다변량데이터분석

과제 2

2021150456 이예지

목차

[다변량데이터분석][과제 2]

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

Logistic Regression 모델 적용을 위해 선정한 데이터셋은 Kaggle 의 Gender Recognition by Voice 이다. 해당 데이터셋은 총 21 개의 변수와 3168 개의 관측치를 가지고 있다.
몇 달 전, 유튜브에서 한 남성이 여성인 것처럼 목소리를 내어 다른 사람들을 속이는 영상을 본 적이 있다. 영상을 보고 나서, 넷카마(인터넷 상에서 여성인 척하는 남성)가 여성인 척 목소리를 내어 금전적인 이득을 취한다면 해당 행위는 단순한 장난에서 그치는 것이 아니라 사기라는 범죄 행위로 이어질 수 있을 것이라는 생각이 들었다. 따라서 음성을 통한 성별 예측 문제 역시 중요한 문제라 생각해 위 데이터셋을 선정하였다.
이에 더해, 예측하고자 하는 종속변수는 'label'로, 남성 혹은 여성의 값만을 가지는 이진형 변수이다. 이는 이진형의 형태를 갖는 종속변수에 대해 회귀식의 형태로 모형을 추정하고자 하는 Logistic Regression 의 목적에 부합하므로 해당 데이터셋을 선정하였다.

데이터셋 다운로드 링크: https://www.kaggle.com/datasets/primaryobjects/voicegender

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

데이터셋의 종속변수는 'label', 설명변수는 그 외 나머지 변수들이다. 변수들에 대한 설명은 다음과 같다. 'peakf'의 경우 kaggle 에는 설명이 나와있으나 csv 파일 내에는 존재하지 않으므로 설명 대상에서 제외하였다.

변수명	변수 설명
meanfreq (설명변수)	음성의 평균 주파수(kHz 단위)
sd (설명변수)	주파수의 표준 편차
median (설명변수)	주파수의 중위값(kHz 단위)
Q25 (설명변수)	주파수의 제 1 사분위수(kHz 단위)
Q75	주파수의 제 3 사분위수(kHz 단위)

[다변량데이터분석][과제 2]

(설명변수)	
IQR (설명변수)	주파수의 제 1 사분위수부터 제 3 사분위수까지의 범위(kHz 단위)
skew (설명변수)	주파수 분포의 왜도
kurt (설명변수)	주파수 분포의 첨도
sp.ent (설명변수)	스펙트럼 엔트로피. 스펙트럼의 무질서도를 나타내는 측정치로, 스펙트럼의 피크가 얼마나 뾰족한지를 나타냄.
sfm (설명변수)	스펙트럼 평활도. 오디오 스펙트럼을 특성화하기 위해 디지털 신호처리에서 사용되는 측정치임. 스펙트럼의 모든 주파수 대역에 거의 동일한 양의 전력이 존재하면 높은 스펙트럼 평활도를 가지며, 낮은 스펙트럼 평활도는 스펙트럼의 전력이 상대적으로 적은 수의 대역에 집중되어 있음을 나타냄.
mode (설명변수)	최빈 주파수
centroid (설명변수)	도심 주파수
meanfun (설명변수)	음향 신호 전반에 걸쳐 측정된 평균 기본 주파수. 주기적 진동의 경우 그 주기에 상당하는 주파수를 말하며, 주기적 진동이 없는 경우에는 그 성분 중 최소 주파수를 의미함.
minfun (설명변수)	음향 신호 전반에 걸쳐 측정된 최소 기본 주파수
maxfun (설명변수)	음향 신호 전반에 걸쳐 측정된 최대 기본 주파수
meandom (설명변수)	음향 신호 전반에 걸쳐 측정된 평균 지배 주파수. 지배 주파수는 주기적인 파형의 주파수를 분석했을 때, 진폭 스펙트럼이 극대로 되는 주파수임. 혹은 지진파와 같은 불규칙파에 포함된 진동수 중 빈도나 진폭이 다른 진동수에 비해 탁월한 진동수.
mindom (설명변수)	음향 신호 전반에 걸쳐 측정된 최소 지배 주파수
maxdom (설명변수)	음향 신호 전반에 걸쳐 측정된 최대 지배 주파수
dfrange	음향 신호 전반에 걸쳐 측정된 지배 주파수의 범위

(설명변수)	
modindx (설명변수)	변조 지수. 데이터 신호 같은 변조 신호에 의해 주파수, 위폭, 위상 등의 변조 파라미터가 변이하는 크기를 말함.
label (종속변수)	성별. 남성 혹은 여성의 값을 가짐.

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

일반적인 성인 남성의 음성은 기본 주파수가 90~155Hz 이며, 일반적인 성인 여성의 음성은 165~255Hz 라고 알려져 있다. 위 설명변수들은 대부분 주파수와 관련된 변수들이므로 종속변수와 관련이 있어 보인다. 그러나 그 중에서도 'meanfreq', 'median', 'centroid', 'meanfun', 'meandom' 변수가 높은 상관관계를 보일 것으로 예상된다. 일반적으로 성별에 따라 음성의 주파수에 차이가 있으므로, 주파수의 평균 역시 차이가 있을 것이라 생각한다. 추가적으로, 순간적으로 비명을 지르는 등의 행위가 포함되었을 경우 평균 주파수가 올라가게 되나, 중위값은 이러한 이상치에 영향을 덜 받기 때문에 'meanfreq'보다도 'median'이 높은 상관관계를 보일 것으로 예측된다. 'centroid' 역시 비슷한 맥락으로, 성별에 따라 주파수가 다르기 때문에, 도심 주파수 역시 성별에 따라 다를 것으로 예상되어 높은 상관관계가 있을 것이라 생각하였다.
설명변수들 내에서 높은 상관관계가 있을 것으로 예상되는 조합은 {meanfreq, median, skew, mode}, {Q25, IQR}, {Q75, IQR}, {kurt, sfm}, {sp.ent, sfm}이다.

- {meanfreq, median, skew, mode}: 왜도는 분포가 대칭을 벗어나 한쪽으로 치우친 정도이며, 평균, 중위값, 최빈값을 통해 분포가 치우친 방향을 알 수 있다. 대칭분포에서는 평균 = 중위값 = 최빈값이나, 이들 사이에 크기 차이가 생기면 한쪽으로 치우친 분포가 된다. 따라서 해당 변수들 사이에 높은 상관관계가 있을 것으로 보인다.
- {Q25, IQR}, {Q75, IQR}: IQR 은 제 3 사분위수에서 제 1 사분위수를 빼는 것으로 구할 수 있다. 이는 IQR 의 값이 Q25 와 Q75 에 의존한다는 이야기이므로 이들 사이에는 높은 상관관계가 있을 것으로 예상된다.
- {kurt, sfm}: 주파수 분포의 첨도가 크다는 이야기는 특정 주파수가 많이 측정되었다는 이야기이며, 이는 스펙트럼 평활도가 낮은 값을 보일 것이라는 것과 동일한 이야기이다. 따라서 둘 사이에는 음의 상관관계가 강하게 나타날 것으로 예상된다.
- {sp.ent, sfm}: 다양한 스펙트럼이 존재하면 스펙트럼 엔트로피가 높아지며, 스펙트럼 평활도 역시 높아진다. 따라서 스펙트럼 엔트로피와 스펙트럼 평활도 사이에는 양의 상관관계가 있을 것으로 보인다.

2) 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

'minfun', 'maxfun', 'mindom', 'maxdom', 'dfrange' 변수들은 필요하지 않을 것으로 예상된다. 음성의 주파수는 음색뿐만 아니라 말하는 속도, 감정 등 다양한 요소와 관련이 있는데, 최솟값과 최댓값은 이러한 다른 요소의 영향을 크게 받은 이상치가 속해 있을 가능성이 높으므로 종속변수를 예측하는데 적절하지 않다고 판단했다. 또한, 'dfrange'의 경우, 음향 신호 전반에 걸쳐 측정된 지배 주파수의 범위로, IQR 과 달리 모든 지배 주파수를 포함하는 범위를 제시하고 있으므로 이 역시 이상치에 민감하여 설명변수로 사용하기 적절하지 않다고 판단했다. 추가적으로, 'Q25'와 'Q75' 역시 필요하지 않을 것으로 예상된다. 'IQR' 변수가 이 두 변수 이용해서 얻어진 새로운 변수라고 볼 수 있으므로, 세 변수를 모두 사용하는 것은 모델의 복잡도만을 증가시킬 뿐, 성능 향상에 큰 도움이 된다고 보기 어렵기 때문이다.

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

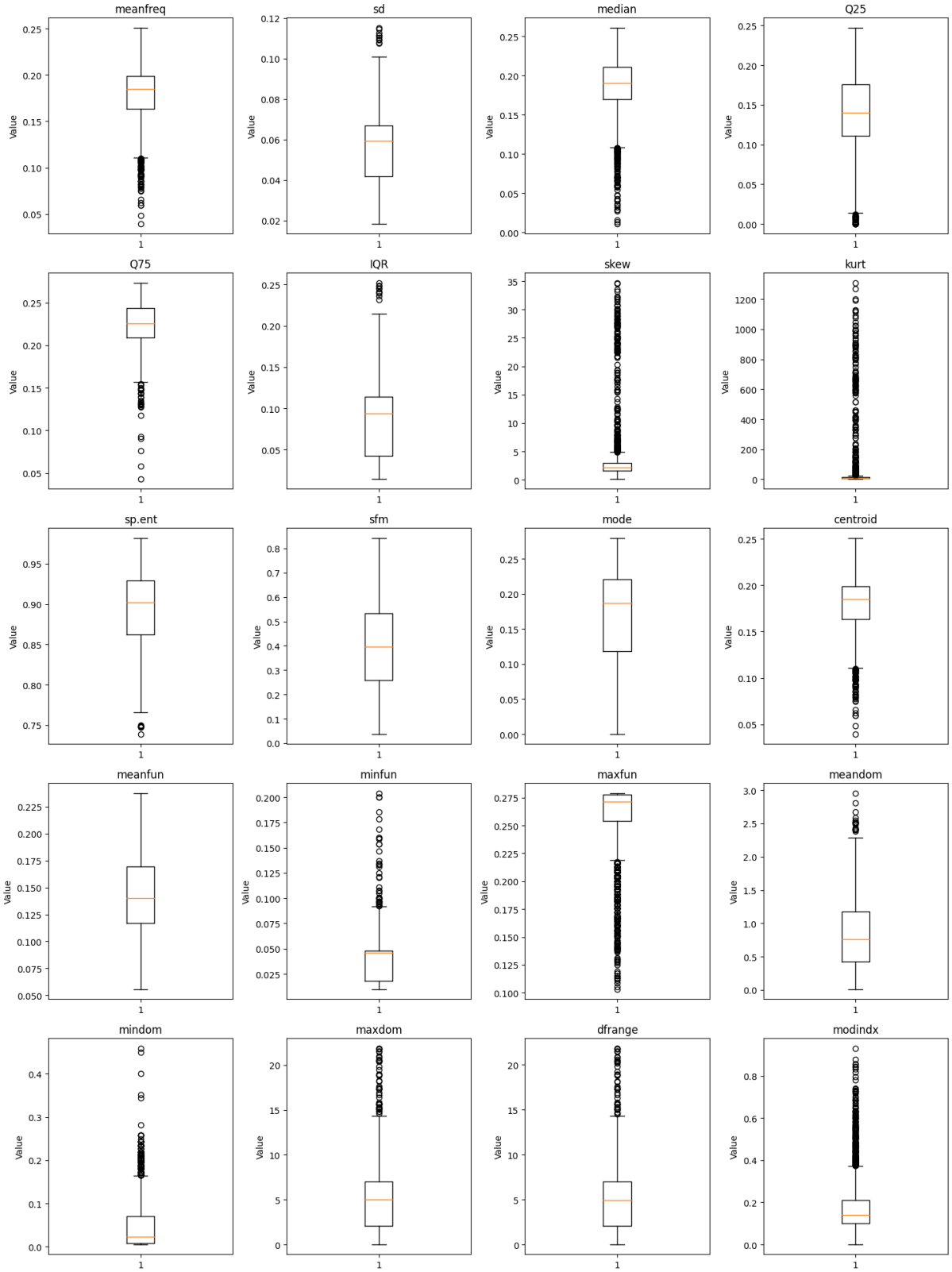
설명변수들 중 범주형 변수는 존재하지 않으므로 1-of-C coding 은 적용하지 않는다.
개별 입력 변수들에 대한 단변량 통계량을 계산하면 다음과 같다.

[다변량데이터분석][과제 2]

	Mean	Standard deviation	Skewness	Kurtosis
meanfreq	0.180907	0.029913	-0.617203	0.801997
sd	0.057126	0.016650	0.136851	-0.522859
median	0.185621	0.036354	-1.012305	1.625037
Q25	0.140456	0.048672	-0.490644	0.016411
Q75	0.224765	0.023636	-0.899884	2.975213
IQR	0.084309	0.042776	0.295292	-0.449347
skew	3.140168	4.239859	4.930978	25.321540
kurt	36.568461	134.907364	5.869805	35.873550
sp.ent	0.895127	0.044972	-0.430730	-0.425149
sfm	0.408216	0.177493	0.339797	-0.836508
mode	0.165282	0.077191	-0.836840	-0.257397
centroid	0.180907	0.029913	-0.617203	0.801997
meanfun	0.142807	0.032299	0.039122	-0.860496
minfun	0.036802	0.019217	1.877115	10.739221
maxfun	0.258842	0.030073	-2.237475	5.193815
meandom	0.829211	0.525122	0.610733	-0.056579
mindom	0.052647	0.063289	1.660327	2.182242
maxdom	5.047277	3.520601	0.725845	1.310770
dfrange	4.994630	3.519484	0.727916	1.314040
modindx	0.173752	0.119436	2.063357	5.913695

개별 입력 변수들의 box plot 시각화 결과는 다음과 같다.

[다변량데이터분석][과제 2]



정규분포를 가정하기 위해서 왜도와 첨도(본 분석에서는 첨도를 구하기 위해 Fisher 의 정의를 사용한다.)가 모두 0 이어야 하지만, 보편적으로는 왜도의 절댓값이 3 보다 작고, 첨도의 절댓값이 8 보다 작으면 정규분포를 따른다고 해석할 수 있다. 따라서 정규분포를 따른다고 할 수 있는

[다변량데이터분석][과제 2]

변수들은 'meanfreq', 'sd', 'median', 'Q25', 'Q75', 'IQR', 'sp.ent', 'sfm', 'mode', 'centroid', 'meanfun', 'maxfun', 'meandom', 'mindom', 'maxdom', 'dfrange', 'modindx'로, 총 17 개이다.

Box plot 을 확인해보았을 때, 데이터들이 정규분포를 따르면 데이터들이 골고루 분포되어 있어, 이상치가 없을 것이라 예상할 수 있다. 따라서 이상치가 없는 Box plot 을 가졌을 경우 정규분포를 따른다고 가정하면, 정규분포를 따른다고 할 수 있는 변수는 'sfm', 'mode', 'meanfun'으로, 총 3 개이다. 이 변수들은 모두 왜도와 첨도를 기준으로 각 변수가 정규분포를 따르는지 확인해보았을 때 포함되었던 변수들이다. 이는 Box plot 과 왜도, 첨도 모두 데이터의 분포를 통해 나오는 결과이며, Box plot 에 조금 더 엄격한 기준을 적용했기 때문으로 보인다.

추가적으로 보다 정확한 정규성 검정을 위해 Shapiro-Wilk Test 를 진행하였다. Shapiro-Wilk Test 는 표본 크기가 5000 을 넘어가면 정확하지 않을 수 있으나, 현재 데이터셋의 크기는 3168 로 5000 을 넘어가지 않기 때문에 Shapiro-Wilk Test 의 결과를 신뢰할 수 있을 것이라 판단했기 때문이다.
Shapiro-Wilk Test 의 가설은 다음과 같다.

H_0: 데이터가 정규분포를 따른다. vs. H_a: 데이터가 정규분포를 따르지 않는다.

	Column	Test Statistic	Critical Value(alpha = 0.05)	Test Result
0	meanfreq	0.976071	1.152381e-22	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
1	sd	0.969054	1.344717e-25	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
2	median	0.946825	2.152655e-32	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
3	Q25	0.977815	7.735911e-22	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
4	Q75	0.954351	2.158879e-30	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
5	IQR	0.942503	1.914607e-33	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
6	skew	0.384877	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
7	kurt	0.231031	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
8	sp.ent	0.976718	2.306296e-22	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
9	sfm	0.966455	1.479177e-26	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
10	mode	0.905056	1.314824e-40	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
11	centroid	0.976071	1.152381e-22	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
12	meanfun	0.981747	8.794537e-20	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
13	minfun	0.785093	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
14	maxfun	0.691370	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
15	meandom	0.965022	4.624858e-27	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
16	mindom	0.702839	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
17	maxdom	0.942088	1.529815e-33	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
18	dfrange	0.941509	1.119995e-33	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
19	modindx	0.819478	0.000000e+00	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)

[다변량데이터분석][과제 2]

모든 입력 변수들에 대하여 Shapiro-Wilk Test 를 진행한 결과, 유의수준 0.05 에서 모든 변수가 정규분포를 따르지 않는다는 결과를 확인할 수 있었다.

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

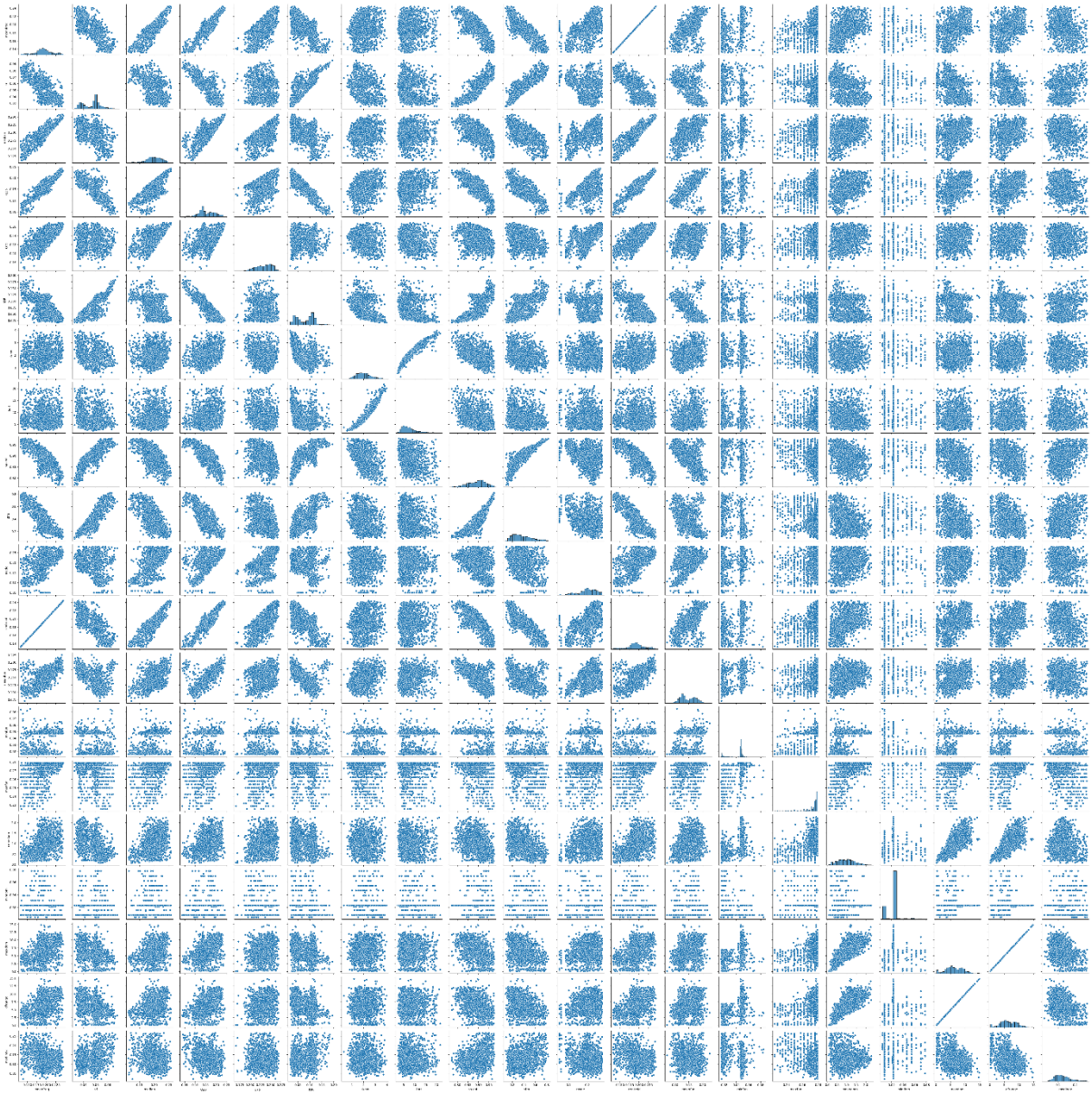
[Q3]의 Box plot 에서 나타난 것처럼, $Q3 + 1.5 * IQR$보다 위쪽에 있거나, $Q1 - 1.5 * IQR$보다 아래쪽에 있는 값을 이상치로 정의하였다. 여기서 $Q3$는 제 3 사분위수, $Q1$은 제 1 사분위수, $IQR = Q3 - Q1$이다. 정의한 이상치 범위에 해당하는 객체들을 제거한 결과, 데이터셋의 크기가 3168 에서 1607 로 감소하였다. 약 절반 정도의 데이터만이 남았지만, 보다 완화된 기준을 통해 이상치를 분석 데이터에 포함시킨다면, 성별이 아닌, 다른 상황적인 요인 등으로 인해 생겨난 노이즈가 분석 결과에 영향을 미칠 수 있을 것으로 생각된다. 또한, 로지스틱 회귀분석을 수행하기에 충분한 양의 데이터라고 판단되므로 본 분석에서는 해당 이상치 기준을 그대로 사용하기로 한다.

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 판단해 보시오.

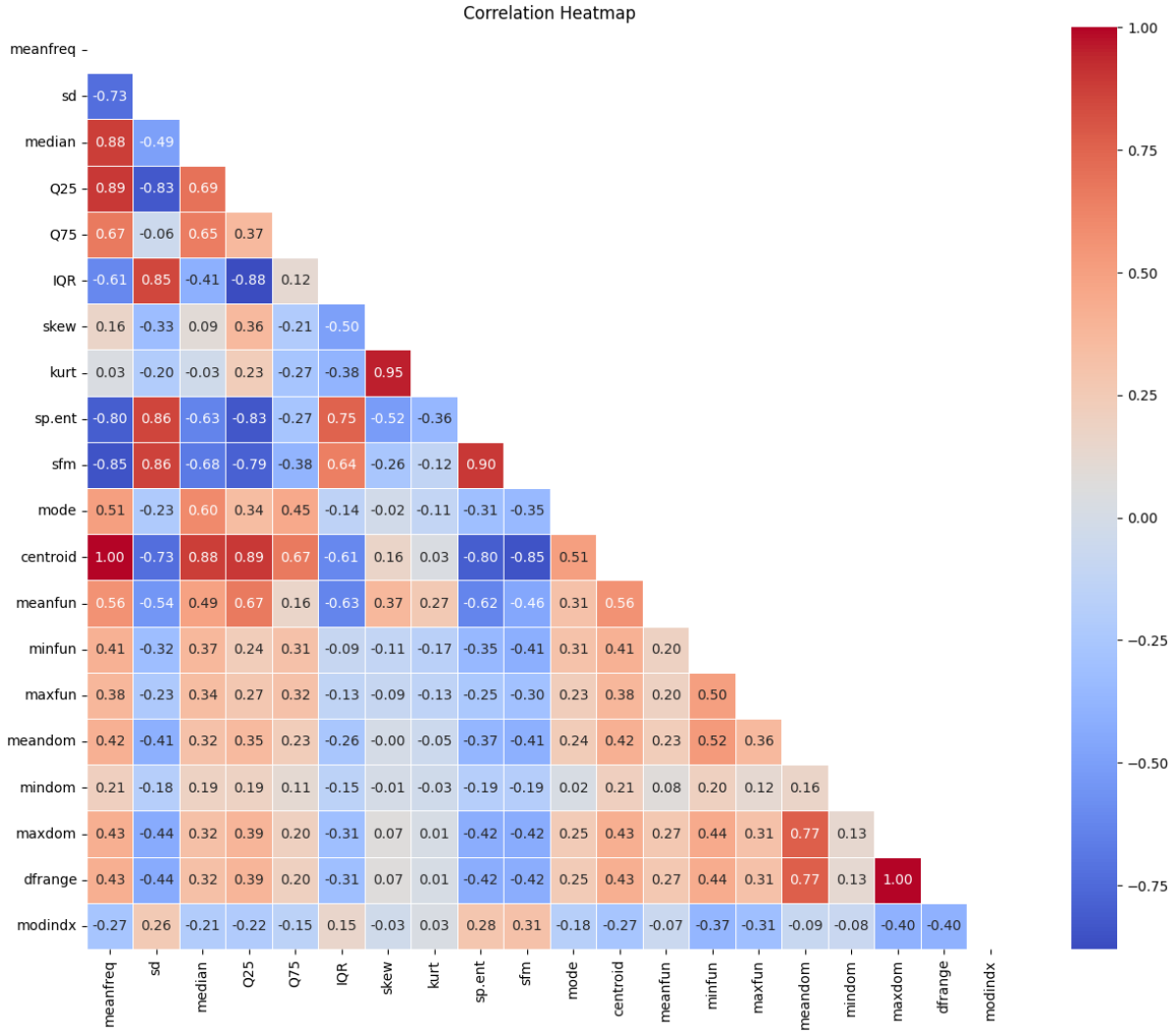
가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도를 그리고자 seaborn 라이브러리의 pairplot 함수를 이용하였다. 그 결과는 다음과 같다.

[다변량데이터분석][과제 2]



추가적으로, 변수들 간의 상관계수를 시각화한 히트맵은 다음과 같다.

[다변량데이터분석][과제 2]



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

산점도를 통해 육안으로 확인해본 결과, 많은 데이터쌍들이 상관관계를 띠고 있는 것을 알 수 있었다. 조금 더 구체적으로 확인하면, {median, meanfreq}, {Q25, meanfreq}, {centroid, meanfreq}, {IQR, sd}, {sp.ent, sd}, {sfm, sd}, {centroid, median}, {centroid, Q25}, {kurt, skew}, {sfm, sp.ent}, {dfrange, maxdom}이 강한 양의 상관관계를, {sp.ent, meanfreq}, {sfm, meanfreq}, {Q25, sd}, {sp.ent, Q25}, {sfm, Q25}, {IQR, Q25}, {centroid, sp.ent}, {centroid, sfm}이 강한 음의 상관관계를 띠고 있다. 다만 이는 해석하는 사람에 따라 달라질 수 있으므로, 상관계수를 통해 강한 상관관계를 가지는 변수쌍을 보다 명확히 규정하고자 한다.

상관계수는 변수 간의 선형관계를 측정하는 단위이며, [-1, 1]의 값을 가질 수 있다. 값이 1 에 가까울수록 양의 상관관계, -1 에 가까울수록 음의 상관관계를 가진다. 값의 절댓값이 클수록 선형성은 강하며, 값이 0 에 가까우면 선형성이 존재하지 않는다고 판단할 수 있다. 상관계수의

절댓값이 [0.0, 0.1)의 값을 가지면 상관관계가 거의 없다, [0.1, 0.5)의 값을 가지면 약한 상관관계가 있다, [0.5, 0.7)의 값을 가지면 상관관계가 있다, [0.7, 0.9)의 값을 가지면 높은 상관관계가 있다, [0.9, 1.0]의 값을 가지면 매우 높은 상관관계가 있다고 해석할 수 있다.
높은 상관관계와 매우 높은 상관관계를 가지는 변수 조합은 아래와 같다.

상관계수의 절댓값	의미	해당하는 변수쌍
[0.7, 0.9)	높은 음/양의 상관관계가 있다.	- 양: {median, meanfreq}, {Q25, meanfreq}, {IQR, sd}, {sp.ent, sd}, {sfm, sd}, {centroid, median}, {centroid, Q25}, {sp.ent, IQR}, {maxdom, meandom}, {dfrange, meandom} - 음: {sd, meanfreq}, {sp.ent, meanfreq}, {sfm, meanfreq}, {Q25, sd}, {centroid, sd}, {IQR, Q25}, {sp.ent, Q25}, {sfm, Q25}, {centroid, sp.ent}, {centroid, sfm}
[0.9, 1.0]	매우 높은 음/양의 상관관계가 있다.	- 양: {centroid, meanfreq}, {kurt, skew}, {sfm, sp.ent}, {dfrange, maxdom} - 음: 존재하지 않음.

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

본 분석에서는 상관계수의 절댓값이 0.7 이상일 때 강한 상관관계가 존재한다는 일반적인 기준을 따르도록 한다.
강한 상관관계를 지녔다는 것은 하나의 변수로 다른 변수를 대부분 혹은 전부 설명할 수 있다는 것을 의미한다. 이때 설명할 수 있는 정도는 상관계수의 절댓값에 따른다. 따라서 두 변수 중 하나의 변수를 삭제하여도 종속변수를 예측하는 모델을 구축하는데 있어 모델의 설명력이 크게 줄어들지 않을 것이라 예상할 수 있다. 또한, 변수의 개수를 줄임으로써 모델 구축을 위한 데이터를 수집하는 비용 감소와 모델 해석의 용이성 증대라는 효용을 얻을 수 있다.
두 변수 중 모델을 구축할 때 사용할 변수를 선정하는 기준은 다음과 같다. [Q3]에서 제시된 Box plot 의 형태를 보았을 때, 데이터가 더 넓게 분포해 있다면 각 객체들의 특성을 잘 나타내는 변수라는 의미이므로 데이터가 더 넓게 분포된 변수를 선택할 것이다. 그러나, 이상치는 데이터의 특성을 잘 나타내는 값으로 보기 어려우므로, 이상치가 포함되지 않는 범위 내에서 데이터가 더 넓게 분포된 변수를 선택한다.

[다변량데이터분석][과제 2]

- {median, meanfreq}: Box plot 을 통해 'meanfreq'보다 'median'의 데이터의 범위가 조금 더 긴 것을 확인할 수 있다. 따라서 'meanfreq'보다는 'median'을 사용하는 편이 더 좋을 것이라 생각된다.
- {Q25, meanfreq}: 마찬가지로 Box plot 을 통해 'meanfreq'보다 'Q25 의 데이터의 범위가 더 긴 것을 확인할 수 있다. 따라서 'meanfreq'보다는 'Q25 를 사용하는 편이 더 좋을 것이라 생각된다.
- {IQR, sd}: 'sd'보다 'IQR'의 데이터 범위가 더 긴 것을 확인할 수 있다. 따라서 'sd'보다 'IQR'을 사용하는 편이 더 좋을 것이다.
- {sp.ent, sd}: 'sd'보다 'sp.ent'의 데이터 범위가 더 긴 것을 확인할 수 있다. 따라서 'sd'보다 'sp.ent'를 사용하는 편이 더 좋을 것이다.
- {sfm, sd}: 두 변수의 데이터 범위가 거의 동일한 것을 확인할 수 있다. 따라서 이상치가 존재하지 않아 데이터가 더 잘 분포되어 있다고 볼 수 있는 'sfm'을 선택하는 편이 더 좋을 것이라 생각된다.
- {centroid, median}: 'median'이 'centroid'보다 조금 더 긴 범위를 가졌으므로, 'centroid'보다 'median'을 사용하는 편이 더 좋을 것이다.
- {centroid, Q25}: 'Q25'가 'centroid'보다 더 긴 범위를 가졌으므로 'Q25'를 사용하는 편이 더 좋을 것이다.
- {sp.ent, IQR}: 'sp.ent'가 'IQR'보다 조금 더 긴 범위를 가졌으므로 'IQR'보다 'sp.ent'를 사용하는 편이 더 좋을 것이다.
- {maxdom, meandom}: 'maxdom'이 'meandom'보다 긴 범위를 가졌으므로 'maxdom'을 사용하는 편이 더 좋을 것이다.
- {dfrange, meandom}: 마찬가지로 'dfrange'가 'meandom'보다 긴 범위를 가졌으므로 'dfrange'를 사용하는 편이 더 좋을 것이다.
- {sd, meanfreq}: 'meanfreq'가 'sd'보다 더 넓은 범위를 가졌으므로 'meanfreq'를 사용하는 편이 더 좋을 것이다.
- {sp.ent, meanfreq}: 'sp.ent'가 'meanfreq'보다 더 넓은 범위를 가졌으므로 'sp.ent'를 사용하는 편이 더 좋을 것이다.
- {sfm, meanfreq}: 'meanfreq' 가 'sfm'보다 더 넓은 범위를 가졌으므로 'meanfreq'를 사용하는 편이 더 좋을 것이다.
- {Q25, sd}: 'Q25'가 'sd'보다 더 넓은 범위를 가졌으므로 'Q25'를 사용하는 편이 더 좋을 것이다.
- {centroid, sd}: 'centroid'가 'sd'보다 더 넓은 범위를 가졌으므로 'centroid'를 사용하는 편이 더 좋을 것이다.

- {IQR, Q25}: 'Q25'가 'IQR'보다 조금 더 넓은 범위를 가졌으므로 'Q25'를 사용하는 편이 더 좋을 것이다.
- {sp.ent, Q25}: 'Q25'가 'sp.ent'보다 조금 더 넓은 범위를 가졌으므로 'Q25'를 사용하는 편이 더 좋을 것이다.
- {sfm, Q25}: 'Q25'가 'sfm'보다 더 넓은 범위를 가졌으므로 'Q25'를 사용하는 편이 더 좋을 것이다.
- {centroid, sp.ent}: 'sp.ent'가 'centroid'보다 넓은 범위를 가졌으므로 'sp.ent'를 사용하는 편이 더 좋을 것이다.
- {centroid, sfm}: 'centroid'가 'sfm'보다 넓은 범위를 가졌으므로 'centroid'를 사용하는 편이 더 좋을 것이다.
- {centroid, meanfreq}: 두 변수의 데이터 분포가 동일하므로, 다른 변수들과의 상관관계에 대한 해석 역시 동일하다. 다만, 'centroid'가 'meanfreq'보다 더 고차원적인 개념이므로, 'centroid'가 'meanfreq'를 포함한다고 해석할 수 있다. 따라서 'meanfreq'를 삭제하는 편이 더 좋을 것이라 생각한다.
- {kurt, skew}: 현 Box plot 으로는 정확히 어느 변수의 범위가 더 긴지 판별하기 어렵고, 어느 변수에 이상치가 많은지도 이야기하기 어렵다. 따라서 해당 변수들의 왜도와 첨도를 통해 생각해보면, 'skew'의 왜도와 첨도 모두 'kurt'보다 작아 데이터가 보다 고르게 분포되어 있다고 볼 수 있으므로 'skew'를 사용하는 편이 더 좋을 것이다.
- {sfm, sp.ent}: 'sp.ent'가 'sfm'보다 넓은 범위를 가졌으므로 'sp.ent'를 사용하는 편이 더 좋을 것이다.
- {dfrange, maxdom}: 두 변수의 데이터 분포가 거의 동일하므로, 다른 변수들과의 상관관계에 대한 해석 역시 동일하게 할 수 있다. 다만, 'dfrange'는 음향 신호 전반에 걸쳐 측정된 지배 주파수의 범위이고, 'maxdom'은 음향 신호 전반에 걸쳐 측정된 지배 주파수의 최댓값이므로 'dfrange'에 'maxdom'의 영향이 미치고 있다고 할 수 있다. 따라서 'dfrange'가 'maxdom'을 포함하는 개념이라고 볼 수 있으므로 'dfrange'를 선택하는 편이 더 좋을 것이라 생각한다.

위 결과들을 종합하였을 때 삭제된 변수는 'meanfreq', 'IQR', 'sd', 'sp.ent', 'sfm', 'centroid', 'maxdom', 'meandom', 'kurt'이다.

따라서 최종적으로 선택된 변수는 'median', 'Q25', 'dfrange', 'skew', 'Q75', 'mode', 'meanfun', 'minfun', 'maxfun', 'mindom', modindx'이다.

[다변량데이터분석][과제 2]

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

모든 입력 변수를 사용하는 경우, {centroid, meanfreq}, {dfrange, maxdom}의 상관계수가 모두 1 이므로 종속성이 존재하여 Singular matrix 가 발생한다. 따라서 분석 진행을 위하여 'meanfreq', 'maxdom' 변수를 제거한 뒤 모델을 학습하도록 하겠다.

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

먼저, 변수가 통계적으로 유의미한지를 검정하기 위한 가설은 다음과 같다.

$$H_0: \beta_i = 0 \quad \text{vs.} \quad H_a: \beta_i \neq 0$$

$p - value < \alpha = 0.05$이면 귀무가설을 기각할 수 있다.

다음은 각 변수들에 대한 p-value 값을 나타낸 것이다.

	P-value
constant	0.8812
sd	0.9560
median	0.9474
Q25	1.0000
Q75	1.0000
IQR	1.0000
skew	0.0078
kurt	0.3057
sp.ent	0.7174
sfm	0.3727
mode	0.9617
centroid	0.9804
meanfun	0.1192
minfun	0.9437
maxfun	0.9655
meandom	0.6896
mindom	0.9870
dfrange	0.0393
modindx	0.3810

위 내용을 참조하면 다음과 같은 결과를 얻을 수 있다.

[다변량데이터분석][과제 2]

변수명	귀무가설 기각 여부	해석
sd	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
median	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
Q25	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
Q75	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
IQR	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
skew	귀무가설 기각	회귀계수가 유의미하다.
kurt	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
sp.ent	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
sfm	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
mode	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
centroid	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
meanfun	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
minfun	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
maxfun	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
meandom	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
mindom	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
dfrange	귀무가설 기각	회귀계수가 유의미하다.
modindx	귀무가설 기각 불가	회귀계수가 유의미하지 않다.

따라서 유의수준 0.05 에서 통계적으로 유의미한 변수는 'skew', 'dfrange' 총 2 개이다.

- 'skew': 주파수의 분포가 얼마나 한쪽으로 치우쳐졌는지 알 수 있는 변수이다. 해당 변수를 통해 비슷한 주파수가 많았는지, 혹은 비교적 다양한 주파수가 존재했는지를 알 수 있다. 비슷한 주파수가 많았다면 일반적인 기준에 따라 성별을 예측하는데 어려움이 적을 것이며, 분포가 한쪽으로 치우쳐진 경우 역시 데이터가 많이 분포해있는 곳을 통해 성별을 예측할 수 있을 것이다. 따라서 해당 변수는 상식 선에서도 유효한 변수라고 판단했다.
- 'dfrange': 음향 신호 전반에 걸쳐 측정된 지배 주파수의 범위를 나타내는 변수이다. 지배 주파수를 통해 음성에서 해당 주파수의 성분이 중요하다는 것을 알 수 있으며, 이는 음성의 특성과 관련이 있다. 따라서 지배 주파수를 아는 것은 성별을 예측하는데 도움이 될 것이라 생각한다. [Q2-2]에서 언급했다시피, 범위는 이상치에 민감하다. 음성의 주파수는 음색, 어조, 감정 상태 등 다양한 요인들의 영향을 받는다. 이는 성별 외 다른 요인들이 지배 주파수에 영향을 미칠 수 있다는 것을 의미하며, 이로 인해 최소 혹은 최대 지배 주파수가 극단적인 값을 띠게 되면 성별을 예측하는데 있어 악영향을 미칠 수 있게 된다. 이러한 이유로 앞선 문항에서는 'dfrange'가 유효하지 않은 변수일 것이라

예상했으나, [Q4]에서 이상치를 제거함으로써 우려하던 문제가 사라졌으므로 상식적인 수준에서도 'dfrange'가 유의미한 변수라고 판단했다.

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

'minfun', 'maxfun', 'mindom', 'maxdom', 'dfrange'를 유의하지 않을 것이라 예상했으나, 'dfrange'는 통계적으로 유의미한 변수로 나타났다. 'minfun', 'maxfun', 'mindom'은 모두 p-value 가 1 에 가까운 값으로 나타나, 모델링 측면에서 실제로 유의하지 않는 것으로 나타났다. 'maxdom'은 'dfrange'와의 종속성이 존재하여 제거한 변수인데, 'dfrange'가 통계적으로 유의하므로, 'dfrange' 대신 'maxdom'을 사용했더라면 'maxdom'이 유의한 변수로 나타났을 것이다.
앞서 언급했듯이, 최대, 최소, 범위가 이상치에 민감하다는 점을 고려하면 본래는 'dfrange'와 'maxdom' 모두 유효하지 않은 변수일 것이라고 생각하나 현 데이터셋의 경우 [Q4]를 통해 이상치를 제거하였으므로 검정 결과처럼 'dfrange'와 'maxdom'이 유의미한 변수가 될 수 있다고 판단했다.

3) 학습 데이터와 테스트 데이터에 대한 Confusion Matrix 를 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure 를 산출하여 비교해 보시오

Confusion Matrix (Train)		**Predicted**	
		Female	Male
Actual	Female	502	62
	Male	100	460

학습 데이터의 경우, 여성을 남성으로 잘못 판별하는 경우보다 남성을 여성으로 잘못 판별하는 경우가 더 많은 것을 확인할 수 있다.

Confusion Matrix (Test)		**Predicted**	
		Female	Male
Actual	Female	229	29
	Male	136	189

테스트 데이터 역시 여성을 남성으로 잘못 판별하는 경우보다 남성을 여성으로 잘못 판별하는 경우가 더 많은 것을 확인할 수 있다. 학습된 모델이 데이터셋에 대해 남성보다 여성으로 판별하는 일이 더 많은 것을 알 수 있다. 이는 몇몇 변수가 성별을 예측하는데 있어 유의하지 않아 성별 예측에 혼란을 주고 있는 것으로 볼 수 있다. 이 모델의 경우, 'label'='Male'인 경우가 많은 데이터셋으로 테스트를 진행한다면 현재 측정된 성능보다 낮은 성능을 보일 것이라 예상할 수 있다.

추가적으로, 두 Confusion matrix 를 통해 여성과 남성이 비슷한 비율로 존재하는 것을 확인할 수 있다. 따라서 simple accuracy 를 성능 판별 지표로 사용하여도 큰 문제가 없다는 것을 아래 지표들을 통해 확인할 수 있다.

	ACC	BCR	F1
Train	0.8559	0.8551	0.8502
Test	0.8654	0.8635	0.8533

학습 데이터와 테스트 데이터에 대한 성능 지표들을 살펴보면, 현재 분할된 데이터셋에 대해서는, 미세하게 테스트 데이터에 대한 성능이 더 좋은 것을 확인할 수 있다. 그러나 이는 설정된 seed 에 따라 달라질 수 있는 부분이므로 무조건적으로 테스트 데이터에 대한 성능이 더 좋다고 이야기하기 어렵다. 현재 학습 데이터와 테스트 데이터에 대해 세 지표가 모두 우수한 성능을 보이고 있으며, 세 지표 사이에 값의 차이가 크지 않은 것을 확인할 수 있다. 이는 데이터셋이 불균형하지 않아, 모든 y 의 예측값이 한 성별로 쏠리는 현상이 발생하지 않았기 때문이다. 테스트 데이터에 대해 세가지 지표 모두 우수한 점수가 나왔다는 것은 모델이 훈련 데이터셋에 대해 과적합되지 않았다는 것을 의미한다. 따라서 우리는 해당 모델을 통해 새롭게 주어지는 데이터에 대해서도 잘 예측할 수 있으리라고 생각할 수 있다.

4) 학습 데이터와 테스트 데이터에 대한 AUROC 를 산출하는 함수를 직접 작성하고 이를 사용하여 학습/테스트 데이터셋에 대한 AUROC 를 비교해 보시오

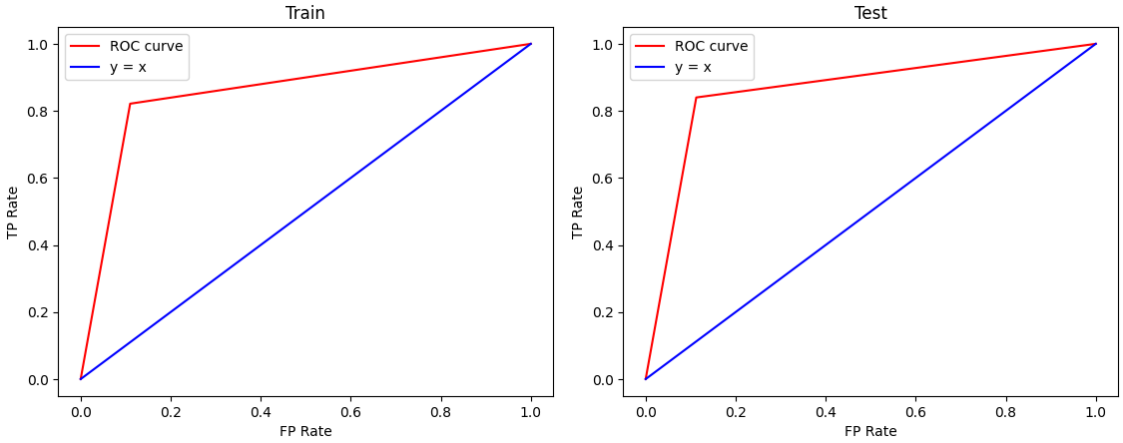
AUROC 를 산출하는 함수는 다음과 같이 작성할 수 있다.

```
def calculate_auroc(y_true, y_pred):

    sorted_idx = np.argsort(y_pred)[::-1]
    sorted_y_pred = y_pred[sorted_idx]
    sorted_y_true = y_true[sorted_idx]

    num_pos = np.sum(y_true == 1)
    num_neg = np.sum(y_true == 0)

    tpr = []
    fpr = []

    tp = 0
    fp = 0

    for i in range(len(sorted_y_true)):
        if sorted_y_true[i] == 1:
            tp += 1
        else:
            fp += 1
        tpr.append(tp / num_pos)
        fpr.append(fp / num_neg)

    auroc = np.trapz(tpr, fpr)
    return round(auroc, 4)
```

이 함수를 통해 계산된 AUROC 값은 다음과 같다.

Data	AUROC
Train	0.8584
Test	0.8577

추가적으로 ROC curve 를 시각화한 결과는 다음과 같다.



두 데이터셋에 대하여 계산된 AUROC 값은 이전에 보았던 세 지표처럼 매우 비슷한 값이 도출되었으며, ROC curve 역시 매우 유사한 형태를 띠고 있는 것을 확인할 수 있다. 현재 계산된 AUROC 값만으로 모델을 비교한다면, 학습 데이터 쪽이 더 좋다고 이야기할 수 있으나, 이는 seed 의 영향일 수 있으므로 함부로 단정지을 수 없다. 어느 데이터셋에 대해 성능이 더 좋은지 이야기하기 위해서는 여러 번의 실험을 통해 신뢰구간을 구하는 과정이 필요하다. 현재로서는 학습시킨 모델이 테스트 데이터셋에 대해서도 좋은 성능을 보이고 있어 모델이 과적합되지 않았다고 판단할 수 있다.

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

[Q5]에서 선택된 변수는 'median', 'Q25', 'dfrange', 'skew', 'Q75', 'mode', 'meanfun', 'minfun', 'maxfun', 'mindom', modindx'이다.

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

먼저, 변수가 통계적으로 유의미한지를 검정하기 위한 가설은 다음과 같다.

$$H_0: \beta_i = 0 \quad \text{vs.} \quad H_a: \beta_i \neq 0$$

$p - value < \alpha = 0.05$이면 귀무가설을 기각할 수 있다.

[다변량데이터분석][과제 2]

다음은 각 변수들에 대한 p-value 값을 나타낸 것이다.

	P-value
constant	0.0294
median	0.8122
Q25	0.0306
Q75	0.6782
skew	0.0000
mode	0.8667
meanfun	0.0352
minfun	0.9352
maxfun	0.8779
mindom	0.9916
dfrange	0.0000
modindx	0.5801

위 내용을 참조하여 아래와 같은 결과를 얻을 수 있다.

변수명	귀무가설 기각 여부	해석
median	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
Q25	귀무가설 기각	회귀계수가 유의미하다.
Q75	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
skew	귀무가설 기각	회귀계수가 유의미하다.
mode	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
meanfun	귀무가설 기각	회귀계수가 유의미하다.
minfun	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
maxfun	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
mindom	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
dfrange	귀무가설 기각	회귀계수가 유의미하다.
modindx	귀무가설 기각 불가	회귀계수가 유의미하지 않다.

[Q6-1]에서 통계적으로 유의미한 변수는 'skew', 'dfrange'가 있었다. 현재는 'skew', 'dfrange'뿐 아니라 'Q25', 'meanfun' 역시 유의미한 변수로 나타나 총 4 개의 변수가 유의한 것으로 나타났다. [Q6]의 데이터는 상관관계가 높은 변수가 여럿 포함되어 다중공선성의 문제가 발생했을 것으로 보인다. 그러나 [Q7]의 경우, 상관관계가 높은 변수들을 제거함으로써 다중공선성 문제가 완화되어 통계적으로 유의미한 변수가 늘어난 것으로 보인다.

일반적으로 성별에 따라 음성의 주파수의 범위가 차이가 있는데, 'Q25'는 측정된 주파수의 제 1 사분위수를 나타내므로 해당 값의 차이가 성별을 예측하는데 유의미하게 작용할 것으로

예상된다. 'meanfun'은 음향 신호 전반에 걸쳐 측정된 평균 기본 주파수로, 평균 역시 이상치에 민감한 통계량이다. 그러나 [Q4]를 통해 이상치를 제거하여 일반적인 수준의 평균 기본 주파수들만이 남아, 성별을 예측하는데 있어 혼란이 줄어든 것으로 보인다. 따라서 이 역시 상식선에서도 유의미한 변수라고 판단할 수 있다.

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

Confusion Matrix (Train)		**Predicted**	
		Female	Male
Actual	Female	482	82
	Male	115	445

학습 데이터의 경우, 여성을 남성으로 잘못 판별하는 경우보다 남성을 여성으로 잘못 판별하는 경우가 더 많은 것을 확인할 수 있다. [Q6]과 동일하게 여성으로 예측한 경우가 남성으로 예측한 경우보다 많은 것을 확인할 수 있다.

Confusion Matrix (Test)		**Predicted**	
		Female	Male
Actual	Female	223	35
	Male	39	186

테스트 데이터 역시 여성을 남성으로 잘못 판별하는 경우보다 남성을 여성으로 잘못 판별하는 경우가 더 많은 것을 확인할 수 있다. 그러나 [Q6-3]과 비교하였을 때, 여성을 남성으로 잘못 판별한 경우와 남성을 여성으로 잘못 판별한 경우의 차이가 매우 작아 유의미하지 않은 변수가 제거되었다고 판단할 수 있다.

추가적으로, 두 Confusion matrix 를 통해 여성과 남성이 비슷한 비율로 존재하는 것을 확인할 수 있다. 따라서 simple accuracy 를 성능 판별 지표로 사용하여도 큰 문제가 없다는 것을 아래 지표들을 통해 확인할 수 있다.

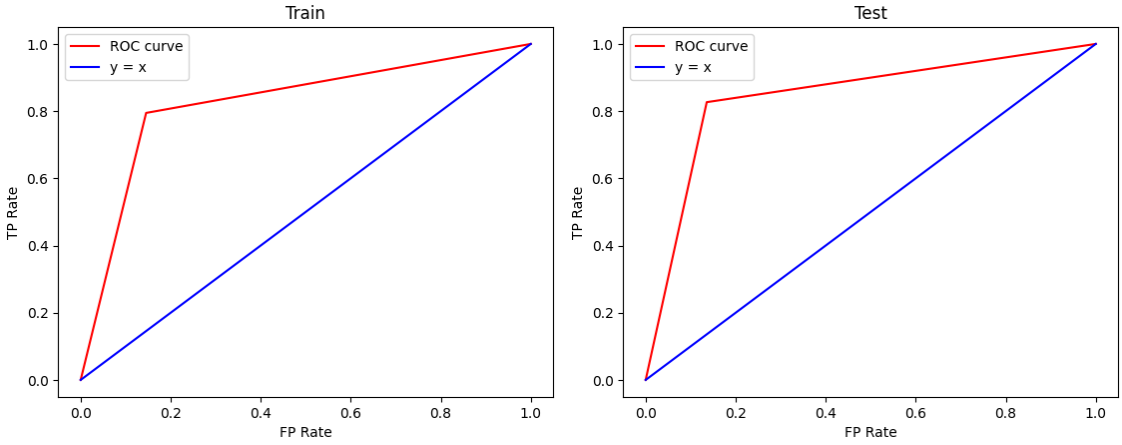
	ACC	BCR	F1
Train	0.8247	0.8241	0.8187
Test	0.8468	0.8453	0.8341

학습 데이터와 테스트 데이터에 대한 성능 지표들을 살펴보면, 현재 분할된 데이터셋에 대해서는 테스트 데이터에 대한 성능이 조금 더 좋은 것을 확인할 수 있다. 그러나 이는 설정된 seed 에 따라 달라질 수 있는 부분이므로 무조건적으로 테스트 데이터에 대한 성능이 더 좋다고 이야기하기는 어려우나, 모델이 훈련 데이터셋에 과적합되지 않았다고 판단할 수 있는 근거가

된다. 현재 학습 데이터와 테스트 데이터에 대해 세 지표가 모두 우수한 성능을 보이고 있으며, 세 지표 사이에 값의 차이가 크지 않은 것을 확인할 수 있다. 이는 데이터셋이 불균형하지 않아, 모든 y 의 예측값이 한 성별로 쏠리는 현상이 발생하지 않았기 때문이다. [Q6-3]과 비교하였을 때, 각 지표들에 대한 값은 약 2~3% 정도씩 감소했다. 그러나 변수의 개수가 거의 절반 수준으로 줄어든 것을 감안하면, 오히려 중요하지 않은 변수들이 잘 제거되었다고 판단할 수 있다. 또, 약 80%의 성능은 절대적인 수치로 보아도 결코 낮지 않은 성능이다. 따라서 [Q6]보다 [Q7]의 모델이 선호된다.

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

Data	AUROC
Train	0.8252
Test	0.8472



[Q6-4]와 마찬가지로, 계산된 AUROC 값은 이전에 보았던 세 지표처럼 매우 비슷한 값이 도출되었으며, ROC curve 역시 매우 유사한 형태를 띠고 있는 것을 확인할 수 있다. [Q6-4]의 AUROC 값과 비교해보았을 때, AUROC 값이 약간 감소한 것을 볼 수 있으나, 7 개의 변수를 제거하였다는 점을 고려하면 매우 준수한 성능을 보여주고 있음을 알 수 있다. 현재는 테스트 데이터셋에 대한 AUROC 값이 더 높으나 seed 를 변경하는 경우에도 동일한 결과가 나올지는 보장할 수 없다.

[Q8] [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.

[다변량데이터분석][과제 2]

다음은 Training dataset 에 대한 AUROC, 각 방법론을 통해 변수 선택 후 모델링을 완료하기까지의 소요 시간, 선택된 변수를 나타낸 표이다.

	AUROC	소요 시간(sec)	선택된 변수
Forward Selection	0.9811	75.1633	'median', 'Q25', 'IQR', 'meanfun', 'minfun'
Backward Elimination	0.9778	334.8818	'sd', 'median', 'Q25', 'Q75', 'IQR', 'skew', 'sp.ent', 'sfm', 'mode', 'centroid', 'meanfun', 'minfun', 'mindom', 'modindx'
Stepwise Selection	0.9801	120.2862	'median', 'Q25', 'Q75', 'IQR', 'mode', 'meanfun', 'minfun', 'maxfun', 'modindx'

가장 먼저, AUROC 값은 Forward Selection > Backward Elimination > Stepwise Selection 이나, 모두 비슷한 수준의 값을 가지고 있는 것을 확인할 수 있다. 소요 시간은 Forward Selection < Stepwise Selection < Backward Elimination 인 것을 확인할 수 있다. 선택된 변수의 수는 Forward Selection = 5, Backward Elimination = 14, Stepwise Selection = 9 인 것을 확인할 수 있다.
위 결과만을 놓고 보면, AUROC 값이 가장 높고 소요 시간이 가장 짧으며 선택된 변수가 가장 적은 Forward Selection 방법론을 선택하는 것이 타당해 보인다. 다만, seed 를 여러 번 바꿔가면서 추가적인 실험을 통해 정말로 방법론들 사이에 성능 차이가 있는지를 따져볼 필요가 있다. 또한, 소요 시간을 측정할 때 모델링 시간까지 포함하였기 때문에 변수의 개수가 적을수록 소요 시간이 짧은 것으로 추정된다. 모델링 과정을 제외하고 소요 시간을 다시 측정한다면 소요 시간이 Backward Elimination < Forward Selection < Stepwise 일 가능성이 존재한다. Backward Elimination 은 4 개의 변수를 제외하는 과정을 거쳤고, Forward Selection 은 5 개의 변수를 선택하는 과정을 거쳤기 때문이다.
세 방법론 모두 모든 입력 변수를 사용한 모델이나, 직접 입력 변수를 선정한 모델보다 약 15% 정도의 성능 향상을 보였다.
추가적으로 눈에 띄는 점은 세 방법론 모두 'median', 'Q25', 'meanfun', 'minfun' 변수를 선택했는데, 이 변수들은 모두 [Q5]에서 선택되었던 변수들이다. 따라서 입력 변수들 중 'median', 'Q25', 'meanfun', 'minfun' 이 네개의 변수들이 성별을 예측하는데 있어 중요한 변수들이라고 판단할 수 있다.

다음은 Validation dataset 에 대한 각 방법론의 AUROC, ACC, BCR, F1 을 나타낸 표이다.

	AUROC	ACC	BCR	F1
Forward Selection	0.9984	0.9956	0.9957	0.9957

Backward Elimination	0.9960	0.9911	0.9914	0.9914
Stepwise Selection	0.9984	0.9956	0.9957	0.9957

세 방법론 모두 모든 평가지표에서 매우 우수한 성능을 보이고 있음을 확인할 수 있다. 이는 종속변수의 클래스가 불균형하지 않기 때문이다. 눈에 띄는 점은, Forward Selection 과 Stepwise Selection 의 평가 지표 값들이 동일하다는 점이다. 선택된 변수들을 확인해보면 Forward Selection 에서는 'median', 'Q25', 'IQR', 'meanfun', 'minfun' 변수들이, Stepwise Selection 에서는 'median', 'Q25', 'Q75', 'IQR', 'mode', 'meanfun', 'minfun', 'maxfun', 'modindx' 변수들이 선택된 것을 볼 수 있다. Forward Selection 에서 선택된 변수들은 모두 Stepwise Selection 에 포함되어 있다. 이를 통해 'Q75', 'mode', 'maxfun', 'modindx' 변수는 성능 향상에 있어 필요하지 않은 변수임을 알 수 있다. 또한 위 표를 통해 Forward Selection 이 모든 지표에 있어 가장 우수한 성능을 보일 뿐 아니라, 가장 단순한 구조를 가지고 있어 비용, 시간적 측면에서도 가장 우수한 모델임을 알 수 있다.

[Q9] AUROC 를 Fitness function 으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오. 작성한 함수를 이용하여 GA 기반 변수 선택을 수행하고, 선택된 변수를 사용한 Logistic Regression 의 Validation dataset 에 대한 분류 성능(AUROC, Accuracy, BCR, F1-Measure), 변수 감소율, 수행 시간의 세 가지 관점에서 Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용한 Logistic Regression 과 비교해보시오.

Genetic Algorithm 에서 선택된 변수는 'sd', 'median', 'Q25', 'Q75', 'IQR', 'skew', 'sfm', 'mode', 'centroid', 'meanfun', 'meandom', 'mindom', 'modindx'이다.

	AUROC	**ACC**	**BCR**	**F1**	**변수 개수**	**소요 시간(sec)**
Forward Selection	0.9984	0.9956	0.9957	0.9957	5	75.1633
Backward Elimination	0.9960	0.9911	0.9914	0.9914	14	334.8818
Stepwise Selection	0.9984	0.9956	0.9957	0.9957	9	120.2862
Genetic Algorithm	0.9918	0.9822	0.9825	0.9828	13	33.0993

위 결과를 보면, Genetic Algorithm 은 가장 빠르게 수행되었지만, 다른 방법론들에 비해 성능이 조금 떨어지는 것을 알 수 있다. 다만, 해당 결과는 하이퍼파라미터에 의해 바뀔 수 있으므로 다양한 하이퍼파라미터 조합을 가지고 실험을 해볼 필요가 있다.
선택된 변수를 살펴보면, Genetic Algorithm 은 다른 세 방법론이 공통적으로 선택했던 'minfun'을 선택하지 않은 것을 볼 수 있다. Genetic Algorithm 이 'minfun'을 선택하지 않았기 때문에 다른 방법론들에 비해 성능이 조금 떨어지는 것으로 예상된다.
또한, Genetic Algorithm 은 변수 감소율 측면에 있어서도 Forward Selection, Stepwise Selection 보다 좋지 못한 것을 알 수 있다.
종합적으로 고려하였을 때, 본 데이터셋에 대해서는 Forward Selection, Stepwise Selection, Backward Elimination, Genetic Algorithm 순으로 사용을 권장한다.

[Q10] Genetic Algorithm 에서 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 27 가지 이상의 조합)하여 각 조합에 대한 변수 선택 결과에 대해 본인만의 생각을 더해 해석해보시오.

Genetic Algorithm 에 존재하는 다양한 하이퍼파라미터 중 population_size, n_gen, mutation_rate 의 값을 변경해가면서 성능이 어떻게 변화하는지 관찰해보고자 한다.
population_size = [10, 30, 50], n_gen = [5, 10, 15], mutation_rate = [0.05, 0.1, 0.2]를 후보 값으로 선정하여 각 조합에 대한 성능을 계산하였다. 그 결과는 아래와 같다.

	Hyperparameter			Performance measure				
	population_size	n_gen	mutation_rate	AUROC	ACC	BCR	F1	Number of Variables
1	10	5	0.05	0.9897	0.9822	0.9828	0.9826	14
2	10	5	0.1	0.9897	0.9822	0.9828	0.9826	15
3	10	5	0.2	0.9897	0.9822	0.9828	0.9826	12
4	10	10	0.05	0.9897	0.9822	0.9828	0.9826	14
5	10	10	0.1	0.9897	0.9822	0.9828	0.9826	15
6	10	10	0.2	0.9897	0.9822	0.9828	0.9826	12
7	10	15	0.05	0.9897	0.9822	0.9828	0.9826	14
8	10	15	0.1	0.9897	0.9822	0.9828	0.9826	12
9	10	15	0.2	0.9960	0.9911	0.9914	0.9914	9
10	30	5	0.05	0.9468	0.9289	0.9296	0.9292	13

11	30	5	0.1	0.9918	0.9822	0.9825	0.9828	14
12	30	5	0.2	0.9918	0.9822	0.9825	0.9828	14
13	30	10	0.05	0.9938	0.9867	0.9868	0.9871	13
14	30	10	0.1	0.9918	0.9822	0.9825	0.9828	13
15	30	10	0.2	0.9918	0.9822	0.9825	0.9828	14
16	30	15	0.05	0.9938	0.9867	0.9868	0.9871	13
17	30	15	0.1	0.9918	0.9822	0.9825	0.9828	13
18	30	15	0.2	0.9918	0.9822	0.9825	0.9828	14
19	50	5	0.05	0.9938	0.9867	0.9868	0.9871	13
20	50	5	0.1	0.9918	0.9822	0.9825	0.9828	12
21	50	5	0.2	0.9938	0.9867	0.9868	0.9871	12
22	50	10	0.05	0.9918	0.9822	0.9825	0.9828	13
23	50	10	0.1	0.9918	0.9822	0.9825	0.9828	6
24	50	10	0.2	0.9602	0.9467	0.9474	0.9465	12
25	50	15	0.05	0.9918	0.9822	0.9825	0.9828	13
26	50	15	0.1	0.9918	0.9822	0.9825	0.9828	6
27	50	15	0.2	0.9602	0.9467	0.9474	0.9465	12

빨간색으로 표기된 9 번 조합은 가장 성능이 좋았던 조합이다. 해당 조합에서 선택된 변수는 'sd', 'Q25', 'Q75', 'IQR', 'skew', 'sp.ent', 'sfm', 'meandom', 'modindx'이다. 거의 모든 변수가 Backward Elimination 에서 선택되었던 변수에 포함되고, Backward Elimination 과 같은 성능을 보이고 있다는 것을 알 수 있다. 그러나 이전에 수행되었던 Forward Selection, Stepwise Seletion 보다는 아직 성능이 조금 떨어지는 것을 확인할 수 있다. 그러나 이전 Genetic Algorithm 에 비해 성능이 향상되었고, 변수가 감소했다는 점, Backward Elimination 과 같은 성능을 보이나 그보다 적은 변수를 사용했다는 점에서 중요하다.

회색으로 칠해진 부분들은 똑같은 성능이 자주 반복되었던 조합들이다. 11 번부터 26 번 사이의 회색 부분들을 보면, n_gen 의 값이 변하더라도 population_size 와 mutation_rate 가 특정한 조합을 이루고 있을 때 성능이 변하지 않는 것을 확인할 수 있다. 이는 generation 이 5 가 되기 전에 안정적인 상태를 취하게 되며, 지금은 generation 의 수가 중요한 하이퍼파라미터가 아니라는 것을 의미한다. 또한, 23 번과 26 번은 'sd', 'Q75', 'IQR', 'sfm', 'centroid', 'modindx', 총 6 개의 변수가 선택되었는데, 이 경우에도 두배 이상의 변수가 선택된 경우와 동일한 성능을 보이고 있는 것을 알 수 있다. 이때 23 번과 26 번에서 선택된 6 개의 변수가 종속변수를 예측하는데 중요한 변수이며, 다른 경우들에서 선택된 추가 변수들은 종속변수를 예측하는데 중요하지 않은 변수일 것이다. 다만, 다른 방법론들을 통해 중요할 것이라고 예측된 변수는 'median', 'Q25', 'IQR', 'meanfun', 'minfun'이며, Genetic Algorithm 이 해당 변수들을 선택하지 않아 성능이 조금 떨어지는 것으로 보인다. 그럼에도 불구하고, 기본적으로 입력 변수들 간 높은

[다변량데이터분석][과제 2]

상관관계를 보이는 경우가 많아 어느 정도 비슷한 성능을 보이는 것 같다. 만약 입력 변수들 간에 상관성이 높지 않은 데이터였다면 Genetic Algorithm 역시 다른 방법론들과 동일한 입력 변수를 선택할 가능성이 있다.
추가적으로 눈에 띄는 점은, population_size = 50 이고, mutation_rate 가 0.2 일 때이다. 즉 21, 24, 27 번을 보면, n_gen 이 커지면서 모델의 성능이 감소하는 것을 볼 수 있다. 이는 mutation_rate 가 높아 세대를 거듭하면서 돌연변이가 많이 생겨 그런 것으로 추정된다. 따라서 population_size 가 클 때, n_gen 이 커지면 mutation_rate 는 작게 설정해주는 편이 좋을 것 같다.
1 번부터 8 번의 회색부분을 보면 population_size 가 10 일 때 다른 하이퍼파라미터들이 바뀌어도 모델의 성능은 변하지 않는 것을 확인할 수 있다. 이는 비교할 염색체의 개수가 적으면 비슷한 수준에서 변수들이 선택되고, 동일한 성능이 측정되는 결과로 이어짐을 알 수 있다.