[다변량데이터분석][과제2][2020170816][김소민]

[Q1] 본인이 스스로 Logistic Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

⇨ Logistic Regression은 이진형 종속변수에 대해 회귀식의 형태로 모형을 추정하는 것을 목적으로 한다. 이를 고려했을 때, 데이터셋을 결정하기 위해 고려할 가장 중요한 요소는 바로 '무엇을 두 가지로 분류할 것인가?'일 것이라 판단했다. 이것을 염두에 두며 data repository를 둘러본 결과, 여러 요인에 따른 질병 발병 여부가 Logistic Regression과 어울린다는 결론을 내렸다. 다양한 데이터셋을 살펴본 후 정한 데이터셋은 Heart Failure Clinical Records Dataset으로, 심부전 환자의 심혈관건강 요소와 사망 여부를 나타낸 데이터이다. 해당 데이터셋의 종속변수를 제외한 칼럼 수는 12개로, Forward Selection, Backward Elimination, Stepwise Selection, 그리고 Genetic Algorithm을 수행하기에 적합한 개수라고 생각했다. 또한, 각 칼럼이 나타내는 나이, 당뇨병 여부, 흡연 여부 등의 건강과 관련된 요소들이 사망 여부와 관련이 있다는 사실이 직관적으로 받아들여진다고 생각하였다. 이러한 근거들을 통해 해당 데이터셋이 사망 여부를 종속변수로 하고, 나머지 변수들을 설명변수로 설정한 Logistic Regression을 수행하기에 적합한 데이터라는 결론을 내렸다.

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 두 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

⇨앞서 언급했듯이, 해당 Logistic Regression의 목표는 심부전 환자의 여러 심혈관건강 지표에 따른 사망여부 판별이다. 따라서 사망 여부인 DEATH_EVENT가 종속 변수이고, 나머지 변수들은 설명변수가 된다. 다음은 변수의 목록이다.

<설명변수>

age (40~95): 심부전 환자의 나이

anaemia (0 or 1): 빈혈 여부 (1이면 빈혈)

creatinine_phosphokinase(23~7861): 혈중 CPK(크레아틴 포스포키나아제) 농도(μg/L)

diabetes(0 or 1): 당뇨병 여부 (1이면 당뇨병)

ejection_fraction(14~80): 좌심실 구혈률(%)

high_blood_pressure(0 or 1): 고혈압 여부 (1이면 고혈압)

platelets(25,100~850,000): 혈중 혈소판수치(개/μL)

serum_creatinine(0.5~9.4): 혈청 크레아티닌수치(mg/dL)

serum_sodium(113~148): 혈청 나트륨수치(mEq/L)

sex(0 or 1): 성별(1이면 남자)

smoking(0 or 1): 흡연 여부(1이면 흡연자)

time(4~285): 심부전 발생 후 사망 여부 확인까지 걸린 기간(일수)

<종속 변수>

DEATH_EVENT(0 or 1): 사망 여부(1이면 사망)

1. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

⇨ [creatinine_phosphokinase, high_blood_pressure, smoking]

혈중 CPK의 농도가 높아지는 경우는 보통 근육의 손상으로 인해 일어난다. 이러한 특성 때문에 CPK 수치 측정은 심근경색의 조기진단으로 유용한 검사라고 한다. 따라서 혈중 CPK의 농도를 나타내는 creatinine_phosphokinase는 심근경색과 깊은 관련이 있어 DEATH_EVENT와 높은 상관관계가 있을 것이라고 예상된다. 또한, 높은 혈압에 지속적으로 노출되면 혈관이 손상되고 동맥경화가 진행하면서 각종 합병증이 발생한다는 점에서 high_blood_pressure 역시 DEATH_EVENT와 높은 상관관계가 있을 것이라고 생각한다. 한편, 흡연은 체내 산소농도를 낮추어 심근에 산소부족상태를 초래할 위험이 있다. 이러한 심근의 상태는 심근경색으로 이어질 수 있으므로 smoking이 종속변수인 DEATH_EVENT와 상당히 높은 상관관계가 있을 것이라 판단했다.

2. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수

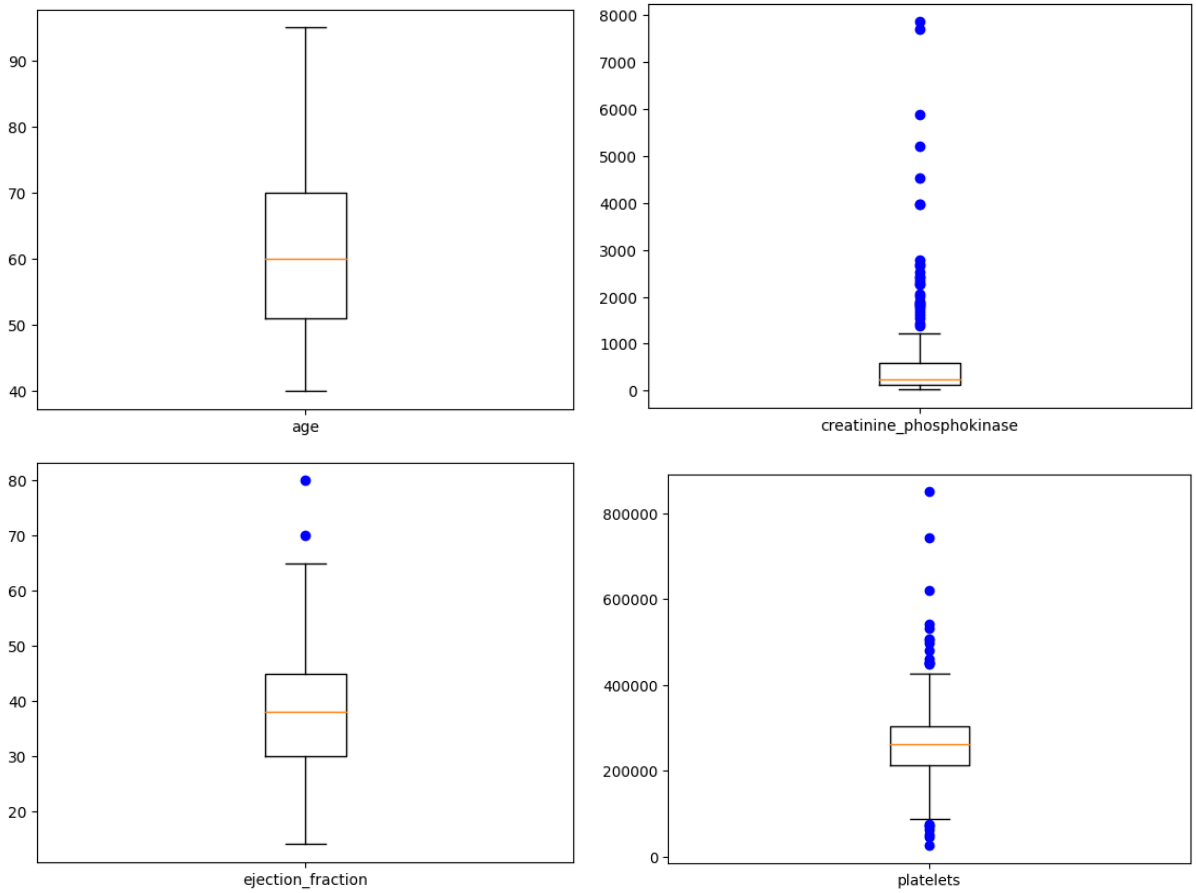
들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

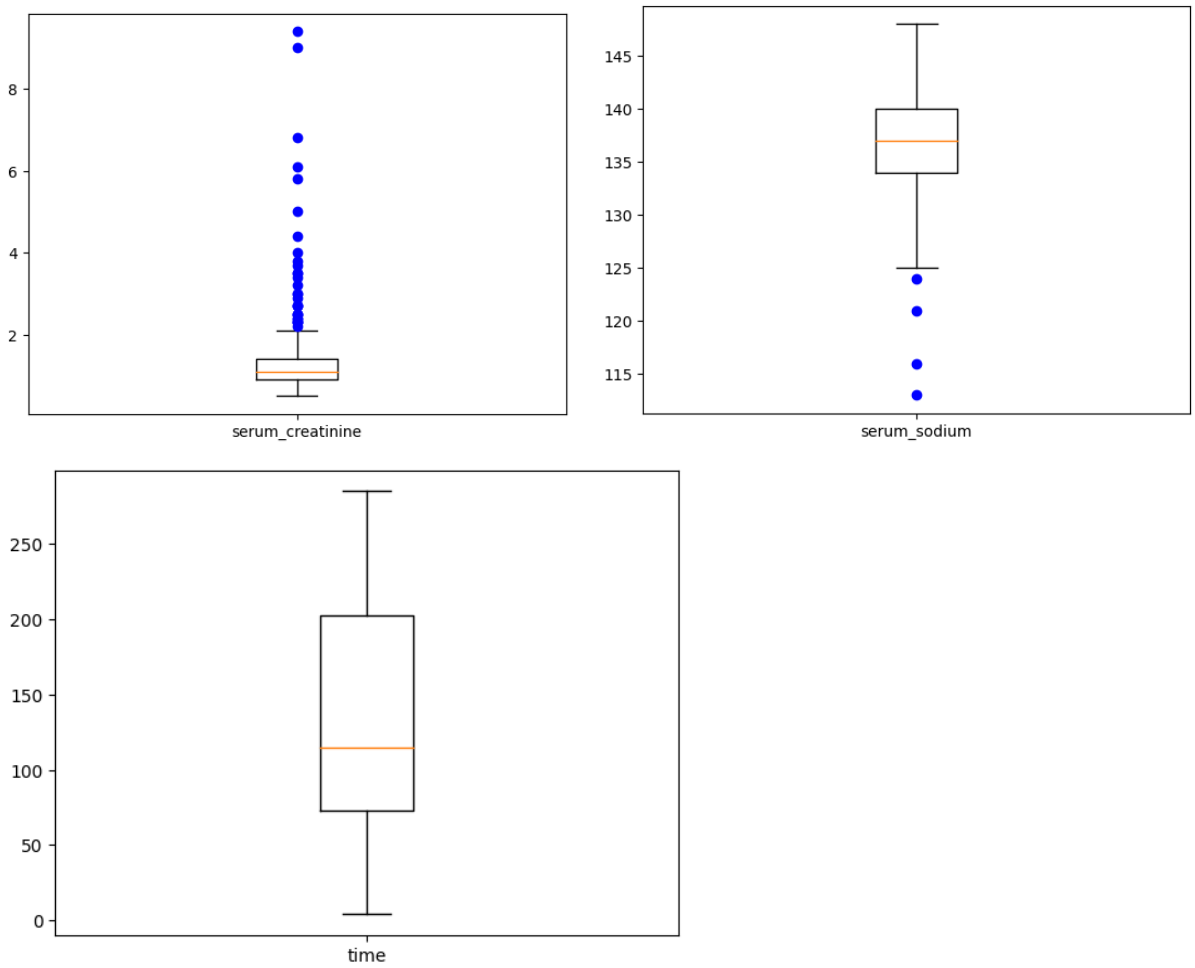
⇨ [age, diabetes, sex, time]

DEATH_EVENT는 심부전 환자를 일정 일수 이후 다시 찾아봤을 때의 사망 여부이므로, 환자의 나이인 age와는 큰 관련이 없을 것이라 생각했다. 여기서의 일수인 time이 4~285일으로 인간의 수명을 좌우할 정도로 길지 않기 때문이다. 또한, 당뇨병 여부인 diabetes는 급성으로 사람의 목숨을 앗아가는 병이 아니기 때문에 종속변수를 예측하는 데 필요하지 않을 것이라 판단했다. 성별(sex) 역시 짧은 시간의 관찰 일수를 고려하면 죽음에 큰 영향을 미치지 않을 것이라 생각했다.

[Q3] 모든 연속형 숫자 형태를 갖는(명목형 변수 제외) 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

	Attribute	mean	std	skewness	kurtosis
0	age	60.833893	11.894809	-0.201793	0.420937
1	creatinine_phosphokinase	581.839465	970.287881	24.710458	4.440689
2	ejection_fraction	38.083612	11.834841	0.020720	0.552593
3	platelets	263358.029264	97804.236869	6.085906	1.454975
4	serum_creatinine	1.393880	1.034510	25.378346	4.433610
5	serum_sodium	136.625418	4.412477	4.031142	-1.042870
6	time	130.260870	77.614208	-1.211874	0.127161





⇨ 데이터가 정규분포를 따르는 지 확인하기 위해 Mean, Standard deviation, Skewness, Kurtosis를 계산하고, Box Plot을 도시하였다. 이상적인 정규분포의 데이터는 skewness와 kurtosis 값이 모두 0이어야 하지만, 현실에서 그 정도로 정규분포를 따르는 데이터는 거의 존재하지 않는다. 그러므로 Kline(2005)의 기준에 따라 Skewness의 절댓값이 3보다 작고, Kurtosis의 절댓값이 8보다 작으면 정규 분포를 따른다고 가정한다. 이것과 Box Plot의 모양을 모두 고려하면 정규분포를 따르는 변수는 age, ejection_fraction, 그리고 time으로 세 개 존재한다고 볼 수 있다.

한편, 정규성을 검정하기 위해 추가적인 Shaprio-Wilks test를 진행하였다. 귀무가설을 데이터의 분포가 정규분포를 따른다고 설정하고 test를 진행한 결과는 다음과 같다.

	test_statistic	p_value
age	0.975469	5.347659e-05
creatinine_phosphokinase	0.514264	7.050557e-28
ejection_fraction	0.947316	7.215172e-09
platelets	0.911510	2.883687e-12
serum_creatinine	0.551466	5.392758e-27
serum_sodium	0.939031	9.220169e-10
time	0.946783	6.284944e-09

모든 변수의 p-value 값이 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 따라서, 어떤 변수도 정규분포를 따르지 않는다고 최종적으로 결론을 내릴 수 있다.

[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해보시오.

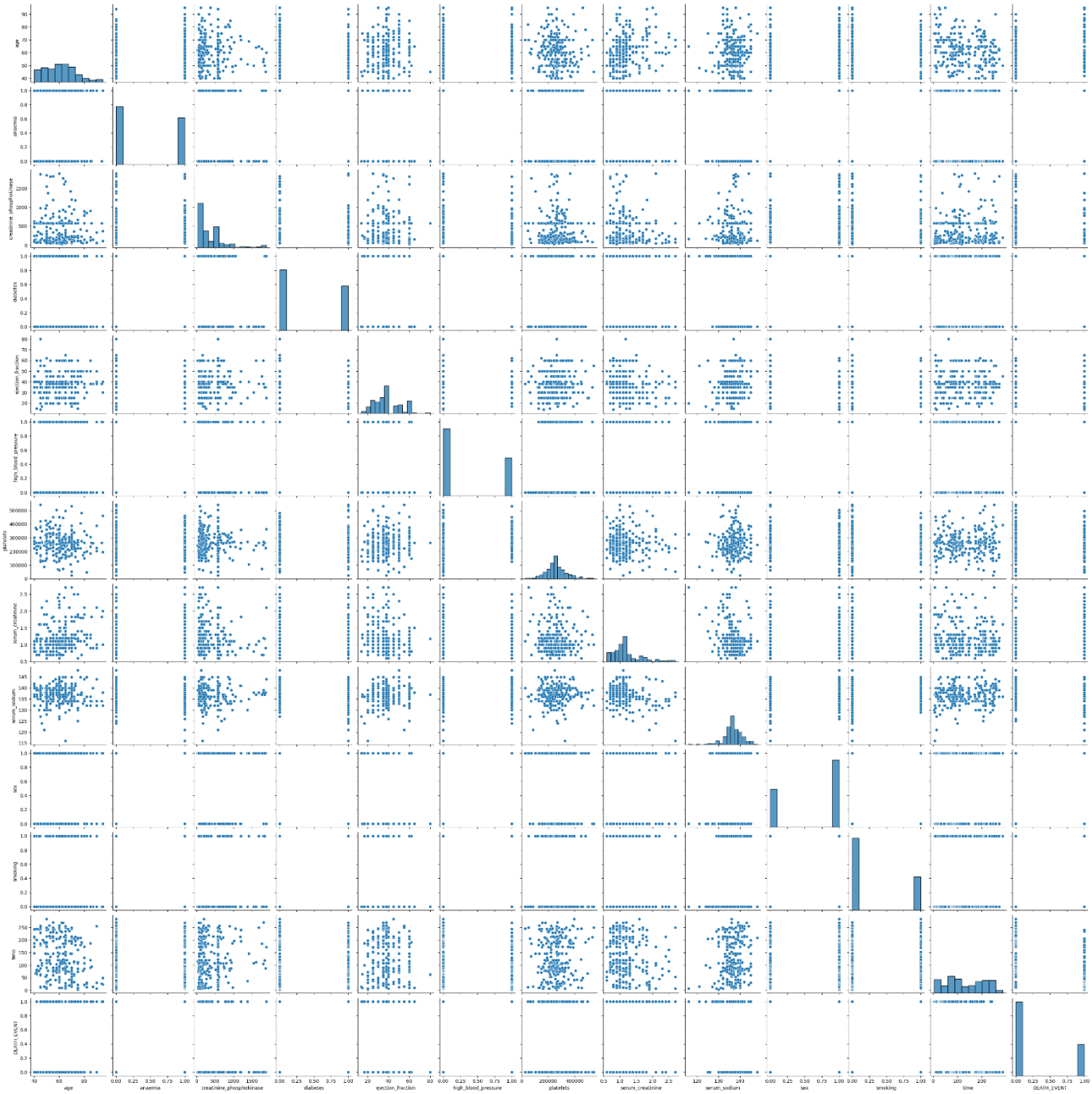
⇨ Box plot에서 whiskers 밖에 존재하는 값들을 이상치라고 부른다. Whisker의 범위는 [Lower quartile(q1) - 1.5 IQR, Upper quartile(q3) + 1.5 IQR]이기 때문에, 일반적인 이상치 조건은 q1 – 1.5 IQR보다 작거나 q3 + 1.5 IQR보다 큰 값이 된다. 하지만 현재 데이터셋의 총 데이터 수는 299개로 상당히 적은 편이다. 따라서 이상치 조건을 q1 – 3 IQR보다 작거나 q3 + 3 IQR보다 큰 객체로 정의하도록 하겠다. 해당되는 객체들을 데이터셋에서 제거해보면 tuple수가 299에서 262로 줄어든 것을 알 수 있다.

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.0	1.9	130	1	0	4	1
1	65.0	0	146	0	20	0	162000.0	1.3	129	1	1	7	1
2	50.0	1	111	0	20	0	210000.0	1.9	137	1	0	7	1
3	65.0	1	160	1	20	0	327000.0	2.7	116	0	0	8	1
4	90.0	1	47	0	40	1	204000.0	2.1	132	1	1	8	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
257	52.0	0	190	1	38	0	382000.0	1.0	140	1	1	258	0
258	63.0	1	103	1	35	0	179000.0	0.9	136	1	1	270	0
259	62.0	0	61	1	38	1	155000.0	1.1	143	1	1	270	0
260	55.0	0	1820	0	38	0	270000.0	1.2	139	0	0	271	0
261	50.0	0	196	0	45	0	395000.0	1.6	136	1	1	285	0

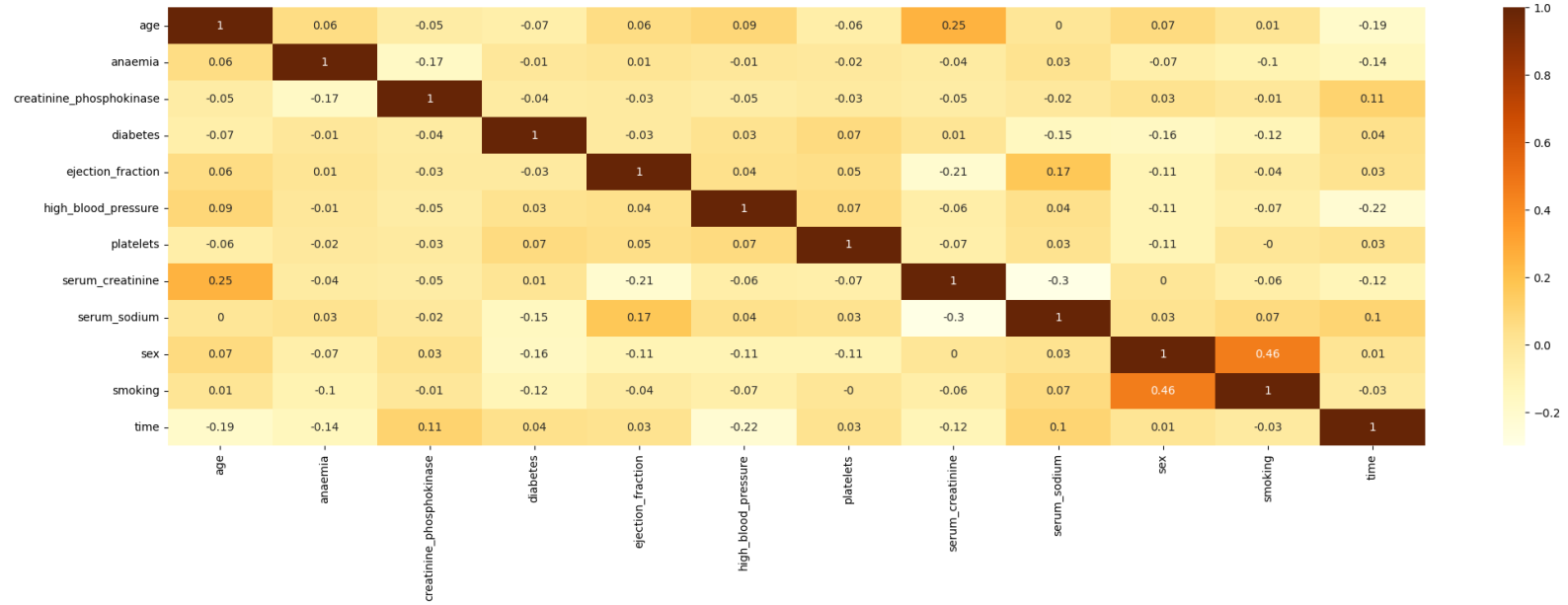
262 rows × 13 columns

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot)를 도시하고 적절한 정량적 지표를 사용하여 상관관계를 판단해보시오.



해당 scatter plot만으로는 변수 사이의 상관관계를 파악하기 어려워 correlation chart를 추가적으로 살펴보았다.



1. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?

⇨ sex와 smoking, 그리고 serum_creatinine과 age가 상대적으로 강한 상관관계가 있다고 할 수 있다. 하지만 이는 상대적일 뿐, 일반적인 기준에 대해서는 강하다고 보기 어렵다. 각각 0.46과 0.25로 1보다는 0에 더 가까운 값이기 때문이다.

2. 강한 상관관계가 존재하는 변수 조합들 중에 대표 변수를 하나씩만 선택해서 전체 변수의 개수를 감소시켜 보시오. ([Q7]에서 사용함)

⇨ 앞서 Q2에서 종속변수와 큰 관련이 없을 것이라고 예상한 sex와 age를 제거하도록 하겠다. 이렇게 감소된 변수의 목록은 다음과 같다.

[anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, smoking, time]

[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.

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

	P-value
constant	1.0000
age	0.9990
anaemia	1.0000
creatinine_phosphokinase	0.8464
diabetes	1.0000
ejection_fraction	0.9977
high_blood_pressure	1.0000
platelets	0.5403
serum_creatinine	1.0000
serum_sodium	0.9996
sex	1.0000
smoking	1.0000
time	0.9540

⇨ 모든 변수가 유의수준 0.05을 크게 상회하는 값을 가진다. 이러한 결과는 그 어떤 변수도 DEATH_EVENT를 예측하는데 유효하지 않다는 것을 의미한다. 하지만 앞서 했던 분석과 상식적인 부분을 생각하면 이러한 결과 자체가 일반적이지 않다는 것을 알 수 있다. 필자는 이러한 결과가 나온 이유를 데이터의 크기가 지나치게 작기 때문이라는 결론을 내렸다. 이상치를 제거하고, 70:30의 비율로 학습 데이터와 테스트를 나누는 과정에서 최종적으로 사용된 학습 데이터의 수가 181개로 상당히 적어졌다. 더욱 정확한 모델을 만들기 위해서는 더 많은 양의 데이터가 필요할 것이다.

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

⇨ [age, diabetes, sex, time] 모두 모델링 측면에서 유효하지 않다. 하지만 앞서 언급했듯이

해당 분석 결과에 따르면 모든 변수가 Logistic Regression 모델링에서 유효하지 않다고 나왔기 때문에 신뢰도가 떨어진다. 크기가 더 큰 데이터로 분석을 진행하면 다른 결과가 나올 가능성이 있다.

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

<학습 데이터>

```
Confusion Matrix:
[[108   0]
 [ 73   0]]
```

```
0.5966850828729282, 0.5, 0.0
```

<테스트 데이터>

```
Confusion Matrix:
[[76  0]
 [ 5  0]]
```

ACC	BCR	F1
0.938272	0.5	0.0

⇨ 학습 데이터와 테스트 데이터의 Balanced Correction Rate와 F1-Measure는 동일하지만, Simple Accuracy는 테스트 데이터가 0.938272로 학습 데이터의 0.596685보다 크다.

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

```
AUROC of Test Data: 0.5
AUROC of Training Data: 0.5
```

⇨ 학습 데이터와 테스트 데이터의 AUROC는 둘 다 0.5로, random classifier와 동일한 성능을 보여준다.

[Q7] [Q5]에서 변수 간 상관관계를 기준으로 선택한 변수들만을 사용하여 [Q6]에서 사용한 학습/테스트 70:30분할 데이터으로 Logistic Regression 모델을 학습해 보시오.

	P-value
constant	1.0000
anaemia	1.0000
creatinine_phosphokinase	0.8450
diabetes	1.0000
ejection_fraction	0.9977
high_blood_pressure	1.0000
platelets	0.5398
serum_creatinine	1.0000
serum_sodium	0.9996
smoking	1.0000
time	0.9536

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

⇨ 여전히 유의수준 0.05에서 유효한 변수는 존재하지 않는다. 즉, 거시적인 관점에서는 [Q6-1]의 결과와 달라진 점이 없다. 하지만 creatinine_phosphokinase의 coefficient가 0.8464에서 0.8450으로 변하고, platelets의 coefficient가 0.5403에서 0.5398로 변하는 등 미세한 차이는 존재한다.

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

<학습 데이터>

```
Confusion Matrix:
[[108   0]
 [ 73   0]]
```

<테스트 데이터>

```
Confusion Matrix:
[[76  0]
 [ 5  0]]
```

0.5966850828729282, 0.5, 0.0 0.938272 0.5 0.0

⇨ 원래의 Confusion Matrix와 달라진 점이 없기 때문에, 해당 Simple Accuracy, Balanced Correction Rate, F1-Measure는 앞선 [Q6-3] 결과와 정확하게 일치한다.

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

```
AUROC of Test Data: 0.5
AUROC of Training Data: 0.5
```

⇨ 앞서 언급한 바와 같이, 원래의 Confusion Matrix와 달라진 점이 없기 때문에, 해당 AUROC값은 앞선 [Q6-4]의 결과와 정확하게 일치한다.

⇨ 이러한 결과들은 age와 sex가 종속변수를 예측하는데 아무런 영향을 끼치지 않았다는 것을 증명한다. 하지만 이러한 결과 역시 지나치게 작은 데이터셋의 영향을 받았을 가능성이 있어, 더욱 row의 수가 많은 데이터셋으로 동일한 Logistic Regression을 진행할 필요성이 있다.

[Q8] [Q6]에서 생성한 학습 데이터를 이용하여 Logistic Regression에 Forward Selection, Backward Elimination, Stepwise Selection을 적용해보시오. 각 방법론마다 Training dataset에 대한 AUROC 및 소요 시간, Validation dataset에 대한 AUROC, Accuracy, BCR, F1-Measure를 산출하시오.

1) Forward Selection

	features	coefs
0	Constant	1.558828
1	ejection_fraction	-0.052597

AUROC of Training Data: 0.656773211567732

Time elapsed: 6.203265428543091

AUROC of Test Data: 0.7539473684210527

Acc	BCR	F1
0.888889	0.753947	0.4

2) Backward Elimination

	features	coefs
0	Constant	0.000311
1	age	0.062736
2	anaemia	-0.000420
3	creatinine_phosphokinase	0.000364
4	diabetes	0.002261
5	ejection_fraction	-0.068162
6	high_blood_pressure	-0.001004
7	serum_creatinine	0.009965
8	serum_sodium	0.000362
9	sex	-0.001874
10	smoking	-0.000710
11	time	-0.022695

```
AUROC of Training Data: 0.791349568746829

Time elapsed: 22.662732362747192

AUROC of Test Data: 0.5
```

Acc	BCR	F1
0.938272	0.500000	0.0

3) Stepwise Selection

	features	coefs
0	Constant	-0.876262
1	age	0.029930
2	anaemia	-0.072260
3	ejection_fraction	-0.073488
4	high_blood_pressure	0.141545
5	serum_creatinine	1.135558
6	smoking	-0.003003

AUROC of Training Data: 0.7571029934043633

Time elapsed: 16.390979051589966

AUROC of Test Data: 0.9078947368421052

Acc	BCR	F1
0.827160	0.907895	0.416667

[Q9] 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.

1) 분류성능

다음은 각 변수 선택 기법의 Accuracy, BCR, F1-Measure이다.

	ACC	BCR	F1
Forward Selection	0.888889	0.753947	0.400000
Backward Elimination	0.938272	0.500000	0.000000
Stepwise Selection	0.827160	0.907895	0.416667
Genetic Algorithm	0.938272	0.500000	0.000000

단순 정확도(Accuracy)는 Backward Elimination과 Genetic Algorithm이 공동 1위로 가장 뛰어나지만, 균형 정확도는 Stepwise Selection이 0.907895를 기록하며 압도적인 1위를 하였다. 추가로 F1-measure 역시 Stepwise Selection이 가장 좋은 값을 보였다. 다음은 AUROC를 살펴보겠다. 우선 Genetic Algorithm의 AUROC는 다음과 같다.

```
AUROC of Test Data: 0.5
```

이를 반영하여 모든 AUROC 를 표로 정리하면 다음과 같다

Forward Selection	0.753947
Backward Elimination	0.5
Stepwise Selection	0.907895
Genetic Algorithm	0.5

여기서도 Stepwise Selection 이 0.907895 로 가장 좋은 값을 보여주고 있다. 이를 모두 고려하면 Backward Elimination 과 Genetic Algorithm 은 동일한 분류성능을 보여주고 있는데, 오직 Accuracy 값만 높고 다른 지표들은 Random Classifier 와 다를 바가 없는 수준을 나타낸다. 그렇다면 의미 있는 결과를 보인 것은 Forward Selection 과 Stepwise Selection 뿐이다. 이 때 Forward Selection 은 Stepwise Selection 보다 단순 정확도 측면에서만 좋은 성능을 보여주고 나머지 AUROC, BCR, 그리고 F1-Measure 에서는

떨어지는 성능을 보였다. 따라서 분류 성능 측면에서 가장 좋은 성능을 보여준 것은 Stepwise Selection 이라 할 수 있다. 각 지표의 의미를 고려하면 Stepwise Selection 은 데이터의 불균형으로 인해 단순 정확도는 떨어지지만, TPR 과 TNR 을 종합적으로 고려한 균형 정확도에서는 높은 값을 보였다. 또한, AUROC 가 매우 높아 평균적으로 좋은 성능을 보이고, F1-Measure 도 높은 값을 보여 상한치도 가장 높았다. 해당 데이터셋에 대해서는 가장 이상적이며 현실적으로도 좋은 기법이 Stepwise Selection 이라는 것이다.

2) 변수 감소율

다음의 표는 각 변수 선택 기법을 통해 선택한 변수를 나타낸다.

Forward Selection	['ejection_fraction']
Backward Elimination	['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes' 'ejection_fraction', 'high_blood_pressure', 'serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time']
Stepwise Selection	['age', 'anaemia', 'ejection_fraction', 'high_blood_pressure', 'serum_creatinine', 'smoking']
Genetic Algorithm	['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking', 'time']

전체 설명 변수 수가 12 개였으므로 Forward Selection 의 변수 감소율은 91.67%, Backward Elimination 의 변수 감소율은 8.33%, Stepwise Selection 의 변수 감소율은 50%, 그리고 Genetic Algorithm 의 변수 감소율은 16.67%이다. 따라서 변수 감소율의 관점에서는 Forward Selection 이 압도적으로 뛰어난 결과를 보였음을 알 수 있다.

3) 수행시간

다음은 Genetic Algorithm 의 수행시간이다.

```
Time elapsed: 8.40589189529419
```

이를 고려하여 모든 변수 선택 기법의 수행시간을 표로 나타내면 다음과 같다.

Forward Selection	6.203 (s)

Backward Elimination	22.663 (s)
Stepwise Selection	16.391 (s)
Genetic Algorithm	8.406 (s)

따라서 수행 시간의 관점에서는 Forward Selection 이 가장 우수한 성능을 보이고, 그 다음으로는 Genetic Algorithm, Stepwise Selection, 그리고 Backward Elimination 순으로 좋은 성능을 보인다.

[Q10] Genetic Algorithm에서 변경 가능한 하이퍼파라미터들(population size, Cross-over rate, Mutation rate 등) 중 세 가지를 선택하고 각각의 하이퍼파라미터마다 최소 세 가지 이상의 후보 값들을 선정(최소 27가지 이상의 조합)하여 각 조합에 대한 변수 선택 결과에 대해 본인만의 생각을 더해 해석해보시오.

⇨ 변경 가능한 하이퍼 파라미터 중 Population size, Cross-over rate, 그리고 Mutation rate를 선택했다. 각 하이퍼 파라미터가 취할 수 있는 값은 다음과 같이 정해두었다.

Population size = {20, 40, 60}
Cross-over rate = {0.3, 0.5, 0.7}
Mutation rate = {0.1, 0.3, 0.5}

3*3*3의 27가지 조합을 시험해본 결과는 다음과 같다. 어느 정도의 가독성을 위해 population size에 따라 다른 색으로 표시하였다.

Population Size: 20, Crossover Rate: 0.3, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.3, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.3, Mutation Rate: 0.5
Attributes: ['anaemia', 'diabetes', 'high_blood_pressure', 'serum_creatinine', 'serum_sodium', 'time']

Population Size: 20, Crossover Rate: 0.5, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.5, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.5, Mutation Rate: 0.5
Attributes: ['anaemia', 'diabetes', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.7, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.7, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'smoking', 'time']

Population Size: 20, Crossover Rate: 0.7, Mutation Rate: 0.5
Attributes: ['anaemia', 'diabetes', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.3, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.3, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'platelets', 'serum_sodium', 'sex', 'smoking']

Population Size: 40, Crossover Rate: 0.3, Mutation Rate: 0.5
Attributes: ['age', 'ejection_fraction', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.5, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.5, Mutation Rate: 0.3
Attributes: ['diabetes', 'ejection_fraction', 'high_blood_pressure', 'serum_creatinine', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.5, Mutation Rate: 0.5
Attributes: ['ejection_fraction', 'high_blood_pressure', 'time']

Population Size: 40, Crossover Rate: 0.7, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 40, Crossover Rate: 0.7, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'platelets', 'serum_sodium', 'sex', 'smoking']

Population Size: 40, Crossover Rate: 0.7, Mutation Rate: 0.5
Attributes: ['age', 'ejection_fraction', 'sex', 'smoking', 'time']

Population Size: 60, Crossover Rate: 0.3, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking']

Population Size: 60, Crossover Rate: 0.3, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'high_blood_pressure', 'platelets', 'serum_sodium', 'sex', 'smoking']

Population Size: 60, Crossover Rate: 0.3, Mutation Rate: 0.5
Attributes: ['age', 'anaemia', 'diabetes', 'ejection_fraction', 'smoking', 'time']

Population Size: 60, Crossover Rate: 0.5, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'ejection_fraction', 'high_blood_pressure', 'serum_sodium', 'sex', 'smoking']

Population Size: 60, Crossover Rate: 0.5, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'high_blood_pressure', 'platelets', 'serum_sodium', 'sex', 'smoking']

Population Size: 60, Crossover Rate: 0.5, Mutation Rate: 0.5
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'high_blood_pressure', 'platelets', 'serum_sodium', 'sex', 'smoking']

Population Size: 60, Crossover Rate: 0.7, Mutation Rate: 0.1
Attributes: ['age', 'anaemia', 'diabetes', 'ejection_fraction', 'serum_sodium', 'sex', 'smoking', 'time']

Population Size: 60, Crossover Rate: 0.7, Mutation Rate: 0.3
Attributes: ['age', 'anaemia', 'creatinine_phosphokinase', 'high_blood_pressure', 'platelets', 'serum_sodium', 'sex', 'smoking']

/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_sag.py:350: ConvergenceWarning: The max_iter was reached which means the coef_ did not converge

  warnings.warn(

Population Size: 60, Crossover Rate: 0.7, Mutation Rate: 0.5

Attributes: ['age', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'smoking', 'time']

우선 주목할 점은 Population Size: 60, Crossover Rate: 0.7, Mutation Rate: 0.5에서 convergence가 일어나지 않았다는 것이다. 이는 원래 mutation rate가 크면 converge하는데 오랜 시간이 걸린다는 사실과 일치한다. 하지만 같은 mutation rate을 가지는 다른 조합들에서는 비슷한 일이 일어나지 않은 것을 보아, 큰 population size와 crossover rate 역시 convergence에 걸리는 시간을 늘리는데 영향을 준다는 것을 알 수 있다.

이번에는 population size가 변수 선택에 미치는 영향을 확인하기 위해 다른 조건을 통일하고 population size만 다르게 한 예시를 몇 개 살펴보겠다.

(Crossover Rate, Mutation Rate) = (0.3, 0.1) 일 때,
(Population Size, Number of Attributes) = (20, 10), (40, 8), (60, 8)

(Crossover Rate, Mutation Rate) = (0.3, 0.3) 일 때,
(Population Size, Number of Attributes) = (20, 9), (40, 8), (60, 8)

(Crossover Rate, Mutation Rate) = (0.3, 0.5) 일 때,
(Population Size, Number of Attributes) = (20, 6), (40, 5), (60, 5)

(Crossover Rate, Mutation Rate) = (0.5, 0.1) 일 때,
(Population Size, Number of Attributes) = (20, 10), (40, 8), (60, 8)

(Crossover Rate, Mutation Rate) = (0.5, 0.3) 일 때,
(Population Size, Number of Attributes) = (20, 9), (40, 7), (60, 8)

(Crossover Rate, Mutation Rate) = (0.5, 0.5) 일 때,
(Population Size, Number of Attributes) = (20, 7), (40, 3), (60, 8)

(Crossover Rate, Mutation Rate) = (0.7, 0.1) 일 때,

(Population Size, Number of Attributes) = (20, 10), (40, 8), (60, 8)

(Crossover Rate, Mutation Rate) = (0.7, 0.3) 일 때,
(Population Size, Number of Attributes) = (20, 9), (40, 8), (60, 8)

(Crossover Rate, Mutation Rate) = (0.7, 0.5) 일 때,
(Population Size, Number of Attributes) = (20, 7), (40, 5), (60, 6)

다른 하이퍼파라미터 값이 고정되어 있을 때, Population size가 20, 40, 60으로 증가함에 따라 변수의 개수가 예외는 있지만 대체적으로 감소하는 것을 볼 수 있다. 따라서 Population size는 클수록 선택되는 변수의 수를 감소시키는 효과가 있다고 할 수 있다.

마지막으로 Crossover rate의 영향을 확인하기 위해 나머지 하이퍼파라미터 값을 고정해 보았다.

(Population Size, Mutation Rate) = (20, 0.1) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 10), (0.5, 10), (0.7, 10)

(Population Size, Mutation Rate) = (20, 0.3) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 9), (0.5, 9), (0.7, 9)

(Population Size, Mutation Rate) = (20, 0.5) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 6), (0.5, 7), (0.7, 7)

(Population Size, Mutation Rate) = (40, 0.1) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 8), (0.5, 8), (0.7, 8)

(Population Size, Mutation Rate) = (40, 0.3) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 8), (0.5, 7), (0.7, 8)

(Population Size, Mutation Rate) = (40, 0.5) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 5), (0.5, 3), (0.7, 5)

(Population Size, Mutation Rate) = (60, 0.1) 일 때,

(Crossover Rate, Number of Attributes) = (0.3, 8), (0.5, 8), (0.7, 8)

(Population Size, Mutation Rate) = (60, 0.3) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 8), (0.5, 8), (0.7, 8)

(Population Size, Mutation Rate) = (60, 0.5) 일 때,
(Crossover Rate, Number of Attributes) = (0.3, 5), (0.5, 8), (0.7, 6)

다른 하이퍼파라미터 값이 고정되어 있을 때, Crossover Rate이 0.3, 0.5, 0.7로 증가함에 따라 어느정도 예외는 있지만 대체적으로 변수의 수가 변하지 않는다는 것을 볼 수 있다. 따라서 Crossover Rate은 선택되는 변수에 수에 큰 영향을 미치지 않는다는 결론을 내릴 수 있다.