CLASSIFICATION AND CLUSTERING USING WISCONSIN BREAST CANCER DATASET

DECEMBER 14, 2023

이민주

Contents

1. 변수 설명

출처: *https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic*

위 데이터는 위스콘슨 주에서 유방암에 걸린 환자들의 유방 상태에 대한 정보를 담은 데이터이다. 총 568명의 환자들의 관측치를 갖고 있으며 (결측값 0), 총 32개의 변수로 구성된다. 첫 변수는 환자들의 ID이며, 두 번째 변수는 Diagnosis(환자들의 진단 상태)로 악성(M)과 양성(B)를 담은 binary 변수이다.

3~12번째 변수는 Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension들의 평균을 담은 데이터이며, 모두 연속형 변수이다.

13~22번째 변수 또한 Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension 변수들을 담고 있지만, 이번에는 이들의 Standard Error을 담고 있다.

23~32번째 변수 또한Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension 변수들을 담고 있으며, 이번에는 이들의 worst case(각 세포 별 구분들에서 제일 큰 3개 값의 평균)를 담고 있다. 즉, 한 명의 환자에 대한 일정한 실험을 20번 했을 때, 20번 관측한 값의 평균을 mean variable에, 그 20번 관측값의 표준오차를 standard error variable에, 마지막으로 20번 관측한 값들 중 가장 값이 안 좋은 3개 관측값의 평균을 worst variable로 구분하였다.

```
> names(bc_d)
 [1] "ID"                  "Diagnosis"             "Radius_mean"           "Texture_mean"          "Parameter_mean"
 [6] "Area_mean"           "Smoothness_mean"       "Compactness_mean"      "Concavity_mean"        "Concave_points_mean"
[11] "Symmetry_mean"       "Fraction_dimension_mean" "Radius_se"           "Texture_se"            "Parameter_se"
[16] "Area_se"             "Smoothness_se"         "Compactness_se"        "Concavity_se"          "Concave_points_se"
[21] "Symmetry_se"         "Fraction_dimension_se" "Radius_worst"          "Texture_worst"         "Parameter_worst"
[26] "Area_worst"          "Smoothness_worst"      "Compactness_worst"     "Concavity_worst"       "Concave_points_worst"
[31] "Symmetry_worst"      "Fraction_dimension_worst"
```

id	환자 식별 번호
dignosis	종양상태(M=악성, B=양성)
radius	세포 반지름(크기)
texture	질감(흑백 사진의 표준편차)
perimeter	세포의 둘레
area	세포의 넓이 (면적)
smoothness	매끄러움
compactness	작은 정도(perimeter^2/area -1)
concavity	오목함(종양의 오목한 부분의 정도)
concave points	오목한곳의 수

symmetry	대칭성
fractal dimension	coastlineapproximation - 1(프랙탈차원)

2. 연구의 목적

각 데이터는 mean, se, worst case 총 3가지로 구분된 사실상 10개의 설명변수로 구성된 데이터이다. 10개의 종양적 특징(Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension)이 Diagnosis(종양 진단상태)에 어떤 영향을 미치는 지 확인하고자 한다.

3. EDA 작업

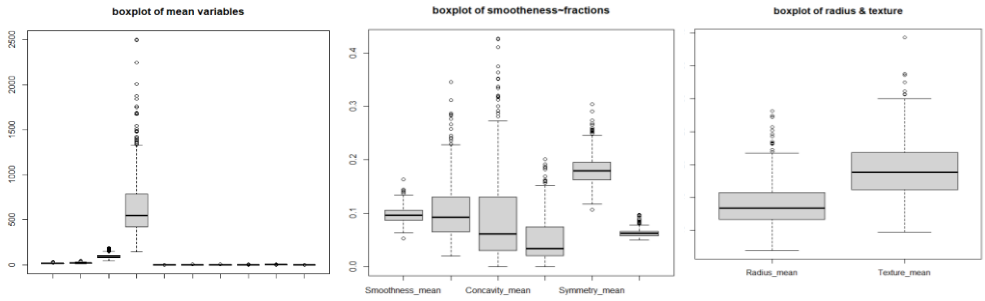


사진 1 Mean 변수들의 Boxplot

사진1은 mean 변수 종양적 특징(Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension) 10개의 boxplot을 그린 것이다. 첫번째 그림에서 보듯이 Area와 parameter 변수의 범위가 다른 변수들에 비해 크기 때문에 둘을 제외해서 살펴보았다.

	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave points	Symmetry	Fraction Dimensior
Min	6.98	9.71	43.79	143.50	0.05	0.02	0.00	0.00	0.11	0.05
1st Q	11.70	16.18	75.14	420.20	0.09	0.06	0.03	0.02	0.16	0.06
Median	13.36	18.86	86.21	548.80	0.10	0.09	0.06	0.03	0.18	0.06
Mean	14.12	19.31	91.91	654.30	0.10	0.10	0.09	0.05	0.18	0.06
3rd Q	15.78	21.80	103.88	782.60	0.11	0.13	0.13	0.07	0.20	0.07
Max	28.11	39.28	188.50	2501.00	0.16	0.35	0.43	0.20	0.30	0.10

사진 2. 각 Mean 변수들의 Summary

두 번째 plot은 Smoothness, Compactness, Concavity, Concave points, Symmetry, Fraction Dimension의 데이터이며, 거의 대부분이 0.1~0.2 근처에서 맴도는 것을 알 수 있다. 하지만,

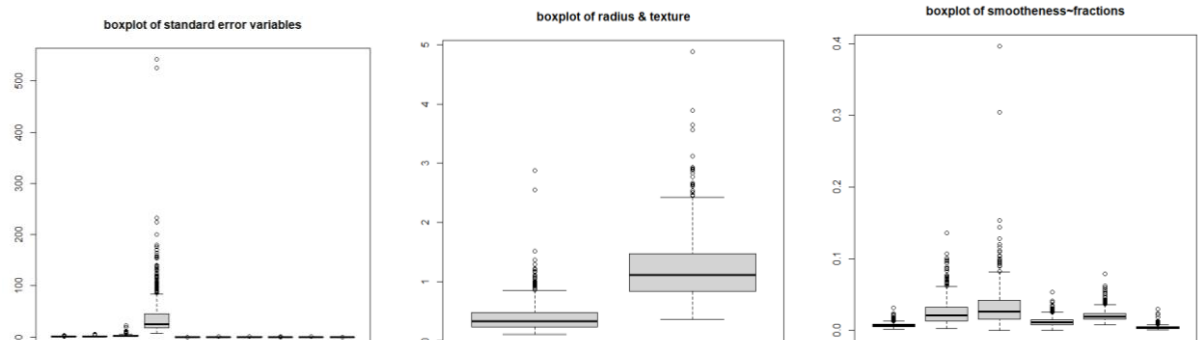


사진 3. Standard Error 변수들의 Boxplot

데이터가 퍼져 있는 정도를 볼 때 Smoothness와 Fraction Dimension의 경우, 분산이 작은 반면 concavity, compactness 변수는 분산이 큰 것을 알 수 있다. Radius와 texture의 경우, 분산이 비슷한 것을 알 수 있으며, 평균이 대략 15, 20 정도에 위치한 것을 알 수 있다.

Standard Error 변수들에서도 같은 현상을 발견할 수 있었다. 이는 Parameter & Area의 scale, Radius & Texture의 scale, 그리고 나머지 (smoothness~fraction dimension) 변수들의 scale이 다르기 때문에 나타나는 현상이다. 이러한 문제를 해결하기 위해 본 연구는 R package의 scale() 함수를 사용하는 대신, 총 두 가지의 standardize method를 적용시켜 데

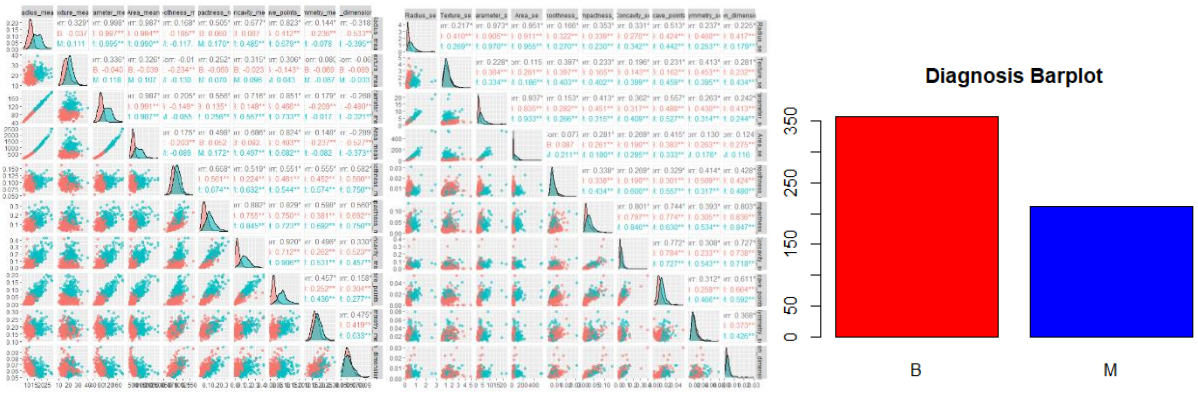


사진 4. Mean 변수, Standard Error 변수의 GGpairs / 실제 반응변수(Diagnosis)의 빈도

이터를 잘 대표할 수 있는 표준화 데이터를 활용해서 분석을 진행하고자 한다.

그에 앞서 먼저 mean, standard error 변수들의 GGpairs plot을 만들어서 변수 간 correlation과 더불어 diagnosis 변수와의 관계성을 파악하고자 하였다. 이를 통해 Perimeter 변수와 Area 변수 간 correlation이 매우 높다는 사실을 파악했다. 이는 일반적 상식으로 볼 때 종양의 둘레와 넓이의 경우, 반지름(radius)에 따라 정비례하기 때문에 나타나는 현상이라고 볼 수 있다. 이처럼, 세 변수(radius, area, perimeter)간 상관관계가 높으며, 이러한 현상을 살펴볼 때 본 변수 중 하나만 남겨놓아야 다중공선성을 해결할 수 있을 것이란 판단이 들었다.

또한, 실제 Diagnosis에서 B(GGpairs에서 빨간색으로 표시)가 많음에도 불구하고 GGpairs에서는 둘의 비율이 비슷해보이며, 때론 M(악성종양)의 비율이 더 커보이는 듯한 양상을 보인다.

4. 데이터 전처리 (Standardization)

4.1 Standardization Method 1

$$\frac{X_{i,j}-mean(X_j)}{mean(X_{i,j+10})} \quad (i = patients(total: 568), j = 3, \dots, 12)$$

평균 변수들을 변수의 평균으로 뺀 후, 표준오차 변수들의 평균으로 나누었다. 즉, 각 환자 개인의 표준오차 정보로 표준화한 것이 아닌 전체 환자에 대한 standardization을 진행하였다.

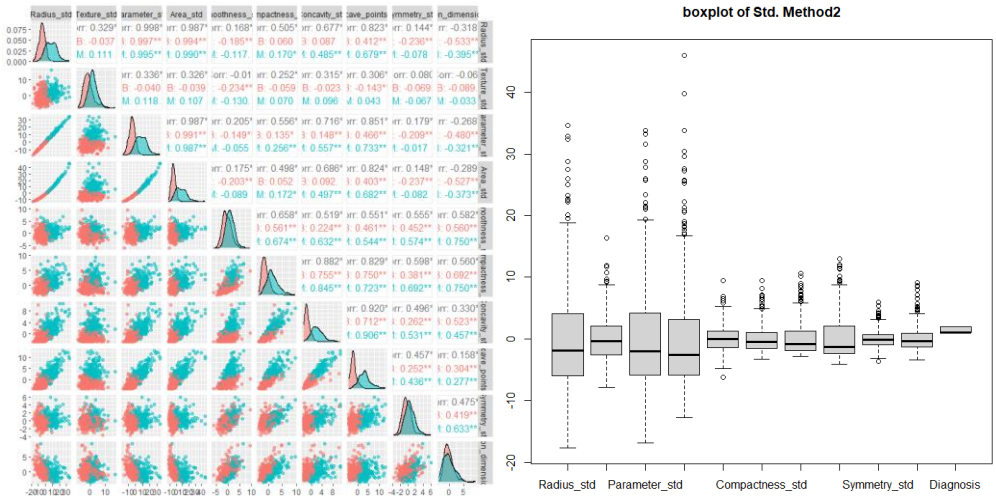


사진 5. Standardization Method 1 결과

결과를 보면, 3개의 변수(area, radius, perimeter) 변수들의 상관관계가 여전히 높다. 또한, 오리지널 데이터가 갖는 반응변수 대표성 부족 문제는 여전히 해결되지 못한 모습을 보인다. Boxplot을 볼 때 평균은 모두 0에 근접하는 모습을 보인다.

4.2 Standardization Method 2

$$\frac{X_{i,j}-mean(X_j)}{X_{i,j+10}} \quad (i = patients(총\,568), j = 3, \dots, 12)$$

이번에는 각 환자들의 standard error 정보로 표준화를 진행하였다. 이를 위해 0의 값을 지닌 standard error 변수들의 관측값을 1로 바꿔준 후, 표준화 작업을 진행하였다.

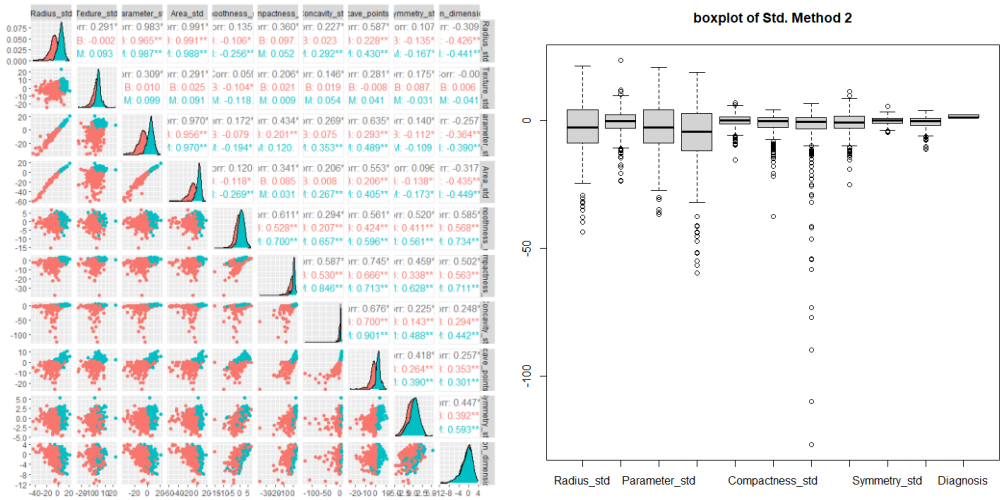


사진 6. Standardization Method 2 결과

표준화 작업 결과, Standardization Method 1의 결과와 비슷하면서 다른 양상을 확인할 수 있다. 먼저, 각 변수들의 분산의 경우, 두 Boxplot에서 비슷한 양상을 보인다. 유일한 차이는 outlier들이 나타나는 변수들의 정보인데, Standardization Method 1의 경우, (Radius, Texture, Perimeter, Area)의 변수에서 큰 이상 값들이 많으며, Standardization Method 2는 Compactness 변수 (뒤에서 5번째) 변수의 음수 이상 값들이 많은 것을 알 수 있다.

이는, 표준화 작업에서 standard error 변수들이 매우 작을 때 (대부분 smoothness ~ fraction dimension처럼 표준오차의 값이 0에 근접한 데이터) 그 값으로 데이터를 나눔으로써 절대값이 커지는 문제가 발생한 것이다.

	Radius_se	Texture_se	Perimeter_se	Area_se	Smoothness_se	Compactness_se	Concavity_se	Concave points_se	Symmetry_se	Fraction Dimension_se
평균	14.12	19.31	91.91	654.3	0.09632	0.10404	0.08843	0.04875	0.1811	0.06277

사진 7. Standard Error 변수들의 평균값

사진 7은 Standardization Method 1로 표준화한 $mean(X_{i,j+10})$ (standard error 변수들의 평균 값)를 표로 작성한 결과이다. Smootheness~Fraction Dimension을 볼 때 값들이 거의 0.05~0.2 정도로 고르게 분포하였으며, 너무 작지 않기 때문에 표준화 후 값이 비이상적으로 불어날 이유가 없다. ***따라서, Standardization Method 1을 사용하여 표준화한 데이터를 통해 clustering 및 classification 진행으로 Diagnosis에 영향을 미치는 변수 및 구조에 대해 파악하고자 한다. 앞서 언급했듯이 Standardization Method 1의 경우 plot화하였을 때 실제 Diagnosis 빈도를 잘 대표하지 못한다는 점에 경각심을 갖고, 시각화 작업 시 변수 하나 하나의 분포를 자세히 살펴보며 데이터의 구조를 면밀히 살펴보고자 한다.***

4.3 Worst Data의 Standardization

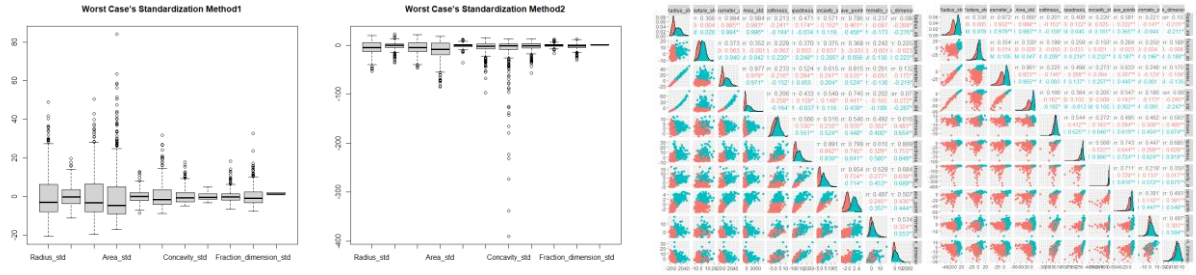


사진 8. Plot of Worst-Case Variables (Standardize Method 1 & 2)

$$\frac{X_{i,j+20}-mean(X_{j+20})}{mean(X_{j+10})} \quad (i = patients(총\,568), j = 3, \dots, 12)$$

마찬가지로 worst case 변수 10개를 Standardization method 1과 2로 표준화했을 때 매우 유사한 결과가 나왔다. Mean, Standard error 변수를 표준화했을 때의 결론처럼, Standardization method 1을 활용한 데이터셋을 제작하여 Mean, Standard error 변수를 표준화한 10개의 데이터와 결합한 총 20개의 설명변수를 활용해 데이터 분석을 진행하고자 한다.

5. EDA (Standardization Ver.)

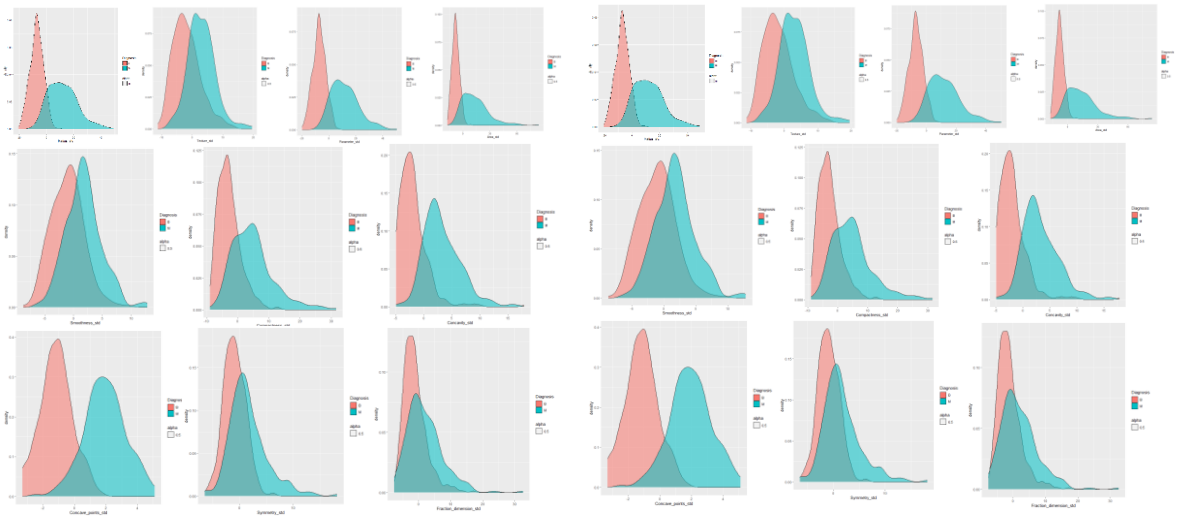


사진 9. Standardize 변수 10개, Worst Case Standardize 변수 10개의 Density Plot

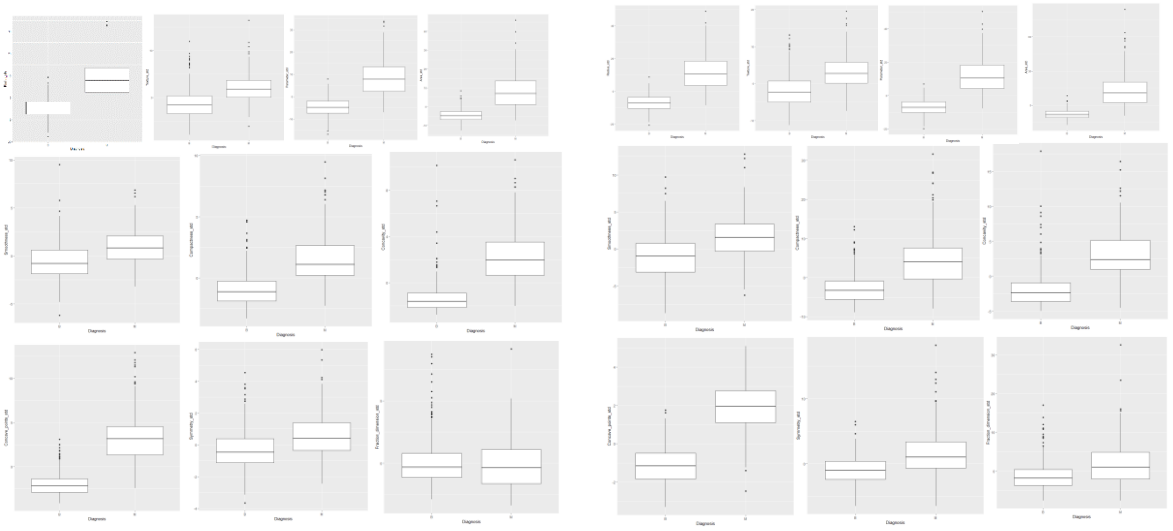


사진 10. Standardize 변수 10개, Worst Case Standardize 변수 10개의 Boxplot

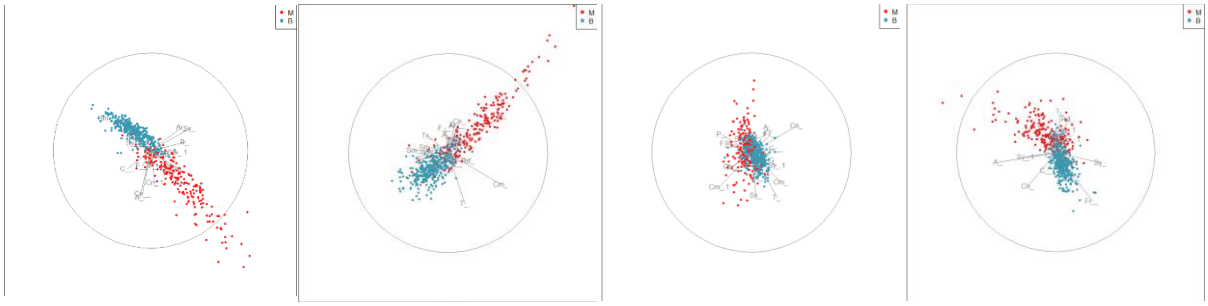


사진 11. 3d Tourr of Standardized Data (20 Variables)

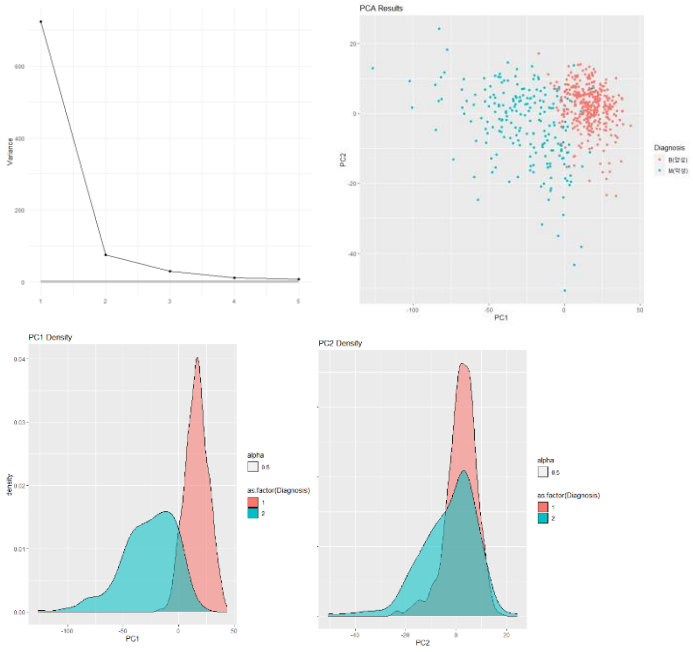
EDA 결과 다음과 같은 특징을 발견하였다.

- 대부분의 density plot에서 Diagnosis 부분이 구분되는 듯한 양상을 보인다. M(악성, 블루)이 뒤쪽에 오는 반면, B(음성)이 앞쪽에 나오는 plot이 그려진다.
- **Radius, Perimeter, Area, Compactness (작은 정도(perimeter^2/area -1)), Concavity(오목함 정도) 변수들의 경우 Diagnosis 구분이 아주 뚜렷**한데, 양성(B) 변수들은 분산이 작고 값도 작은 반면, 악성(M) 변수들은 분산도 크며 값이 크게 나타난다. → 실제 Radius, Perimeter, Area, Compactness 네 변수들은 서로 연관되어 있다는 점이 영향을 미쳤을 가능성 多.
- 반면, **Texture(흑백 사진의 표준편차), Smoothness(매끄러움), Symmetry(대칭성), Fraction Dimension(프랙탈차원)**의 경우, 상대적으로 악성 plot과 양성 plot이 겹치는 듯한 양상을 보이며, 4개의 변수만으로 Diagnosis 상태를 구분하는 데 어려움이 있을 것처럼 보인다.

- Boxplot의 경우 마찬가지로 악성과 양성 간 차이가 뚜렷한 변수들(Radius, Texture, Perimeter, Area, Compactness, Concavity, Concave points)들이 있다.
- 반면, Smoothness, Symmetry, Fraction Dimension 와 같은 변수들의 경우, 두 Diagnosis 상태의 차이가 명확하지 않으며, 이러한 결과는 density plot의 결과와 유사하다는 점을 이해할 수 있다.
- Boxplot은 **Worst Case Standardize** 변수와 Standardize 변수들 간 차이가 존재하는 것을 알 수 있는데, Fraction Dimension의 경우 Standardize 변수에서는 명확한 차이가 없는 반면, worst case에서는 두 boxplot의 분산 및 평균의 차이가 존재하는 것이 보인다.
- 마지막으로 3d Plot으로 우리는 M(악성) 변수들의 경우 outlier 및 분산이 커 보이는 반면, B(양성) 변수들은 분산이 작고 모여 있는 듯한 양상을 보인다.
- **이처럼, 우리는 악성 종양의 형태가 다양한 형태로 나타난다는 것을 알 수 있다.**

6. Dimensionality Reduction

6.1 PCA



	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	26.901	0.830	0.830
PC2	8.688	0.087	0.917
PC3	5.504	0.035	0.952
PC4	3.527	0.014	0.966

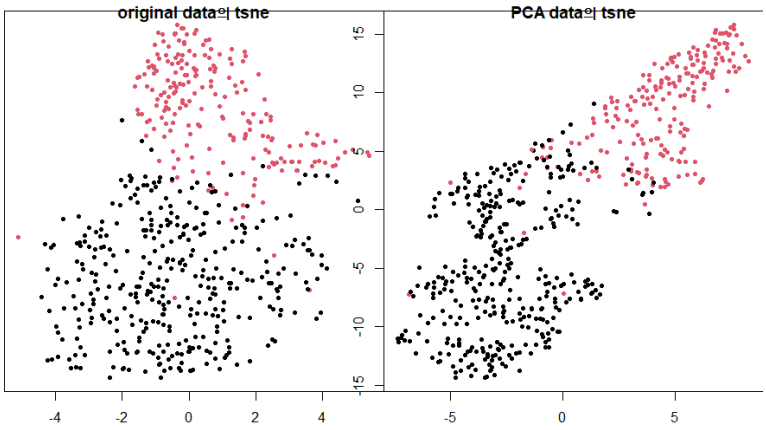
20개의 x변수들을 축소할 수 있는 PCA 분석을 진행하였다. 그 결과 PC2에서 약 91.7%의 설명력을 갖는 것을 알 수 있었으며, 2개의 PC를 활용한 2d plot을 만들 수 있었다.

- EDA 분석에서 확인했듯이 B(양성)는 모여 있는 반면, M(악성)이 퍼져 있는 듯한 양상

을 보인다.

- PC1의 경우, M(블루, 악성)의 분산이 커 보이는 동시에, B(양성, 레드)와 분포의 범위가 달라 확연히 구분되는 양상을 보인다. **(Radius, Perimeter, Area, Compactness (작은 정도(perimeter^2/area -1)), Concavity(오목함 정도) 변수와 유사)**
- 반면, PC2의 경우 M의 분산이 B보다 크지만 분포의 범위가 확연히 구분되지는 않는다. **(Texture(흑백 사진의 표준편차), Smoothness(매끄러움), Symmetry(대칭성), Fraction Dimension(프랙탈차원)와 유사)**
- 이는 우리가 EDA 분석에서 확인한 변수들의 특성과 유사한 점을 보이며, ***PCA 결과가 좋다는 점을 입증한다.***

6.2 T-SNE



원본 데이터(표준화)를 T-sne로 차원 축소한 것과 PCA 데이터를 T-sne로 차원 축소한 것의 차이가 있다는 것을 알 수 있다. 원본 데이터의 경우, 20개의 변수들 간 뭉쳐 있는 부분이 퍼지면서 Diagnosis의 대략적인 분포가 표현되는 것을 알 수 있다. 검은색으로 표시된 M(악성)이 B(양성)보다 분산이 크고,

반면, PCA 데이터의 T-sne는 linear한 관계를 보인다. 본래 2차원의 PCA 데이터는 선형적 특성이 전혀 없는 것을 볼 때, 본 결과는 이상하다고 볼 수 있으며, 이미 2차원으로 축소된 데이터를 변형시킴에 따라 나타난 결과로 볼 수 있다. 하지만, 원본 데이터의 EDA 투어 결과, 일부 측면에서 데이터의 선형성을 발견할 수 있었는데(참고: p8 사진11), 이러한 점을 고려해서 PCA T-sne 데이터와 원본 데이터의 T-sne를 모두 활용해 클러스터링 작업을 진행하고자 한다.

7. Clustering

7.1 Model Based Clustering

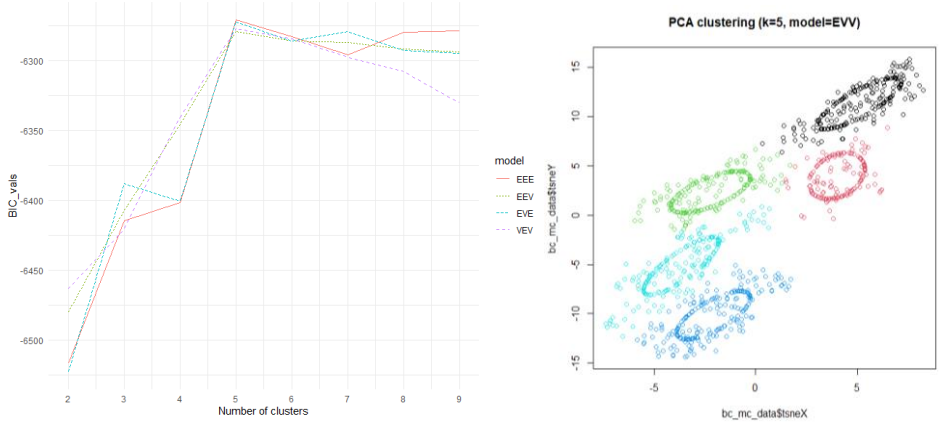


사진 12. PCA 데이터의 Clustering (K=5, EVV)

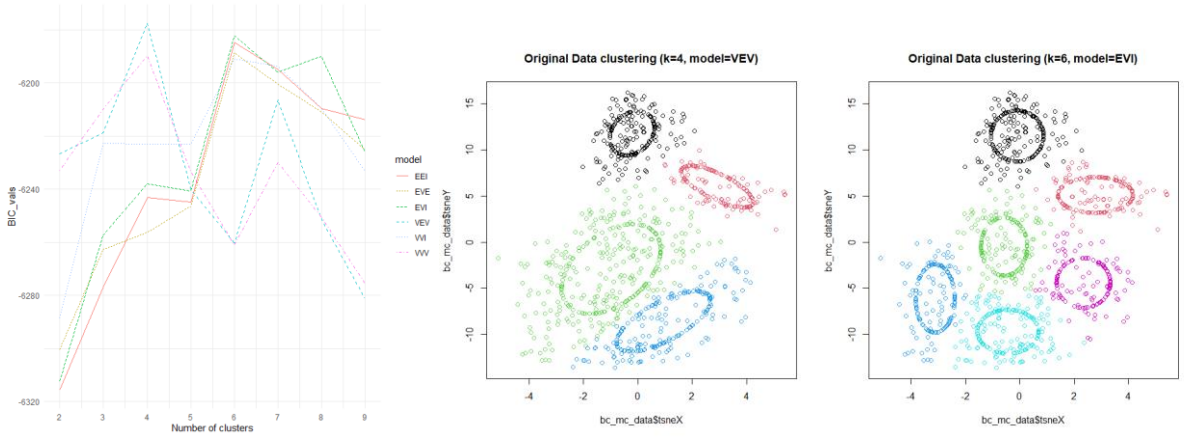


사진 13. Original Data의 Clustering (k=4, VEV) (k=6, EVI)

T-sne로 축소한 PCA데이터와 원본 데이터를 군집화한 결과이다. 이를 통해 우리는 다음과 같은 데이터의 양상을 확인할 수 있었다.

- PCA 클러스터링의 경우, k=5로 군집화하였지만, 좋은 클러스터링이라고 보기 어렵다.
- 원본 데이터의 경우 조금 더 나은 양상을 보인다. 이는 선형적 관계를 갖는 PCA와 달리 Original Data의 경우 더 뭉뚱그려진 형태로 구성되어 있기 때문에 군집화가 잘 되는 듯하다.
- Model-Based Clustering을 통해 우리는 축소 데이터와 원본 데이터 모두 대략 4~6개의 군집화가 가장 이상적이라는 것을 알 수 있다. **즉, 총 20개로 구성된 설명변수들이 대략 4~6개로 구분될 수 있다는 결론을 유추해볼 수 있다.**

7.2 K-means Clustering

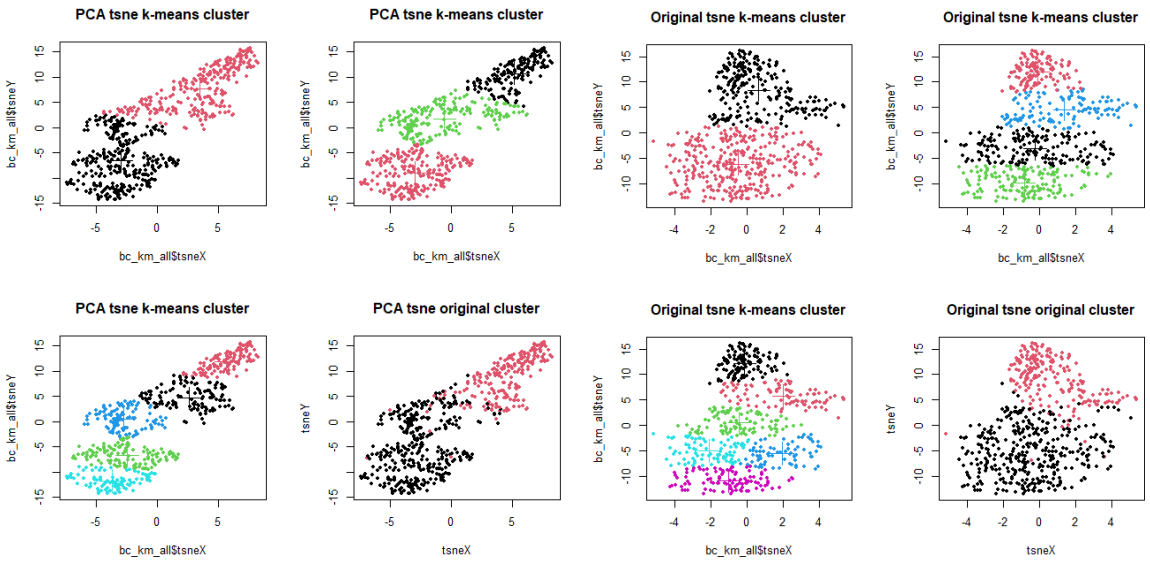


사진 14. PCA 데이터와 Original 데이터의 K-means Cluster

T-sne로 축소한 PCA데이터와 원본 데이터를 k-means로 군집화한 결과이다. 다음과 같은 정보를 확인할 수 있다.

- PCA 데이터의 경우, k=3일 때 가장 깔끔하게 구분된 듯해 보인다. 또한, 실제 Diagnosis 분포 (4번째 사진)와 유사하게 k=2일 때 클러스터링이 잘 구분되는 것을 알 수 있다.
- 원본 데이터의 경우, k=4, k=6 (model clustering 결과 optimal한 클러스터 개수) 로 클러스터링을 진행할 때 결과가 좋지 않은 것을 알 수 있다. 반면, k=2일 때 실제 Diagnosis 분포 (4번째 사진)과 유사하게 잘 군집화되는 것을 알 수 있다.
- 따라서, 우리는 k-means 클러스터링의 경우, 클러스터의 개수가 작을 때 군집화가 잘 일어난다는 것을 알 수 있다.

7.3 Hierarchical Clustering

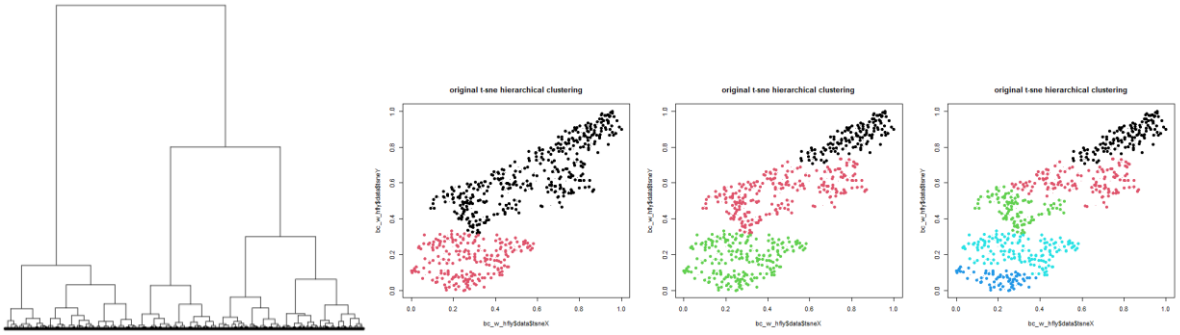


사진 15. PCA 데이터의 Hierarchical Clustering 결과

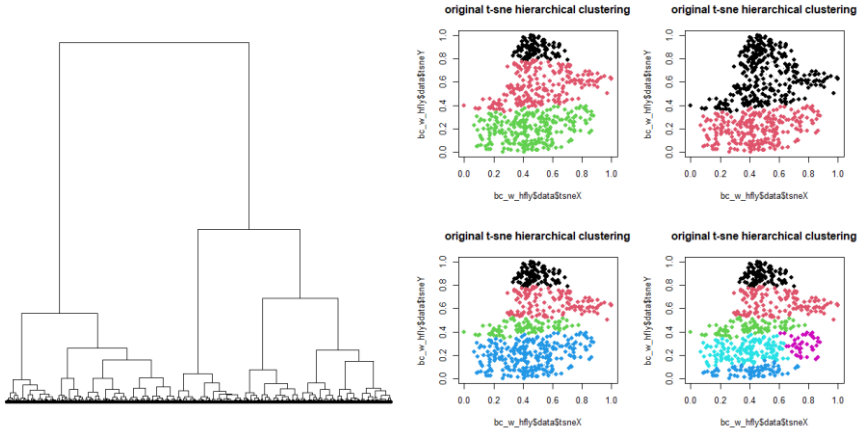


사진 16. Original Data의 Hierarchical Clustering 결과

Wards Linkage로 Dendrogram을 만든 결과이다. 다음과 같은 정보를 유추할 수 있다.

- 전반적으로 데이터가 잘 balanced 된 결과를 보여주며, PCA와 원본 데이터 모두 대략 3개의 cluster로 구성되면 좋을 것 같다는 판단을 할 수 있다.
- 실제 클러스터링 결과 k=3일 때 가장 이상적인 모습의 클러스터링 결과가 나온 것을 알 수 있다. 두 케이스 모두 k=3일때 가장 작은 머리부분(검은색)과 몸통(빨간색), 그리고 다리부분(초록색)이 잘 구별된 것을 알 수 있다.
- 반면, 실제 Diagnosis 상태인 k=2일때 결과는 실제 Diagnosis 상태와 상이한 것을 확인할 수 있다.

8. Classification

8.1 LDA

	B	M
B	351	25
M	6	186

	B	M
B	355	17
M	2	194

	오분류율
B	0.055
M	0.033

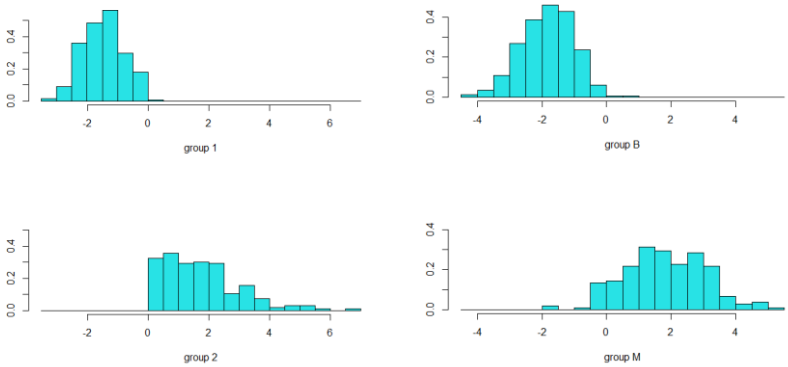


사진 17. PCA 데이터 및 원본(표준화)데이터의 confusion matrix, 오분류율, LDA Histogram

LDA 결과를 진행했을 때 PCA 데이터와 표준화된 원본 데이터 모두 오분류율이 매우 작은 것을 알 수 있다. Histogram을 통해 LDA에서 구분된 group의 plot을 볼 때에도 매우 뚜렷한 separation이 진행된 것을 알 수 있다. 이는 결과적으로 Diagnosis 두 그룹의 차이가 명확하며, classification하기 용이한 변수적 특성을 지닌 것을 의미한다.

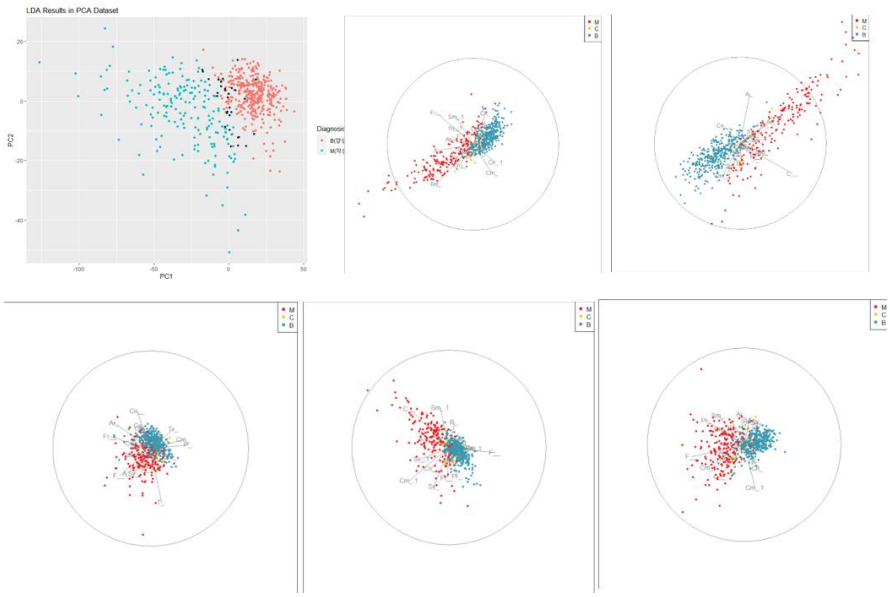


사진 18. LDA Visualization of PCA & Original(표준화) Dataset

LDA classification 결과를 2d, 3d plot으로 시각화한 결과이다. 첫 번째 2d plot은 PCA 데이터 기반의 LDA 결과이며, 오분류된 부분을 검은색 점으로 칠했다. 두 번째 3d plot은 오리지널 표준화 데이터 기반의 LDA 결과이며, 이 또한 오분류된 부분을 노란색 점으로 칠했다. 이를 통해 기본적으로 Diagnosis의 양상이 뚜렷한 것처럼, 대부분의 오분류율 역시 두 그룹(M, B)의 boundary 부근에 나타나는 것을 알 수 있다.

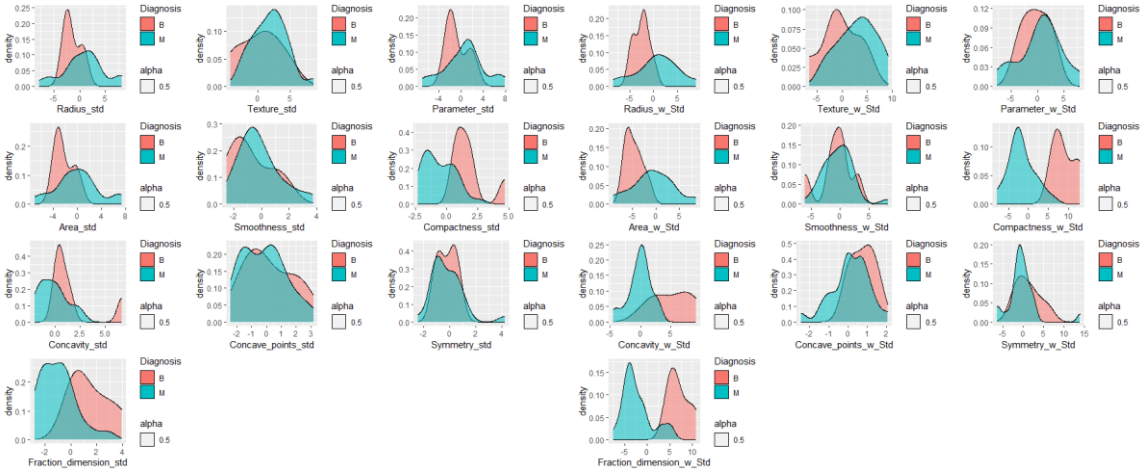


사진 14. 오분류된 원본데이터(표준화)의 Density Plot(오른쪽: standard 변수10개/왼쪽: worst cast standard 변수 10개)

<사진14>를 통해 알 수 있듯이, 기존 EDA를 통해 파악한 것과 다른 Density Plot이 그려지는 것을 알 수 있다. PCA 데이터의 오분류 관측치가 원본데이터의 오분류 관측치에 포괄되기 때문에 본 연구는 원본데이터의 오분류 값들에 대한 시각화를 통해 특성을 파악해 보았다.

- 대부분 M이 뒤쪽, B가 앞쪽에 나오는 EDA 결과와 달리, Compactness, Fraction Dimension, Concavity, Compactness 와 같은 변수들은 M(악성)이 앞쪽에 분포하는 경향을 보였다.
- EDA에서 Radius, Perimeter, Area, Compactness, Concavity 변수들의 양성(B) 분산이 작고 악성(M)의 분산이 컸던 것과 마찬가지의 양상을 보이지만, 이번에는 둘의 분포가 비슷한 범위에 있거나 반대로 M이 앞쪽에 분포(Concavity, Compactness) 경향을 보였다.
- Worst Case에서 분포의 범위는 일반적인 standard 변수와 동일하지만, M(악성)의 분산이 더 작게 나타나는 등(concavity, fraction dimension, compactness)의 현상을 보였다.
- 이러한 경향성은 일반적인 유방암 진단에 포괄되지 못하는 이상값 (특이 케이스)로

해석할 수 있으며, 후에 각 이상 관측치의 변수 간 연관성을 파악한다면 유방암 진단의 중요한 새로운 지표로 활용될 수 있을 것이라 생각된다.

9.2 Tree

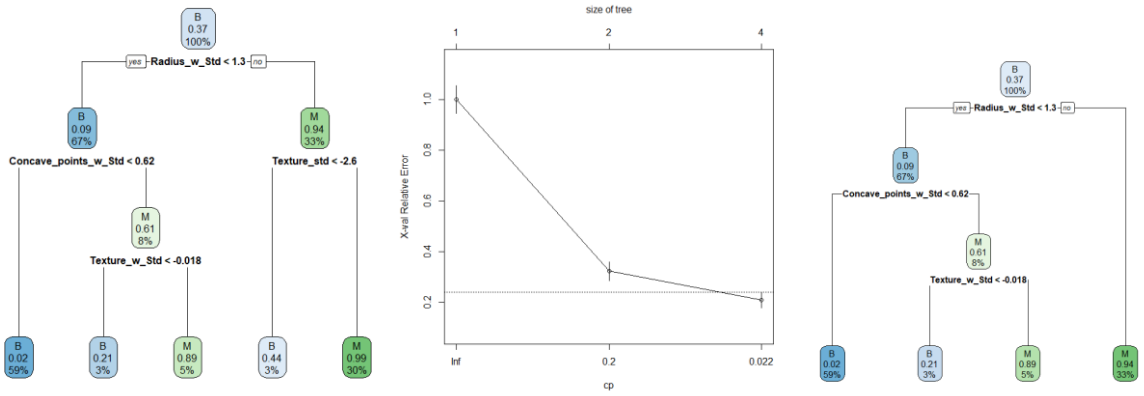


사진 19. 원본데이터(표준화)의 Tree 결과

첫 번째 plot은 최대 depth를 8, 최소 bucket 개수를 5, cp=0(No penalty)로 설정한 Tree의 결과이다. 그 결과 Radius의 Worst Case가 1.3이 넘을 때 94%의 확률로 M(악성) 종양을 구분할 수 있게 된다. 거기에 Texture가 -2.6 미만이면 거의 100%의 확률로 악성 종양을 진단할 수 있게 된다.

반면, 이상적인 페널티 값과 Tree의 개수 및 depth를 설정하기 위해 원본 데이터(표준화)의 70%를 Train Dataset으로 설정한 후, cp plot를 그려보았다. 그 결과, 대략 3~4개의 size tree를 갖는 것이 이상적이라는 판단을 내릴 수 있다. Cp=0.022로 설정한 후 plot을 그려본 결과, 세 번째 plot과 같은 결과를 얻을 수 있다. 첫 번째 그림과 유일한 차이점은 texture 변수의 유무이며, 그 외 나머지 변수들 (radius worst, concave points worst, texture worst)은 Diagnosis 진단에 중요한 변수임을 보여주고 있다.

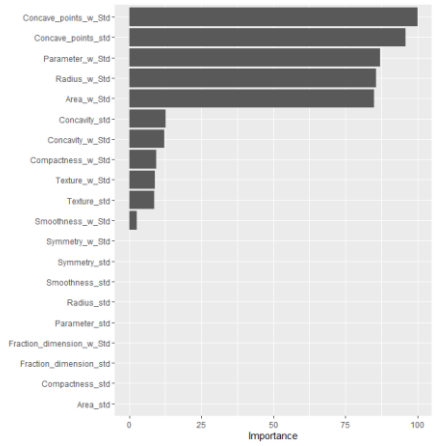


사진 20. Important Features 결과

마지막으로 변수 간 영향력을 파악하기 위해 각 변수를 기준으로 node가 나뉠 때 SSE가 줄어드는 기여도를 보여주는 vip plot을 그렸다. 그 결과, 앞선 tree에서 본 것처럼, radius worst, concave points worst 변수가 상위 5개 영향력 있는 변수로 선정된 것을 알 수 있다. 반면, texture 변수는 기여도가 상대적으로 낮게 나왔는데, 이는 실제 Tree 모형을 보았을 때에도 texture은 실질적인 Diagnosis를 구분 짓기보단 이미 구분된 group의 misclassify된 관측값을 filter하는 역할을 하는 것을 알 수 있다.

또한, 앞선 EDA 분석에서 Radius, Perimeter, Area 변수들의 분포에서 Diagnosis 구분이 뚜렷하게 나타났는데, Importance Feature plot에서도 이 세 변수들의 worst case들이 node 구분(즉, diagnosis 진단)에 중요한 변수로 작용한다는 것을 알 수 있다.

9. Summary

데이터 전처리(Page: 5~7)

본 데이터는 설명변수 간 스케일의 크기가 큰 데이터로, 표준화 작업을 통해 실제 분포를 잘 대변할 수 있는 표준화 방법을 선택했다. 실제 Diagnosis 변수에서 양성의 비율이 1.7배 많음에도 불구하고 악성 종양의 분산이 커서 plot화하였을 때 상대적으로 양성 비율이 적어 보이는 양상을 보였다. 이러한 문제를 해결하기 데이터를 시각화하는 작업에 있어 각 개별 변수의 분포를 파악하고 세심하게 구조를 살펴보기 위한 작업을 진행하였다.

EDA (page 7~9)

EDA 결과, Standard 변수 10개와 worst case standard 변수 10개의 분포가 거의 비슷하게

나타난다는 것을 알 수 있다. 변수들이 크게 두 가지 특성을 지녔는데, Radius, Perimeter, Area, Compactness, Concavity 변수들의 경우, Diagnosis 그룹의 분산 및 분포 범위가 상이하다는 특징을 지녔다. 반면, Texture, Smoothness, Symmetry, Fraction Dimension 변수들의 경우 또한 분포의 범위가 상이하지만, 그룹 간 분산이 비슷하다는 특징을 지닌다. 이러한 특징은 Radius, Perimeter, Area, Compactness, Concavity 변수들이 Diagnosis를 진단하는 데 더 유의미한 영향을 미칠 것이란 예측을 하게 만들었다.

PCA (Page: 9~10)

PCA 분석 결과 약 2개의 component로 데이터를 축소할 수 있었다. PC1과 PC2의 분포 또한 상이한 결과를 보였는데, PC1의 경우, 양성 그룹과 악성 그룹 간 분산과 분포의 범위가 다르다는 특징을 지녔다 (EDA의 Radius, Perimeter, Area, Compactness, Concavity 변수들과 유사). 반면, PC2의 경우, 두 그룹 간 분산은 다르지만 분포의 범위가 확연히 구별되지 않았다 (EDA의 Radius, Perimeter, Area, Compactness, Concavity 변수와 유사). 이처럼, EDA 결과에서 확인한 변수의 분포적 특성을 잘 반영한 모습을 통해 PCA 차원축소가 잘 되었다는 것을 알 수 있다

Clustering (Page: 11~13)

전반적으로 K-means> Hierarchical> Model Based 순으로 클러스터링이 잘 된 모습을 확인하였다. Model Based 클러스터링의 경우 많은 군집화 수를 요구했지만, 결과적으로 유의미하거나 좋은 plot이 만들어지지는 못했다. 반면, K-means나 Hierarchical 클러스터링의 경우, 의 개수가 적을 때 실제 Diagnosis 분포와 유사한 군집화가 진행된 것을 알 수 있다. PCA와 Original 변수들의 T-Sne 데이터로 클러스터링을 진행했는데, PCA T-SNE 데이터의 경우, k=3일 때 데이터가 잘 쪼개지는 듯한 양상을 보인다.

Classification (Page: 14~17)

앞서 EDA 결과에서 봤듯이 본 데이터는 Diagnosis 간 차이가 명확하게 나타나기 때문에 기본적으로 Classification의 어려움이 적었다. LDA의 경우, PCA 와 원본데이터(표준화) 모두 오분류율이 매우 적었다. 오분류된 LDA 데이터만 모아서 시각화하여 구조를 살펴보았을 때, EDA 결과와 다른 분포적 양상을 보였으며, 오분류 변수들 중 비슷한 구조를 보이는 변수들

이 존재했다. Tree의 경우, Radius worst case가 악성종양을 결정하는 중요한 변수로 작용했으며, concave point worst case는 양성종양을 판단하는 중요한 변수로 확인되었다. Importance plot을 통해 두 변수 매우 중요하다는 것을 확인할 수 있었으며, EDA에서 Diagnosis 그룹 간 상이한 분포를 보였던 concave point, perimeter worst case, area worst case도 중요 변수로 선정되었다. 이처럼, 유방암 종양 판단에 있어 종양의 크기(radius, perimeter, area)와 오목한 점이 매우 중요한 영향을 미친다는 것을 알 수 있다.