMI	0.5442	0.149	3.641	0.000	0.251	0.837
DiabetesPedigreeFunction	0.4128	0.124	3.340	0.001	0.171	0.655
Age	0.3000	0.161	1.864	0.062	-0.016	0.615

유의수준 0.05에 대해 유의한 변수

➔ Pregnancies, Glucose, BMI, DiabetesPedigreeFunction

앞서 Q5.에서 Preganancies 변수를 제거하고, 로지스틱 회귀 분석을 수행한 결과 위와 같이 유의수준 0.05에 대해 4개의 변수가 유의함을 확인할 수 있었다.

전체 변수에 대해서 동일하게 로지스틱 회귀 분석을 수행한 결과 Pregnancies 변수를 제거하고 수행한 결과와 동일하게 4개의 변수가 유의함을 확인할 수 있었다.

이 때, BMI, DiabetesPedigreeFunction는 [Q7]에서 선택한 변수 조합에 관계 없이 모두 유효함을 확인할 수 있었다. 이를 통해 BMI, DiabetesPedigreeFunction가 당뇨병 발병을 예측하는데 어느 정도 중요한 변수임을 짐작할 수 있다.

2. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출한 뒤, [Q6-3]의 결과와 비교해 보시오.

-학습 데이터에 대한 confusion matrix

Confusion matrix		Predicted	
		0	1
Actual	0	278	28
	1	64	75

-테스트 데이터에 대한 confusion matrix

Confusion matrix		Predicted	
		0	1
Actual	0	115	18
	1	22	36

		TPR	Precision	TNR	ACC	BCR	F1
[Q6-3]	Train	0.5108	0.7172	0.9085	0.7843	0.6812	0.5967
	Test	0.6034	0.6604	0.8647	0.7853	0.7223	0.6306
[Q7-2]	Train	0.5396	0.7282	0.9085,	0.7933	0.7002	0.6199
	Test	0.6207	0.6667	0.8647	0.7906	0.7326	0.6429

표의 [Q 6-3]이 전체 변수를 가지고 Logistic Regression을 수행한 결과이고, [Q 7-2]가 Pregnancies를 제외하고 Logistic Regression을 수행한 결과이다.

Train dataset과 Test dataset 모두 Pregnancies 변수를 제거한 뒤의 평가 지표가 그렇지 않은 경우(변수 전체를 사용한 경우)보다 좋게 나왔다. (단, TNR 값은 동일하게 나왔다.)

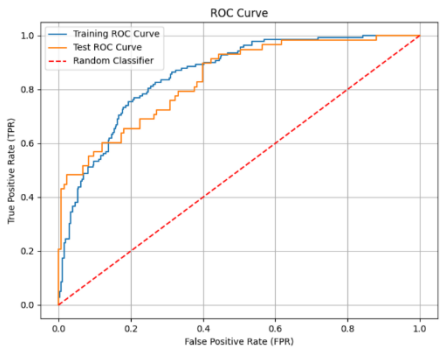
이는 변수의 개수를 줄임으로써 모델의 단순성을 보장하는 동시에 모델의 성능 또한 향상되었기에 변수를 삭제함으로써 모델을 단순화시키는 것이 합리적인 판단이라고 생각된다.

3. 학습/테스트 데이터셋에 대한 AUROC를 산출하여 [Q6-4]의 결과와 비교해 보시오.

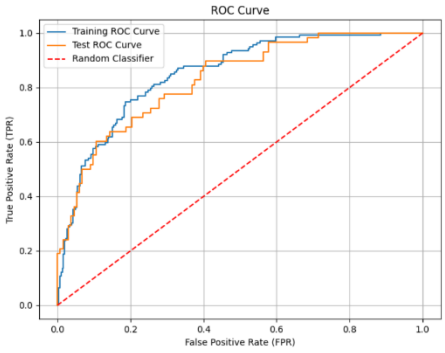
	AUROC(Before)	AUROC(After)
Train	0.8539	0.8503
Test	0.8386	0.8281

분류 모형의 또 다른 성능 평가 지표 중 하나인 AUROC의 경우 Pregnancies 변수를 제거하기 전(전체 변수를 사용했을 경우)이 Pregnancies 변수를 제거한 후보다 약간의 높은 값을 보인다는 것을 위 표를 통해 확인할 수 있다.

앞서 7-2에서 살펴본 정오 행렬 기반 지표와는 상반된 결과이다. 하지만, AUROC 값의 차이가 매우 미미하고 정오 행렬 기반 지표는 오히려 변수를 제거한 뒤의 성능이 향상되었기에 변수 제거를 통한 모델의 단순화를 고려하는 것이 바람직하다고 생각한다.



<전체 변수 사용 ROC Curve>



<변수 제거 ROC Curve>

위 그림의 좌측은 전체 변수를 사용했을 때의 ROC Curve를 우측은 Pregnancies 변수를 제거했을 때의 ROC Curve를 나타낸 것이다.

변수 제거 전과 후의 Train dataset과 Test dataset의 ROC Curve의 개형은 거의 유사한 형태를 띄기에 ROC Curve의 밑 면적인 AUROC 값 또한 유사하게 나왔다는 사실을 시각적으로 다시 한 번 확인할 수 있다.

[Q8] 변수 선택 방법론 적용

[Q6]에서 생성한 학습 데이터를 이용하여 Logistic Regression 에 Forward Selection, Backward Elimination, Stepwise Selection 을 적용해보시오. 각 방법론마다 Training dataset 에 대한 AUROC 및 소요 시간, Validation dataset 에 대한 AUROC, Accuracy, BCR, F1-Measure 를 산출하시오

각 변수 선택법 방식의 시간 측정 시 이전 선택법이 그 다음 선택법의 시간에 영향을 준다고 판단하여(실제로 단일 수행 시 약 9초 걸리는 작업이 1초 걸리는 현상이 발생하였다) 각각의 변수 선택법을 주석 처리하고 시간을 측정하였다.

8-1. Forward Selection

1) Forward Selection을 통해 선택된 변수는 다음과 같다.

['Glucose', 'BMI']

2) Forward Selection을 활용 Training dataset에 대한 AUROC 및 소요 시간은 다음과 같다.

Time taken for forward selection: 9.2621 seconds

Training AUROC: 0.7154

8-2. Backward Selection

1) Backward Selection을 통해 선택된 변수는 다음과 같다.

['Glucose', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']

2) Backward Selection을 활용 Training dataset에 대한 AUROC 및 소요 시간은 다음과 같다.

Time taken for backward selection: 9.7046 seconds

Training AUROC: 0.7288

앞서 Forward Selection으로 선택한 변수의 개수보다 3개의 변수를 더 선택했고, 소요 시간과 AUROC 값은 소폭 증가한 것을 확인할 수 있다.

즉, 3개의 변수를 제거하는 과정을 거쳤기에 Forward selection 방식보다 시간이 약간 더 걸렸다고 해석할 수 있다.

선택된 변수의 개수는 5개로 Forward selection 방식보다 3개 더 많기에 이에 따라 AUROC 값이 증가한 것으로 예상된다.

8-3. Stepwise Selection

1) Stepwise Elimination을 통해 선택된 변수는 다음과 같다.

['Glucose', 'BMI']

앞서 Forward selection과 동일한 변수가 선택되었다. 우연히 Stepwise selection 방식을 통해 Glucose, BMI 변수가 선택되었고 이후 또 다른 변수를 추가하거나 추가한 변수 중 1개를 제거하는 방식이 step 진행 기준을 만족하지 못해 나타난 결과라고 해석할 수 있다.

2) Stepwise Elimination을 활용 Training dataset에 대한 AUROC 및 소요 시간은 다음과 같다.

Time taken for backward selection: 9.5254 seconds

Training AUROC: 0.7261

위 세 변수 선택법의 Training dataset에 대한 소요시간과 AUROC 값을 비교하여 정리하면,

소요시간: Forward<Stepwise<Backward

AUROC: Forward<Stepwise<Backward

이론적으로 최대로 탐색 가능한 경우의 수는 Forward=Backward<Stepwise이다. 그러나 소요시간은 Forward<Stepwise<Backward 순으로 나왔다. 이에 대한 이유로 Backward selection 수행 시 총 3개의 변수가 제거되는 3단계를 반복했으나 Stepwise selection 수행 시에는 총 2개의 변수가 선택되는 2단계를 반복했기 때문이라고 생각한다. 그럼에도 두 변수 선택법 간의 시간 차이가 크지 않은 이유는 Stepwise selection 수행 시 각 step마다 변수 선택과 제거 두 가지 방식을 모두 고려하기에 단계가 적음에도 시간 차이는 얼마 나지 않는 것이다.

8-4. Test dataset에 대한 평가 지표

	AUROC	Accuracy	BCR	F1-Measure
Forward	0.7525	0.7801	0.6717	0.58
Backward	0.7835	0.8063	0.7195	0.6408
Stepwise	0.7855	0.8115	0.7225	0.6471

학습 데이터셋에 대해서는 AUROC 기준으로 Backward Elimination이 가장 좋은 성능을 보였지만, 테스트 데이터셋에서는 AUROC 및 3개의 균형도 값 기준으로 Stepwise Selection이 가장 좋은 성능을 보였다.

Backward Elimination에서 삭제된 변수가 3개임을 고려하면 학습 데이터셋에서는 Backward Elimination 상대적으로 Forward selection 및 Stepwise selection보다 많은 변수를 포함하고 있어 높은 성능을 보이 지만, 과적합 및 다중공선성 등의 문제로 테스트 데이터셋에서는 상대적으로 성능이 낮게 나타났을 것으로 예상된다.

[Q9] GA 기반 변수 선정

AUROC를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오. 작성한 함수를 이용하여 GA 기반 변수 선택을 수행하고, 선택된 변수를 사용한 Logistic Regression의 Validation dataset에 대한 분류 성능(AUROC, Accuracy, BCR, F1-Measure), 변수 감소율, 수행 시간의 세 가지 관점에서 Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용한 Logistic Regression과 비교해보시오

AUROC를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 기반으로 선택된 변수는 다음과 같았다.

'Glucose', 'BloodPressure', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'

-분류 성능

	AUROC	Accuracy	BCR	F1-Measure
Forward	0.7525	0.7801	0.6717	0.58
Backward	0.7835	0.8063	0.7195	0.6408
Stepwise	0.7855	0.8115	0.7225	0.6471
Genetic	0.8373	0.7958	0.7464	0.6846

Validation dataset에 대한 분류 성능의 관점으로는 Genetic>Stepwise>Bakward>Forward 순으로 나타났다고 할 수 있다. 이는 기존 휴리스틱 기법들보다는 최적 변수 집합을 찾을 가능성이 높은 방식이 유전 알고리즘이기 때문이다. 또한 Stepwise 방식은 Forward, Backward 방식이 혼재된 변수 선택법이기에 Backward, Forward 방식보다 분류 성능이 좋게 나왔다고 볼 수 있다. 따라서 분류 성능은 이론적인 관점과 동일하게 나왔다고 볼 수 있다.

한편, 모든 방법론에 대해서 AUROC>Accuracy>BCR> F1-Measure의 관계를 보였다. 이는 데이터셋에 범주 불균형이 있기 때문에 나타난 현상이라고 생각된다.

-변수 감소율

	선택된 변수 개수	변수 감소율
Forward	2	75%
Backward	5	37.5%
Stepwise	2	75%
Genetic	6	25%

변수 감소율 관점에서는 Forward, Stepwise> Backward> Genetic 순으로 효율적인 방법론이라고 할 수 있다. 같은 성능의 모델에 대해서는 변수의 개수가 적을수록 효율적인 모델이라고 할 수 있기 때문이다.

-수행 시간

	수행시간
Forward	9.2621 seconds
Backward	9.7046 seconds
Stepwise	9.5254 seconds
Genetic	1.0669 seconds

수행시간의 측면으로 봤을 때, Genetic>Forward>Stepwise>Backward 방식 순으로 효율적인 모델이라고 할 수 있다. 특히 Forward, Backward, Stepwise 세 가지 방식은 수행시간이 거의 유사하게 나와 차이가 없다고 봐도 무방하다고 생각한다.

그러나 Genetic 알고리즘의 경우 다른 변수 선택법과 다르게 매우 짧은 시간을 기록하였다. 이론적으로 봤을 때, 고려해야 할 경우의 수가 상대적으로 많은 Genetic 알고리즘의 수행 시간이 가장 길 것이라는 예상과는 상반된 결과이다. 이 때 시간 차이가 무시할 수 있는 정도로 작지 않기 때문에 재측정이 필요할 것으로 생각된다.

-종합 평가

분류 성능의 관점으로 봤을 때, Stepwise방식이나 Genetic 알고리즘 방식을 선택하는 것이 옳다고 생각한다. 변수 감소율 측면에서 봤을 때는 Forward 방식이나 Stepwise 방식을 선택하는 것이 옳다고 생각한다. 수행 시간의 측면에서 봤을 때는 Genetic 알고리즘의 경우 재측정이 필요하다고 판단되며 재 측정 시 이론상으로 다른 선택법에 비해 시간이 많이 나올 것으로 예상되기에 세 가지 변수 선택법인 Forward, Backward, Stepwise 방식 중 어느 방식을 선택해도 무방하다고 생각된다. 따라서 종합적으로 고려해봤을 때 분류 성능이 좋고, 변수 개수의 감소가 많으며, 수행 시간이 적은 Stepwise selection을 사용하는 것이 가장 효율적인 방식이라고 생각된다.

[Q10] GA 하이퍼파라미터 변경

Genetic Algorithm에서 변경 가능한 하이퍼파라미터들의 값인 population size, Cross-over rate, Mutation rate을 각각 (10, 30, 50), (0.1 ,0.3,0.5), (0.05, 0.1, 0.2)에서 하나씩 뽑아 총 27가지의 하이퍼파라미터 조합에 대해 GA를 적용하였다. 27가지 하이퍼파라미터 조합에 대한 GA 결과는 크게 변수 관점, 하이퍼파라미터 관점 2가지 관점으로 설명하겠다.

-변수 관점

빈도	변수 수	변수 선택	비율
3	7	Pregnancies, Glucose, BloodPressure Insulin, BMI, DiabetesPedigreeFunction, Age	11.1%
12	6	Pregnancies, Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age	44.4%
6	6	Glucose, BloodPressure, Insulin BMI, DiabetesPedigreeFunction, Age	22.2%
3	5	Pregnancies, Glucose, Insulin, BMI, Age	11.1%
3	5	Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age	11.1%

하이퍼파라미터 값의 변화에 관계없이 모든 경우에서 공통적으로 선택된 변수는 위 표에서 빨간색으로 나타냈다. 그 변수들은 Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age이다.

한편, 하이퍼파라미터 값의 변화에 관계없이 모든 경우에서 공통적으로 제외된 변수는 SkinThickness였다.

종합하여 정리하면 Genetic 알고리즘을 통해 변수 선택 시 hyperparameter와 상관없이 Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age는 선택되고, SkinThickness는 가장 먼저 제외된다는 것을 확인할 수 있다.

따라서 **혈당, 인슐린 수치, 체질량 지수, 가족력, 나이는 robust하게 중요한 변수임을 알 수 있고, 중요하지 변수 1가지를 뽑으라면 피부 두께를 뽑을 수 있을 것이다**.

-하이퍼파라미터 관점

Population size를 고정하고(10, 30, 50), 각각의 경우에 대해 총 9가지(Cross-over rate, Mutation rate 조합 수)를 관찰하면 10, 30의 Population size에 대해서는 mutation rate가 증가함에 따라 선택되는 변수의 수가 감소함을 확인할 수 있다. (mutation rate가 0.2인 경우에만 5개의 변수가 선택되었다). 그러나 Population size가 50일 때 mutation rate가 증가함에 따라 선택되는 변수의 수가 줄어들 것이라는 예상과는 다르게 오히려 증가하였다 (mutation rate가 0.2로 변함에 따라 선택되는 변수의 수가 6개에서 7개로 늘어났다).

그럼에도 Population size, Cross-over rate에 관계없이 mutation rate가 증가함에 따라 선택되는 변수의 수는 감소한다고 생각한다. 왜냐하면 Population size가 50인 한 가지 경우에서만 이러한 현상이 발생하여 Population size가 특정 값 이상이 되면 mutation rate가 증가함에 따라 선택되는 변수의 수가 증가한다고 결론 짓기에는 표본이 적다고 판단하였기 때문이다. 이에 반해 상대적으로 Population이 10, 30 두 가지 경우에서 mutation rate가 증가함에 따라 선택되는 변수의 수는 감소하는 현상이 일어났기에 더 합리적인 결론이라고 할 수 있다.

그럼에도 경우의 수가 2개로 여전히 크진 않기에 더 정확한 결론을 내리기 위해서는 Population size를 70, 90일 때 동일한 작업을 수행 및 관찰하는 것이 합리적이라고 생각된다.

-종합 결론

	공통 변수
[Q8]	Glucose, BMI
[Q10]	Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age

[Q8]에서 3가지 변수 선택법으로 추출된 공통 변수는 Glucose, BMI였다. 그리고 하이퍼파라미터 값을 변화시키며 찾아낸 공통 변수는 Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age였다. 또한 GA를 통해 선택된 변수 조합은 [Q8]의 Backward Selection 방식에서 선택된 변수 조합과 동일하였다.

Glucose, BMI는 변수 선택법이 변화하더라도 GA의 하이퍼파라미터 값이 변화하더라도 항상 선택되는 robust한 변수임을 알 수 있다. 따라서 혈당과 체질량지수는 종속 변수인 당뇨병 유무를 예측하는데 상당히 중요한 변수임을 알 수 있다**. 즉, 혈당이 높고 체질량지수가 높을수록 당뇨병 발병 확률이 높다고 할 수 있다.**

그 외 Insulin, DiabetesPedigreeFunction, Age는 Forward selection, Stepwise selection에서 선택되지는 않았으나 나머지 Stepwise, GA에서 선택되었기에 당뇨병 유무를 예측하는데 어느 정도 중요한 변수임을 확인할 수 있다. 즉, 인슐린 수치가 낮을수록 가족력이 있을수록 나이가 높을수록 당뇨병 확률이 높다고 할 수 있다(앞선 Glucose, BMI보다는 설명력이 떨어진다).

모든 변수 선택법, GA 알고리즘에서 유일하게 공통적으로 선택되지 않은 변수는 SkinThickness였다. 이를 통해 **피부 두께는 당뇨병 발병을 예측하는데 중요하지 않은 변수라고 할 수 있다**. 이는 앞서 살펴본 [Q2]의 예상과 일치하는 결과이다.