머신러닝 및 딥러닝
최종발표

20181110493 정정룡

목차

Part 1

EDA

데이터 설명

➡ National Cancer Institute(국립암연구소)에서 진행되는 프로젝트 중 하나인 TCGA(The Cancer Genome Atlas) 에서 기탁된 유방암 조직 샘플데이터

➡ mRNA 발현 정도를 나타내며 전부 로그 변환이 이루어짐

➡ 590명의 환자로부터 17814개의 유전자의 발현정도 측정

데이터 표

	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	C10orf90	ZHX3	ERCC5	GPR98	RXFP3	...	GRIP2	GPLD1	Label
0	1.535833	0.88800	0.118000	0.00000	0.035167	-2.14600	0.623167	-0.33600	-0.574750	0.9755	...	-0.09300	0.796500	1
1	0.493667	1.88875	0.406375	-0.24650	0.104000	-1.69925	-0.298000	0.07025	0.339714	0.1580	...	-0.32050	-0.989667	1
2	0.132250	0.60950	0.548125	1.01875	0.755000	-1.40075	0.209500	-0.69950	0.309000	0.8710	...	-0.03375	0.163500	1
3	0.601750	0.90825	0.483875	0.28275	-1.548667	-2.49100	0.056000	-0.32950	1.104750	0.2645	...	-0.09100	0.980000	1
4	0.833250	1.42750	0.018625	-0.31250	1.443167	-1.78375	-0.806000	-0.26825	0.690750	0.5010	...	-0.00200	-0.816667	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
585	-0.496500	-0.22400	1.350500	-0.20175	-0.327333	0.93175	-0.464000	0.47375	0.264750	0.4695	...	0.18400	-0.067167	1
586	-0.155000	0.85375	0.014125	-0.34125	-0.615500	-1.57675	-0.389667	-0.10775	-1.570375	0.9865	...	-0.14925	0.105000	1
587	0.664000	0.82125	0.663750	-0.74575	0.134667	-1.58775	0.450833	0.51850	0.839750	0.5010	...	-0.08750	-0.393167	1
588	0.213083	1.38550	0.428625	0.26900	-0.201333	-1.93400	-0.371167	-0.33000	0.507125	0.1695	...	-0.30675	1.305833	1
589	0.939833	-0.02300	-0.100250	0.50475	-1.090333	-2.17625	-0.324833	-0.33075	0.149750	0.0045	...	-0.35075	0.502333	0

590 rows × 17815 columns

전체 데이터는 590개의 행과 17814개의 열을 가짐

열 별 결측치 개수

결측치 개수	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
해당 되는 열 개수	212	92	64	44	30	26	18	16	7	8	1	2	8	4	0	1	1

- 총 534개의 열에서 최소 1개 이상의 결측치 존재
- 본 발표에서는 결측치가 존재하는 열 제거하고 진행

Levene's test

- 정규성을 만족할 수 없어 Levene's test로 등분산 검정 진행

Shapiro_Wilks test

- 표본수가 2000미만이므로 Shapiro_Wilks test 로 정규성 검정 진행

- 독립성과 등분산성을 가정하고 진행

=> 등분산과 정규성 둘 다 만족하는 변수

=> 등분산을 만족하지 않거나,
등분산을 만족하나 정규성을 만족하지 못하는 변수

Student t-test

- 정규성
- 등분산성

Wilcoxon rank sum test

- 정규성 X

=> 정규성 따르며 평균 차이가 유의한 변수 : 874개

=> 정규성 따르나 평균 차이가 유의하지 않은 변수 : 262개

=> 정규성 따르지 않지만 평균 차이가 유의한 변수 : 12182개

=> 정규성 따르지 않고 평균 차이도 유의하지 않은 변수 : 3962개

Part 2

Modeling 1

- Logistic, Random Forest, XGBoost 사용

- 표준화 진행

- train: test=7:3 으로 진행

```
X_train , X_test, y_train , y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

```
scaler=StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

Logistic Regression

- 기본 모델로 penalty ='none', solver = 'saga'

Label	Precision	recall	F1-score
0	0.48	1.00	0.65
1	1.00	0.89	0.94

Accuracy = 0.90

Random Forest

- 기본 모델로 진행

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

Accuracy = 0.99

XGBoost

- 기본 모델로 진행

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

Accuracy = 0.99

Part 3

PCA, ICA Modeling 2

PCA

➡ 정규분포 가정 필요

➡ 주성분은 분산이 최대가 되는 축을 추출

➡ 주성분끼리 직교하는 구속조건하 분산을 최대화

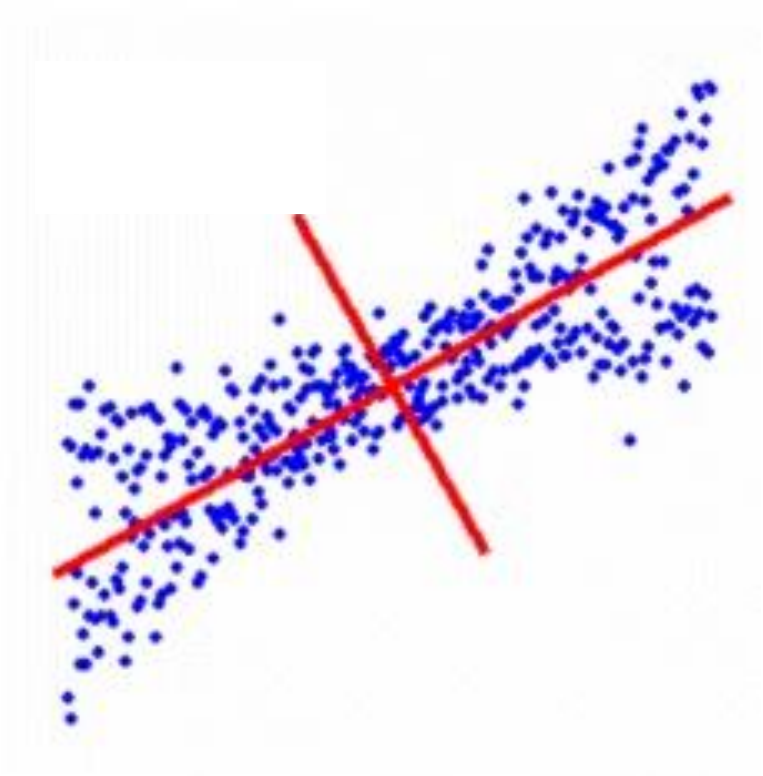
ICA

➡ 데이터가 통계적으로 독립하여 정규분포를 따르지 않는다는 가정

➡ 독립성분은 독립성이 최대가 되는 축을 추출

PCA, ICA Modeling 2

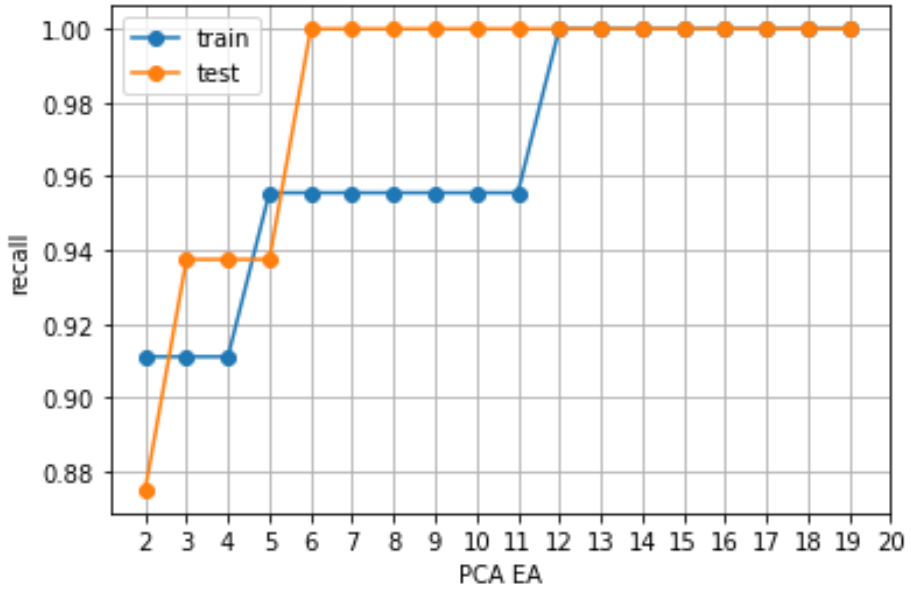
PCA



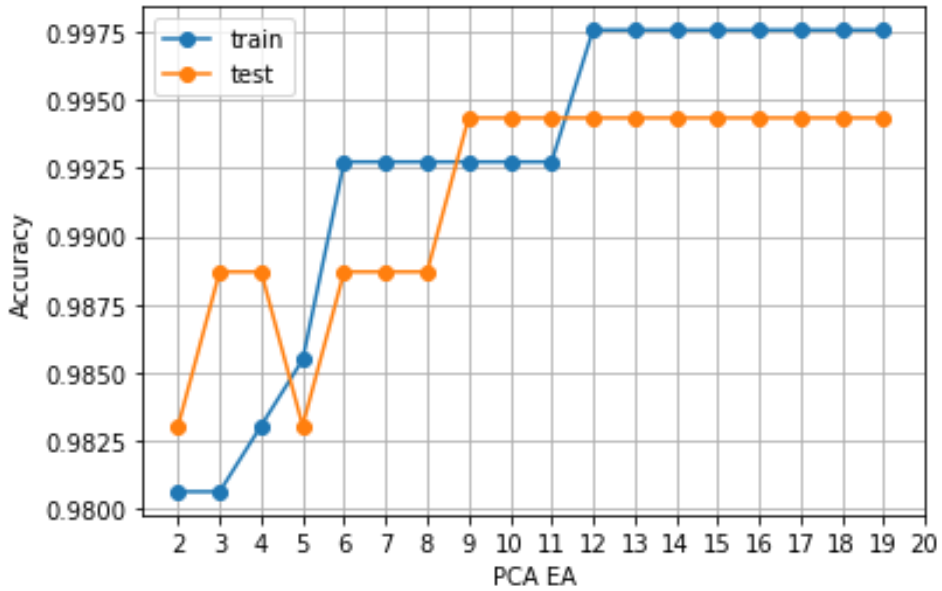
ICA

Logistic Regression - PCA

recall



Accuracy



주성분 12개가 적절할 것으로 보임

Label	Precision	recall	F1-score
0	0.48	1.00	0.65
1	1.00	0.89	0.94

원변수
(17280개)

time : 14.534939초
Accuracy : 0.90

Label	Precision	recall	F1-score
0	0.80	1.00	0.89
1	1.00	0.98	0.99

정규성가정만족하는변수
(874개)

time : 0.483662초
Accuracy : 0.98

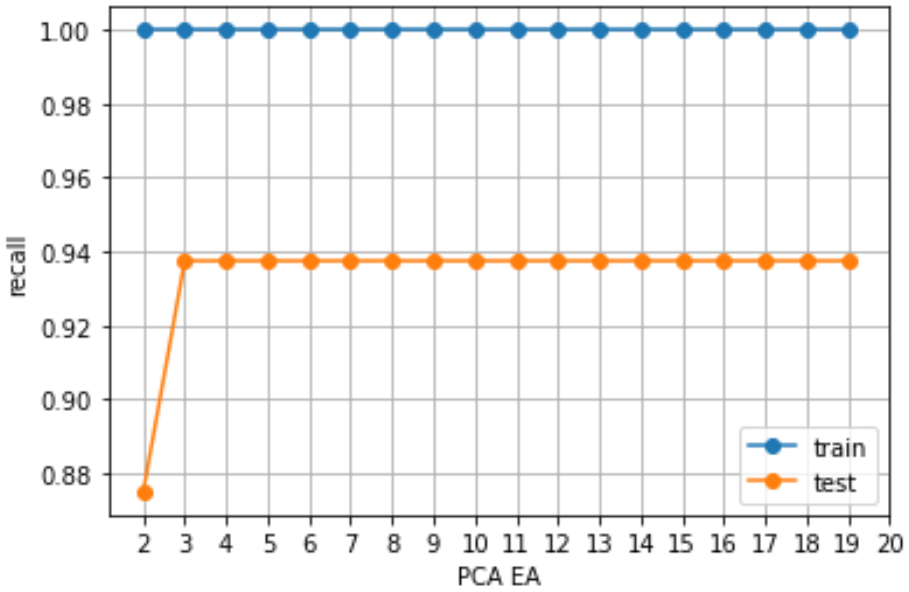
Label	Precision	recall	F1-score
0	0.94	1.00	0.97
1	1.00	0.99	1.00

주성분12개

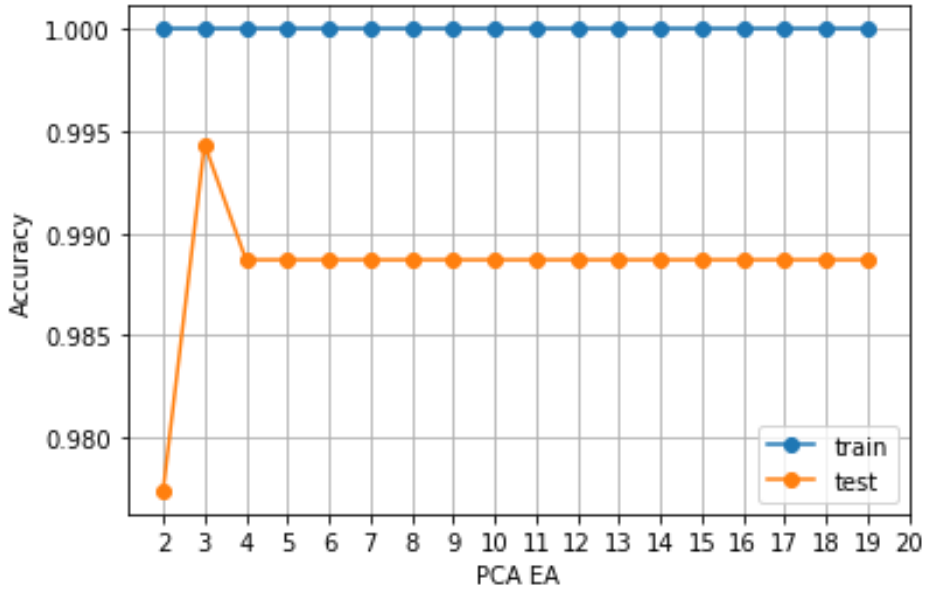
time : 0.032828초
Accuracy : 0.99

Random Forest - PCA

recall



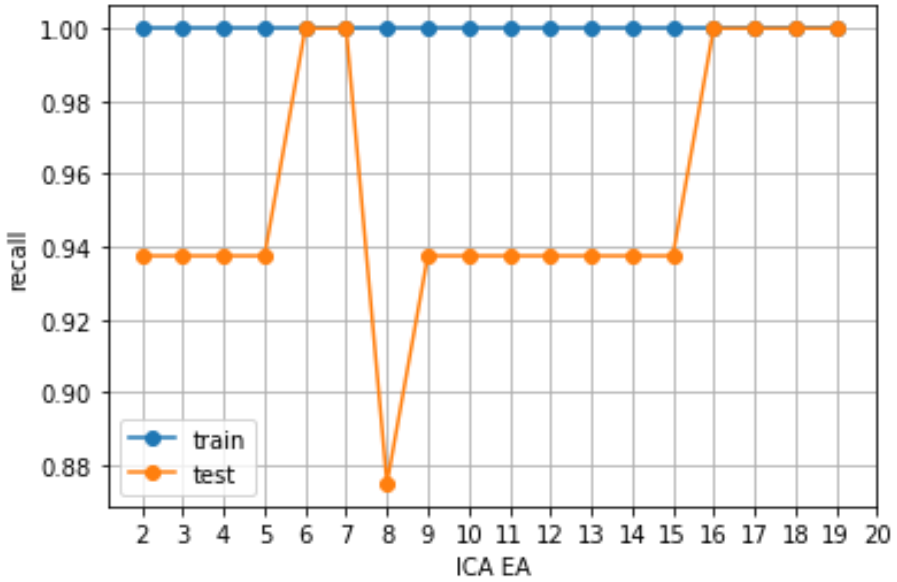
Accuracy



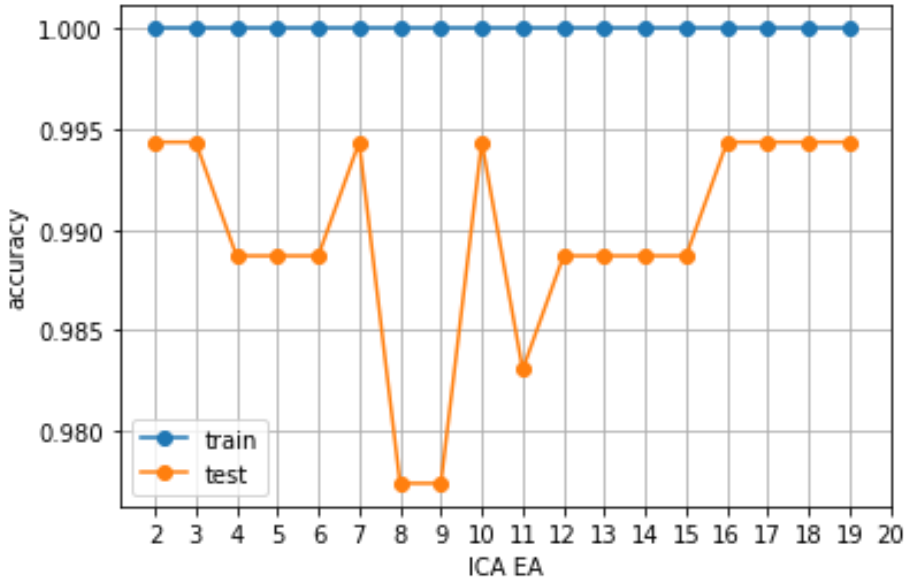
주성분 3개가 적절할 것으로 보임

Random Forest - ICA

recall



Accuracy



독립성분 7개가 적절할 것으로 보임

PCA, ICA Modeling 2

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

원변수
(17280개)

time : 1.625587초
Accuracy : 0.99

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

주성분 3개

time : 0.158364초
Accuracy : 0.99

Label	Precision	recall	F1-score
0	0.94	1.00	0.97
1	1.00	0.99	1.00

독립성분 7개

time : 0.577426초
Accuracy : 0.99

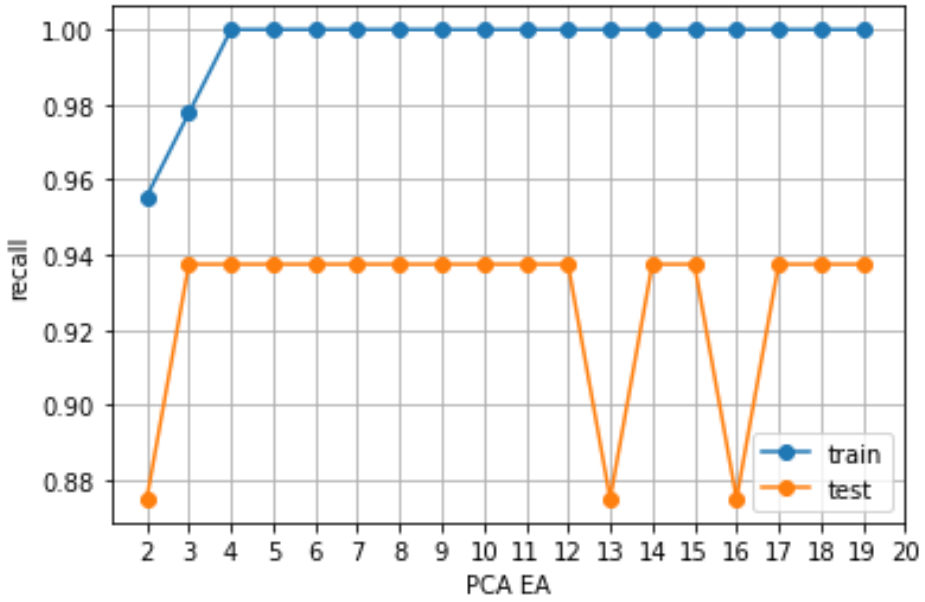
Label	Precision	recall	F1-score
0	0.94	1.00	0.97
1	1.00	0.99	1.00

주성분 3개 + 독립성분 7개

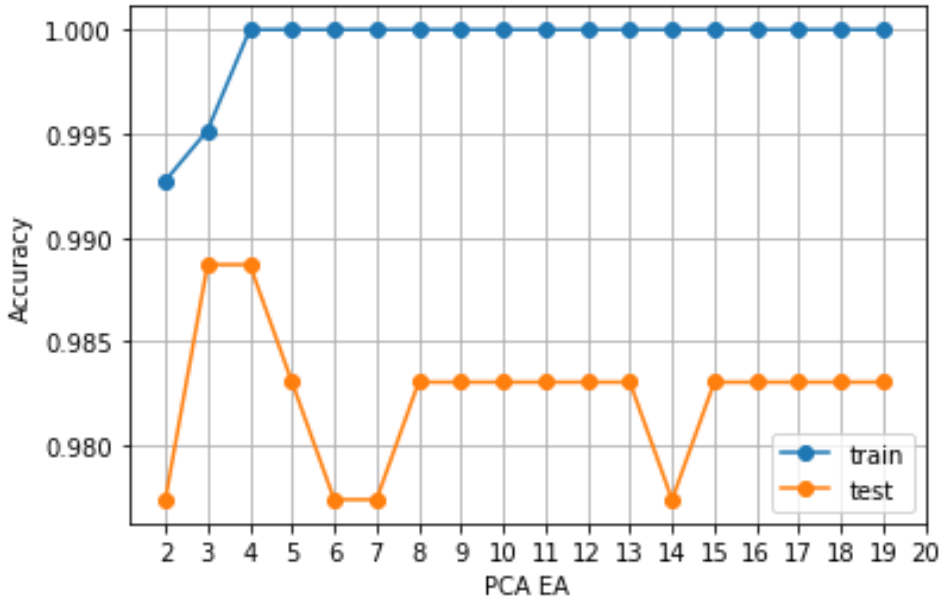
time : 0.668702 초
Accuracy : 0.99

XGBoost- PCA

recall



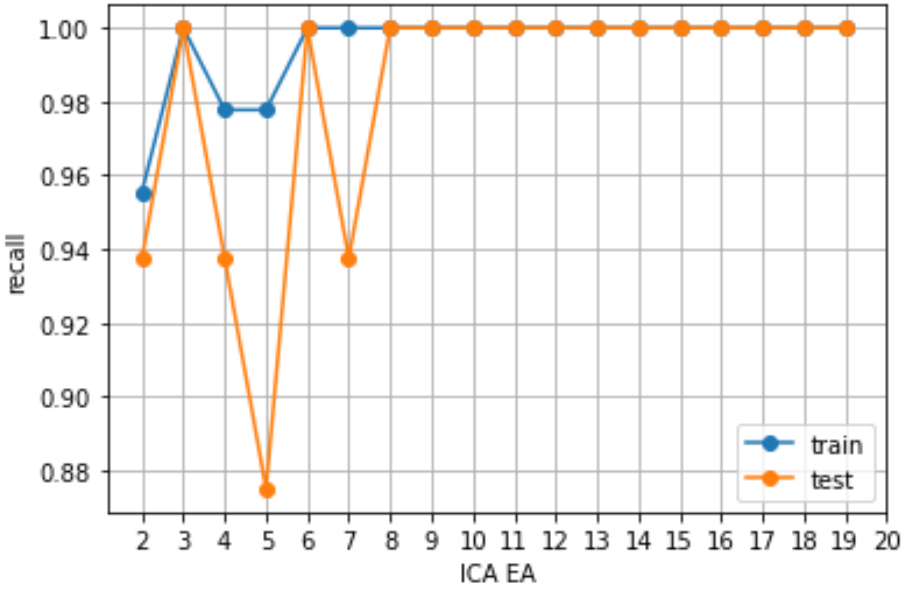
Accuracy



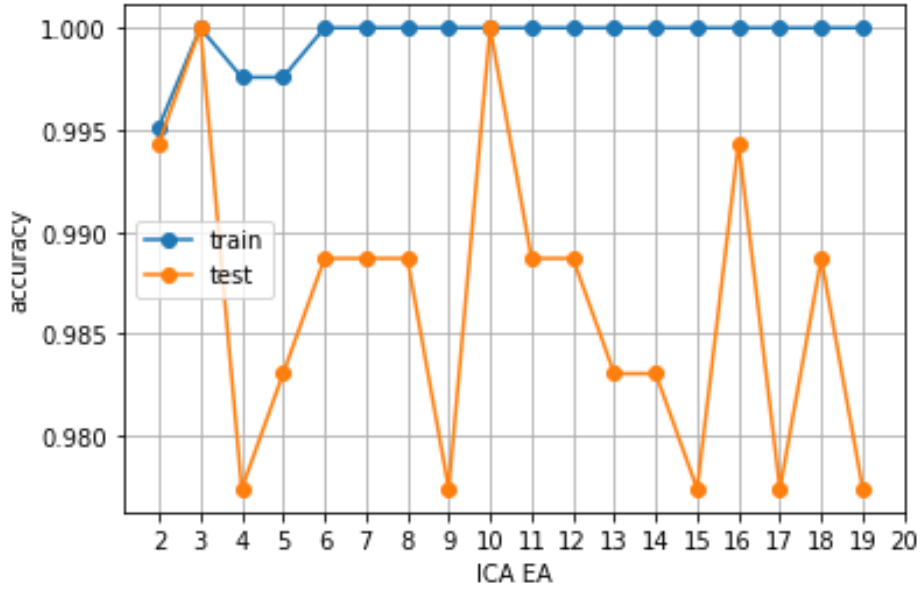
주성분 4개가 적절할 것으로 보임

XGBoost- ICA

recall



Accuracy



독립성분 3개가 적절할 것으로 보임

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

원변수
(17280개)

time : 4.271164초
Accuracy : 0.99

Label	Precision	recall	F1-score
0	0.94	0.94	0.94
1	0.99	0.99	0.99

주성분 4개

time : 0.068520초
Accuracy : 0.99

Label	Precision	recall	F1-score
0	1.00	0.94	0.97
1	0.99	1.00	1.00

독립성분 3개

time : 0.517424초
Accuracy : 0.99

Label	Precision	recall	F1-score
0	0.94	0.94	0.94
1	0.99	0.99	0.99

주성분 4개 + 독립성분 3개

time : 0.500155초
Accuracy : 0.99

Part 4

Kmeans

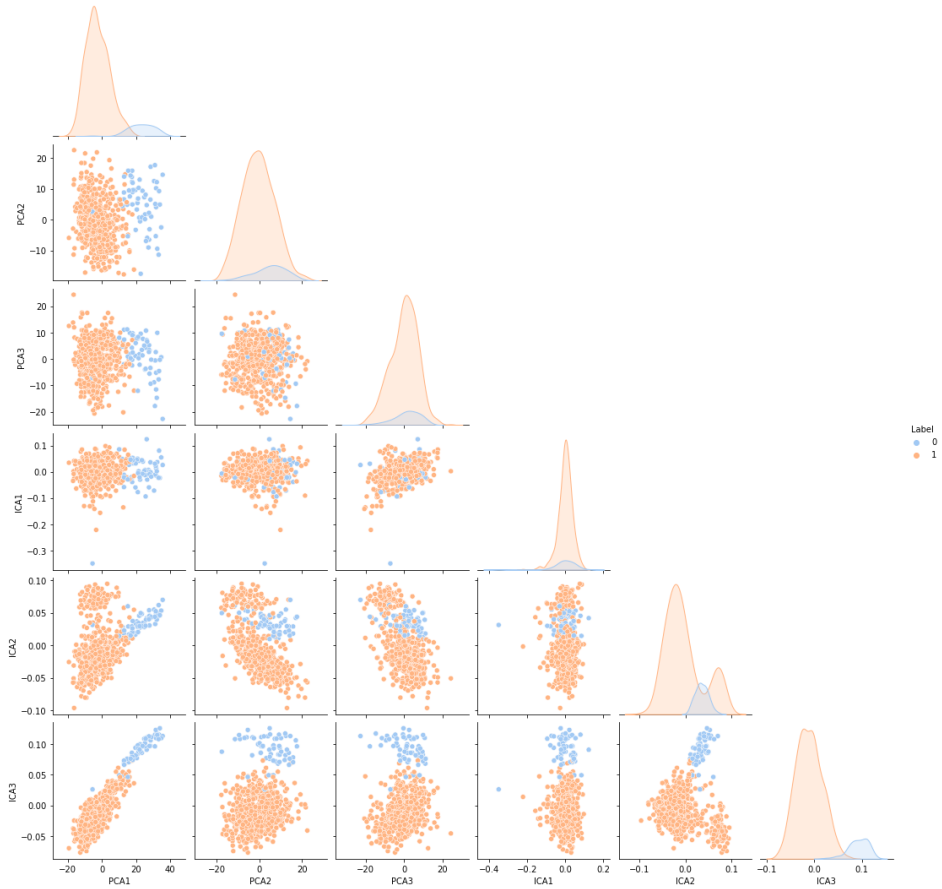
Scaling, PCA, ICA

➡ 표준화

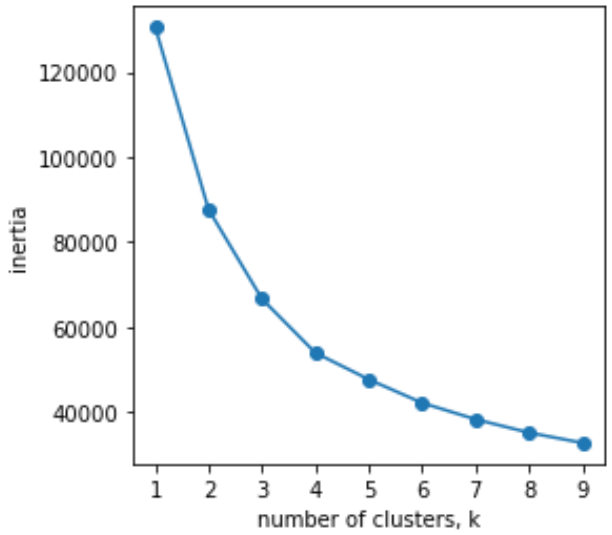
➡ 정규성 따르는 변수는 PCA,
3개의 주성분 사용

➡ 정규성 따르지 않은 변수는 ICA,
3개의 독립성분 사용

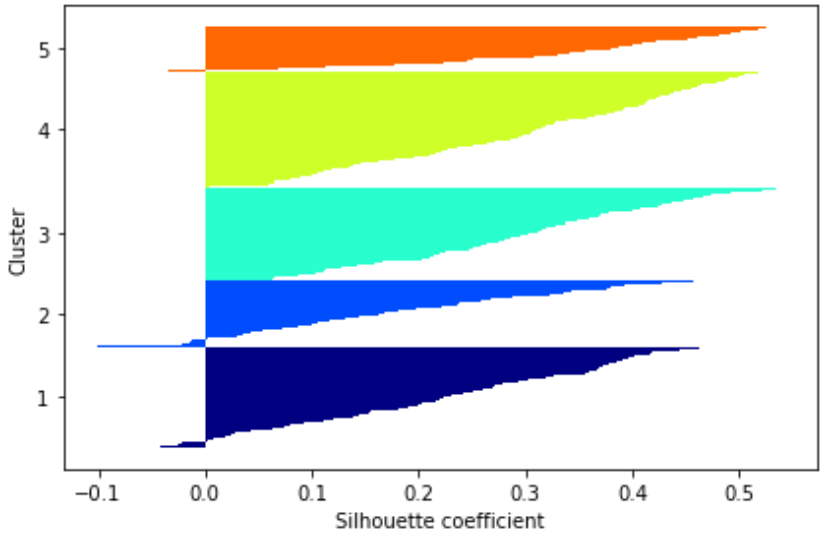
Pair plot



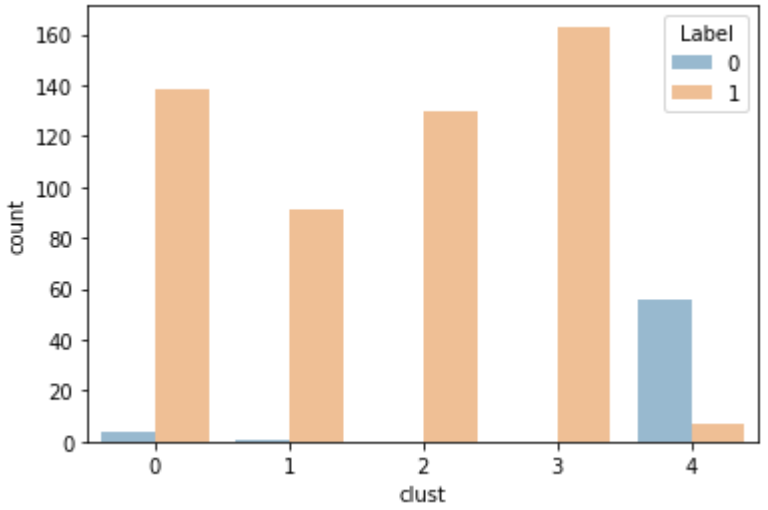
Elbow plot



Silhouette plot



Bar plot



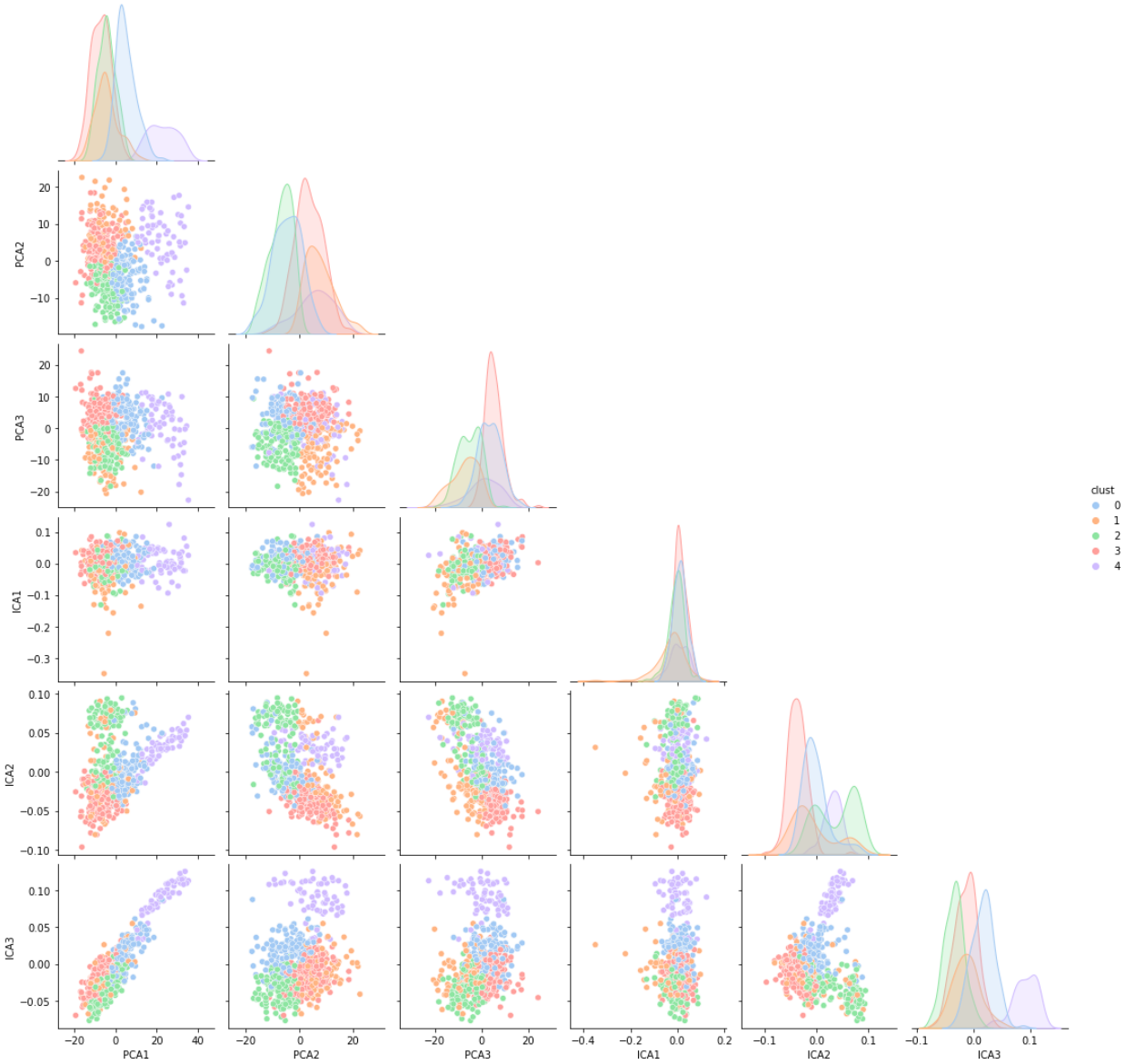
Cluster table

	PCA1	PCA2	PCA3	ICA1	ICA2	ICA3
0	4.97	-4.76	3.37	0.01	-0.002	0.016
1	-4.48	7.56	-7.397	-0.029	-0.006	-0.012
2	-4.3614	-7.132	-5.255	-0.005	0.03	-0.344
3	-7.183	3.425	5.232	0.009	-0.03	-0.014
4	22.9241	5.541	0.508	0.007	0.03	0.09

4번 집단의 특징으로

PCA1이 압도적으로 높으며

ICA3에서 가장 높은 수치를 가진다



Pair plot

PCA1 과 ICA3이
유의하게 다른 변수와 결합하여
유방암에 걸린 집단이 나뉘어짐을
확인 할 수 있다

PCA1 해석

	PCA1
ECM2	0.762121
EBF3	0.749919
CCNF	0.748439
CUGBP2	0.731100
SPARCL1	0.712368
CLIC2	0.708374
...	...

EBF3 의 경우 다형 교모세포종 및 위암을 비롯한 여러 악성 종양에서 역할을 하는 것으로 알려짐.

SPARCL1 의 경우 췌장암과 관련 된 것으로 알려짐

최종 결론

결론 및 향후 과제

➡ Logistic 모형의 경우 PCA를 활용했을 때 극대화된 성능향상과 시간 단축이 이루어 짐

➡ Kmeans를 통해 암에 걸린 사람들의 집단을 도출해낼 수 있었음

➡ Shapiro test의 경우 매우 엄격하기에 정규성을 가정할 수 있는 변수들을 잃어버렸을 가능성이 존재. 또한 매우 많은 변수를 검정했기 때문에 1종 오류의 증가를 감안.
다중 비교법을 활용 한다면 더 좋은 결과를 도출 할 수 있을 것.

➡ Clustering 에서 각 성분들에 대한 해석을 도메인을 가지고 더 진행한다면 다채로운 결과 해석이 가능할 것. 또한 집단 별로 더 자세한 해석을 진행한다면 암에 걸릴 가능성이 있는 집단도 찾을 수 있을 것.

감사합니다