유방암을 진단하기 위한 데이터 분석

Breast Cancer Wisconsin (Diagnostic) Data Set Predict whether the tumor is benign or malignant.

1. 김태형(19931011)
2. taehyeong826@naver.com
3. 김태형 (19931011) / 최준석 (19940404)

우리 팀은 통계학 전공자와 생물·화학 전공자로 이루어진 듀오이며, 전공을 잘 살리기 위해 네이버 측에서 제시한 데이터 중 유방 종양의 특성이 담긴 데이터를 골라 유방암을 분류하기로 했다. 이 데이터는 유방 종양 세포를 FNA (아주 얇은 빨대로 특정 세포를 뽑아내 염색해서 현미경으로 관찰하는 방법)을 통해 촬영한 것인데, 종양 세포들을 각각 10가지의 특성으로 서술했으며 우리는 이 중 어떤 특성이 유방암과 가장 관련이 있는 지 알아볼 것이다.

종양은 비정상적인 세포 덩어리인데 여러 요인에 의해 생길 수 있다. 양성 종양 (benign tumor)과 악성 종양 (malignant tumor = cancer)으로 구분하며, 전자는 인체에 해가 되지 않는 경우도 있어 무조건 치료하는 것은 아니지만 후자는 매우 해롭고 증식 속도와 전이 속도가 빠르므로 치료해야만 한다. 다행히도 유방암 환자의 생존율은 다른 암 환자에 비해 높은데 기존의 유방암 환자들의 데이터로 미래의 환자들에게 더욱 도움이 될 수 있음을 기대한다.

분석에는 Kaggle에서 제공하는 'Wisconsin Diagnostic Breast Cancer Data Set'의 위스콘신 유방암 데이터를 이용하였다.

- 출처: UC Irvine Machine Learning Repository (위스콘신 대학교에서 제공한 유방암 진단결과 데이터)
  (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)
- 32개의 변수와 569(양성 357, 악성 212)개의 레코드를 갖는 진단결과 데이터
- 양성 유방 종양과 악성 유방 종양의 특성을 10가지 변수로 분류한 데이터를 자동 분류하는 모델 구축하기.

*** 변수 이름과 설명 (Columns = 32)**

1. id: 환자 식별 번호
2. diagnosis: 악성 여부 (M = malignant(악성), B = benign(양성))

10가지의 세포에 대한 정보들

3. radius: 세포 중앙에서 둘레의 점까지 거리 (반경)
4. texture: 질감 (Gray-Scale 값들의 표준편차)
5. perimeter: 둘레
6. area: 면적
7. smoothness: 매끄러움 (반경 길이의 국소적 변화)
8. compactness: 밀집도 (둘레^2/면적 -1)
9. concavity: 오목한 면 (윤곽의 오목한 부분의 정도)
10. concave points: 오목면의 수
11. symmetry: 대칭 여부
12. fractal dimension: 프랙탈 차원 (해안선근사 -1)
    ➔ 종양 세포의 둘레가 x배가 되었을 때 유사 세포의 수가 n배가 된다면 그 세포의 프랙탈 차원은 $\log_x n$. 종양 세포의 악성 정도나 진행 정도에 따라 클수록 프랙탈 차원이 낮아지는 경향이 있다.

_mean: 3 ~ 12 번까지는 평균값

_se: 13 ~ 22 번까지는 표준오차

_worst: 23 ~ 32 번까지는 제일 큰 3개의 값의 평균값

어떤 변수가 종양 세포의 양성, 악성 여부를 가릴 수 있는 좋은 변수인가?

R을 활용하여 데이터 분석을 하였으며 결측치는 없었다. 악성 세포와 양성 세포의 성질은 어떻게 다를까? 212개의 악성 세포와 357개의 양성 세포의 차이를 알아보기 위하여 변수들의 특성을 분류해보았다.

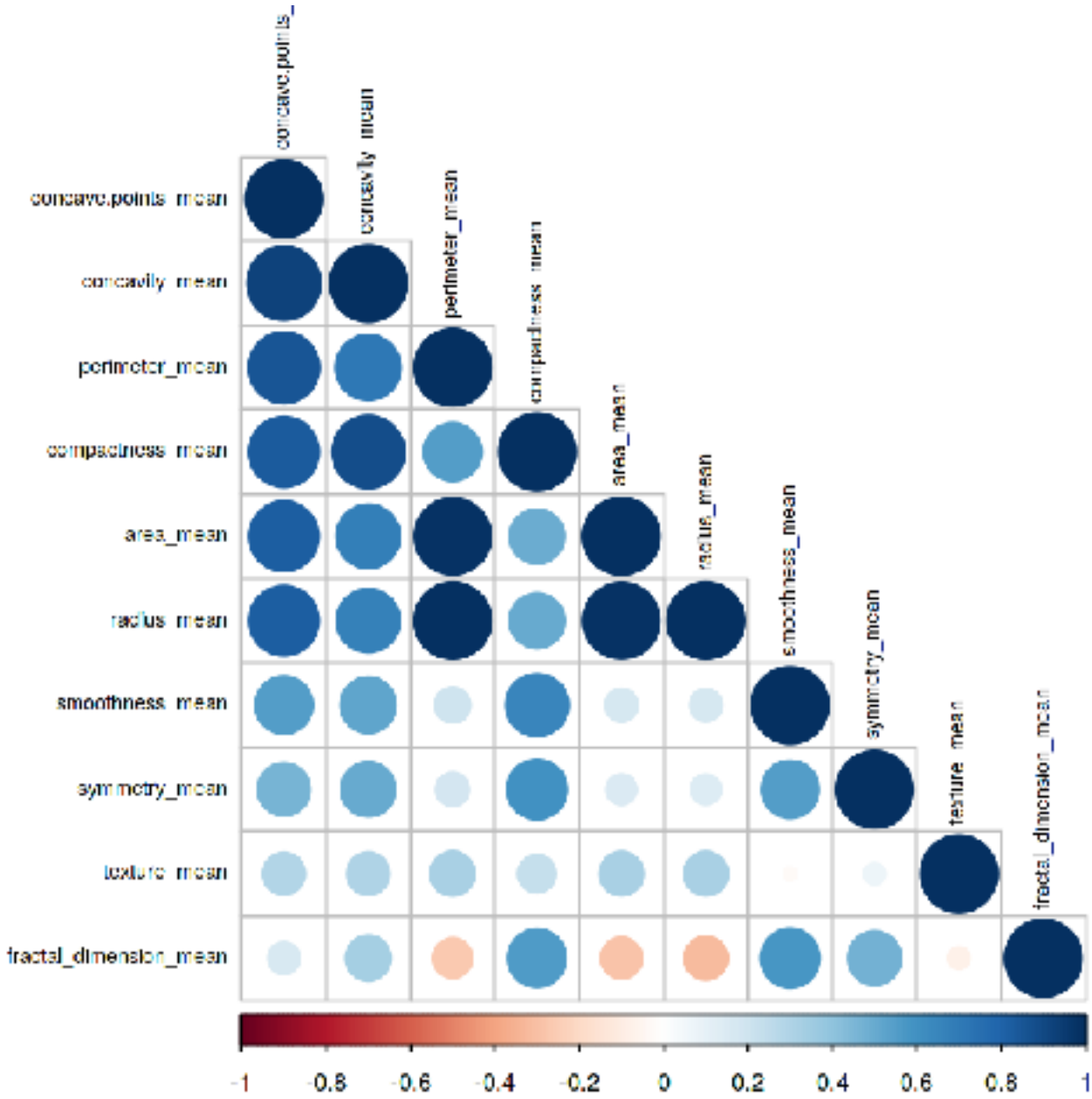
악성 세포와 양성 세포의 구분 기준은 어떤 것이 좋을까? 10개의 특성의 대푯값으로 평균을 이용하여 악성 세포와 양성 세포의 차이를 변수 별로 비교하였다. diagnosis를 기준으로 mean 변수들의 데이터 대푯값을 비교하여 보았다.

		Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
양성 B 357	Min	6.981	9.71	43.79	143.5	0.05263	0.01938	0.00000	0.00000	0.1060	0.05185
	Median	12.200	17.39	78.18	458.4	0.09076	0.07529	0.03709	0.02344	0.1714	0.06154
	Mean	12.147	17.91	78.08	462.8	0.09248	0.08008	0.04606	0.02572	0.1742	0.06287
	Max	17.850	33.81	114.60	992.1	0.16340	0.22390	0.41080	0.08534	0.2743	0.09575
악성 M 212	Min	10.95	10.38	71.90	361.6	0.07371	0.04605	0.02398	0.02031	0.1308	0.04996
	Median	17.32	21.46	114.20	932.0	0.10220	0.13235	0.15135	0.08628	0.1899	0.06157
	Mean	17.46	21.60	115.37	978.4	0.10290	0.14519	0.16077	0.08799	0.1929	0.06268
	Max	28.11	39.28	188.50	2501.0	0.14470	0.34540	0.42680	0.20120	0.3040	0.09744

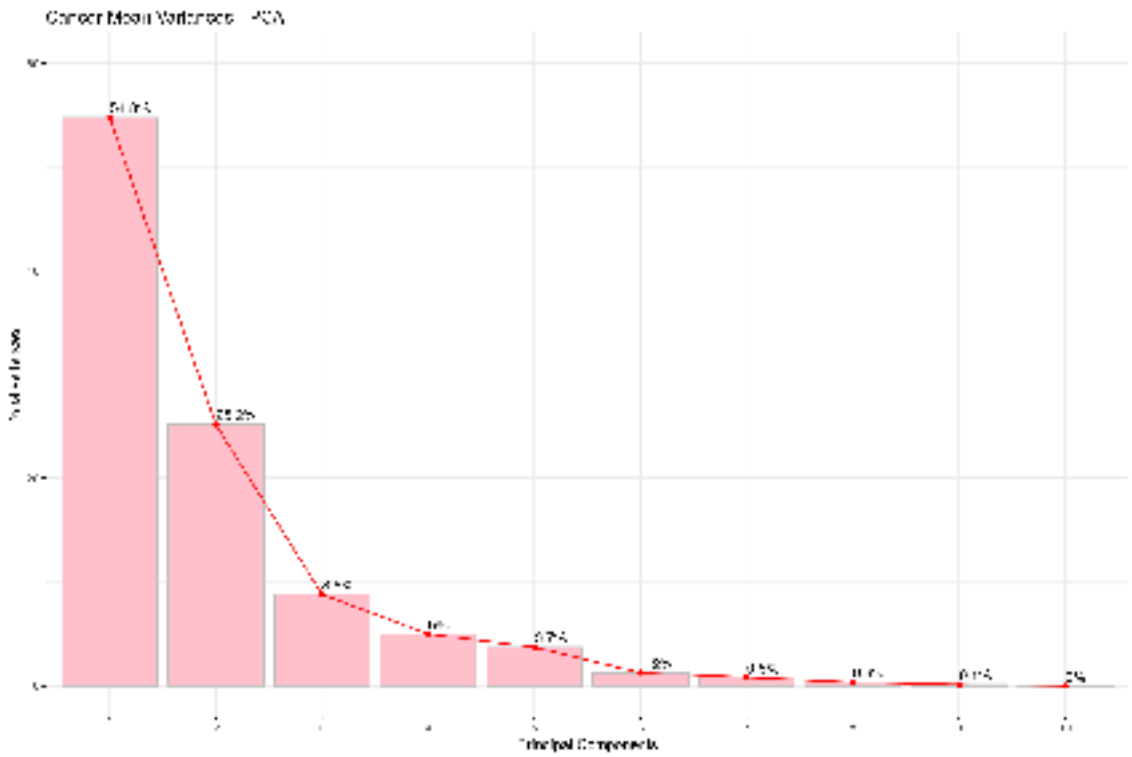
전체적으로 악성 유방암 데이터의 수치가 양성 유방암 데이터의 수치보다 값이 큰 것을 알 수 있다. (Smoothness, Symmetry, Fractal dimension 제외). 허나 이 표를 보고 수치가 별 차이 없는 변수들이 영향을 미치지 않는다고 단정지을 수 없다. 정규화가 되지 않았고 변수들 간의 상관관계로 인한 다중 공선성을 고려하지 못했기 때문이다. 그래서 변수들 간의 상관관계를 알아보았다.

변수들의 상관관계를 알아보고 주성분 분석을 실시.

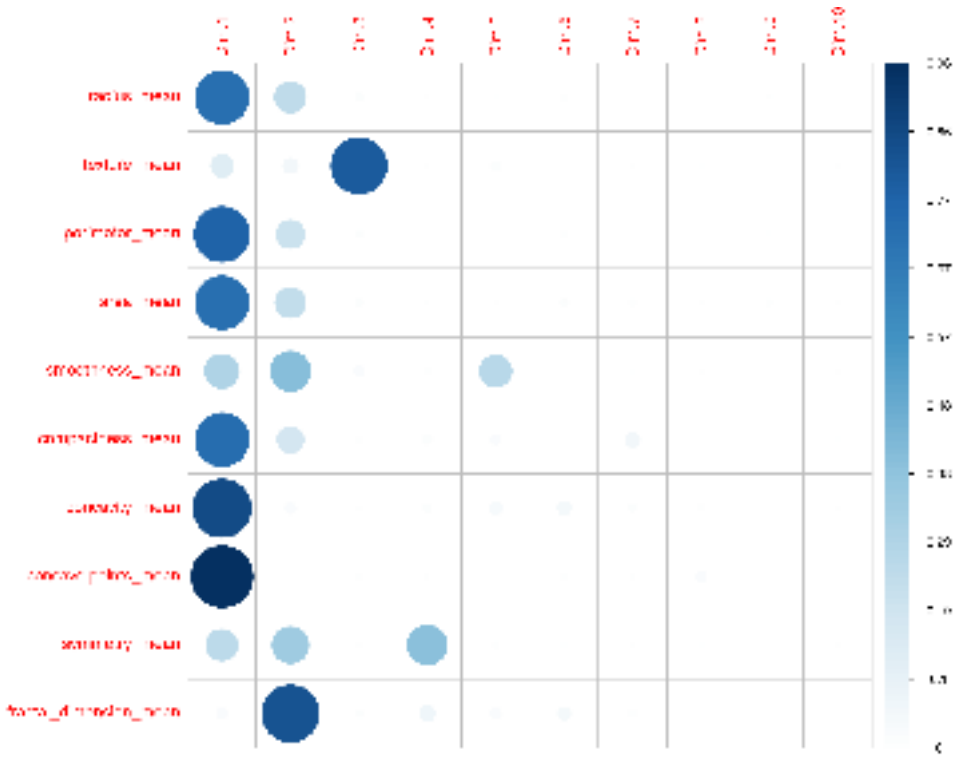
Mean 값을 가진 변수들의 상관관계를 구하고 그림으로 나타냈다.



Radius, Perimeter, Area 변수는 서로서로 상관관계가 매우 높은 것을 알 수 있다. 상식적으로 당연하다. Texture 변수는 다른 변수와의 상관관계가 대부분 낮았다. 위 그림처럼 주성분 순서로 변수 간의 상관관계를 분석한 결과 전체적으로 높은 상관관계를 가진 변수들이 많으므로 다중 공선성을 제거하고 효율적인 변수들로 요약하기 위해 주성분 분석을 하였다. cumulative proportion을 85% 이상으로 보이는 성분까지를 주성분으로 본다. 결과를 보면 Comp 3에서 85%를 처음 초과했으므로 주성분은 1 ~ 3인 것이다.



주성분 분석을 통하여 Screeplot을 그려보았다. 주성분이 3개일 때 약 88%를 설명할 수 있었다.



주성분 변수1은 radius, perimeter, area, compactness, concavity, concave points를 설명하는 변수들로 요약이 되었고 주성분 2는 fractal dimension, 그리고 주성분 3은 texture을 설명하고 있다. 주성분 4부터는 큰 의미를 가진 변수들이 없다고 생각하여 분석에 넣지 않았다.

*** 유방암 진단 모형**

진단결과 분류를 하는 것을 목적으로 KNN 알고리즘과 SVM을 사용하였다. 두 가지의 모형을 사용한 이유는 분석 데이터가 크지 않고 상대되는 관점으로 분류를 하여 결과를 설명할 때 보완할 부분이 있기 때문이다.

분류 모형을 만들고 test하기 위해 전체 569개 데이터를 398개 train 데이터(70%)와 171개 test 데이터(30%)로 나누었다.

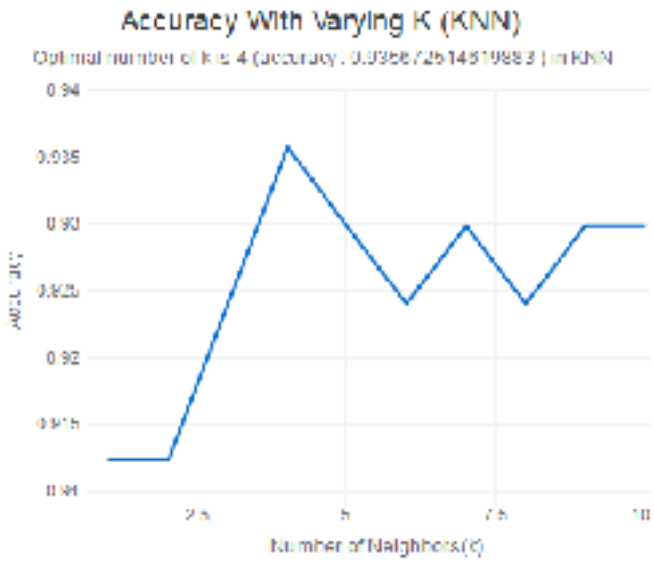
전체 데이터 셋이 잘 나뉘었는지 Train data set과 Test data set의 diagnosis의 비율을 알아보자.

	Benign	Malignant
Train	0.6281507	0.3718593
Test	0.625731	0.374269

B는 63%, M은 37%로 train과 test data set을 잘 나누었다.

1) KNN 알고리즘

Target variable은 Diagnosis(진단결과)이다.




```
Confusion Matrix and Statistics

          Reference
Prediction  Benign Malignant
  Benign       105         9
  Malignant      2        55

               Accuracy : 0.9357
                 95% CI : (0.8878, 0.9675)
    No Information Rate : 0.6257
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8596
 Mcnemar's Test P-Value : 0.07044

            Sensitivity : 0.9813
            Specificity : 0.8594
         Pos Pred Value : 0.9211
         Neg Pred Value : 0.9649
             Prevalence : 0.6257
         Detection Rate : 0.6140
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.9203
```

Mean값을 가진 변수들을 갖고 k값을 구해보았다.

K의 값은 4로 지정한다.

양성이라고 진단했을 때 음성인 결과 = 9개

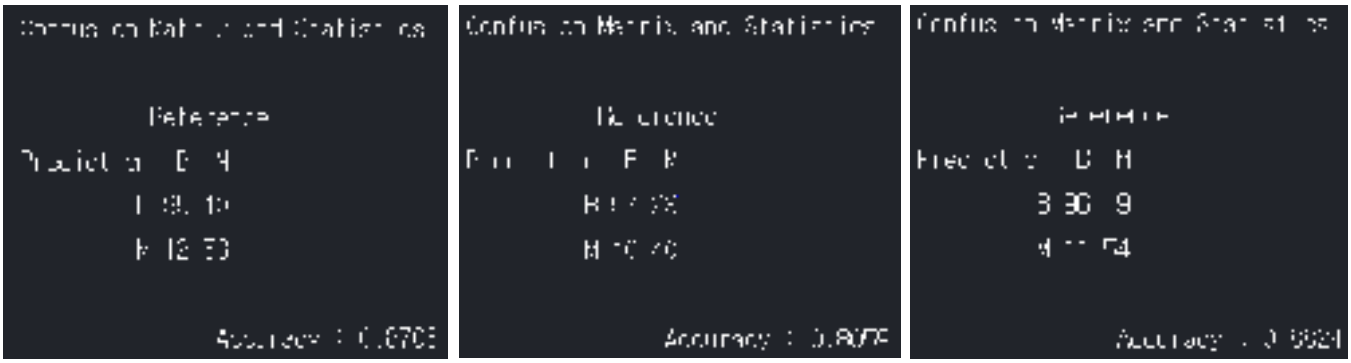
음성이라고 진단했을 때 양성인 결과 = 2개 로 160/171 = 약 94%의 예측 정확도를 보여준다.

2) SVM

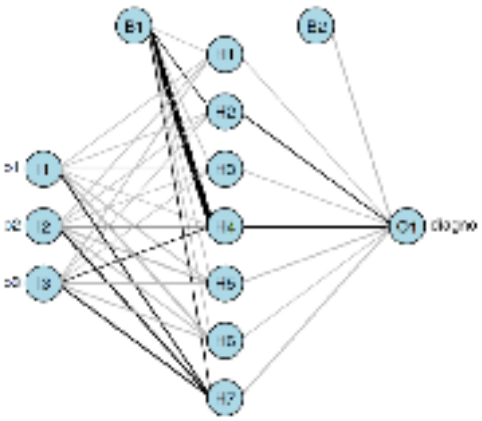
KNN과 더불어 가장 유명한 기계학습 모델인 SVM으로도 분류해 본다. KNN과 마찬가지로 주성분 3개로 SVM을 이용한다.

데이터 전처리에 유용한 caret 패키지로 데이터를 7:3의 비율로 훈련용 데이터과 테스트용 데이터로 나누었다. 그리고 SVM 훈련은 e1071 패키지의 tune 함수로 했으며, tune 함수로 SVM의 파라미터 gamma와 cost의 최적의 값을 찾은 후 테스트를 했다.

왼쪽부터 커널이 linear, polynomial, gaussian일 때의 결과이며 정확도는 당연히 새로운 시행마다 다르다.



3) Neural Net



이번에는 주성분들을 신경망에 탑재해 분류해 볼 것이며 nnet 패키지의 nnet 함수를 이용한다. 마찬가지로 훈련용 데이터와 테스트용 데이터를 7:3의 비율로 해서 신경망을 학습했다. 히든 노드는 7개 단일층이며 반복 회수는 777회, 오버피팅을 방지하기 위한 decay 파라미터 값은 $7e^{-7}$, 초기 랜덤 가중치의 값은 0.7으로 해서 학습용 데이터를 신경망에 학습시킨 후에 테스트용 데이터를 분류했고 정확도는 87.06%가 나왔다. 그리고 devtools 패키지의 plot.nnet 함수로 신경망을 왼쪽과 같이 표시해 보았다. 생각 외로 위의 두 기법들보다 코드가 간단해서 쉽게 할 수 있었는데 최적의 파라미터 값을 찾는 방법은 완벽히 생각하지 못해서 이는 후에 있을 오프라인 교육으로 넘기겠다.

*** 끝을 맺으며**

유방암 세포를 진단하기 위한 분석을 하면서 악성과 양성의 차이를 알아보기 위해 변수의 특징을 살펴보고 분류하는 모형을 만들어보았다. 데이터를 유의미하게 분석하기 위하여 주성분분석을 활용하고 여러가지 모델을 만들었다. KNN 모형의 분류 확률은 매우 높았으나 주성분 변수가 아닌 기존 변수로만 분석을 하여 아쉬움이 남았다. 하지만 데이터가 소량인 점에 비추어서 굳이 주성분을 활용하지 않아도 괜찮다고 판단하였고 예측 확률 또한 높게 나와서 적합한 모형이었다고 생각한다. SVM 모형 같은 경우는 linear와 gaussian이 좋은 결과를 나타냈다. 이 모형은 주성분 변수를 활용하였다. Neural Net 모형도 구현하였다. 모형은 모두 85% 이상의 좋은 분류 결과를 보였다. 주어진 데이터 셋이 정확했다고 생각한다. 부족한 분석과 설명이었지만 앞으로 남은 시간동안 edwith 강의를 잘 참고하여 더 좋은 데이터 분석을 하고 싶다.