Convergence Informatics, Final Exam

Your ID#: [다시 풀어보는]

Your Name: [기말고사_조수현]

[Instruction] 안내

(1) Please read and answer the questions carefully. Write your answers in Korean or English.
(문제를 잘 읽고 신중하게 답변하십시오. 영어 또는 한국어로 답안을 작성하세요).

(2) Please compress (.zip) and submit this report (.docx), and the entire R-code (.R file) you wrote into the BlackBoard.
(이 시험지와 답변에 사용된 R코드를 압축하여 블랙보드에 지정된 시간까지 반드시 제출하십시오.)
Final Exam: 10:00 AM ~ 11:30 AM (90 Min.)
*I highly recommend that you start submitting to the Blackboard at 11:25AM, at least 5 minutes before the end. If you do not submit by the given time, you will receive 0 points. (시험 끝나기 5분전인 11시 25분에는 블랙보드에 제출을 시작하길 권장합니다. 11시 30분에는 제출할 수 없도록 닫히며, 제출을 못했을 시 당연 0점 처리됩니다.)

(3) The R-code should be executable when the TA runs. The submitted compressed file (.zip) must be named Final_YourID_Yourname.zip.
(R 코드는 TA가 돌렸을 시 깔끔하게 돌아가야 하며, 제출될 압축파일은 반드시 Final_YourID_Yourname.zip 로 명명하여 제출하십시오.)

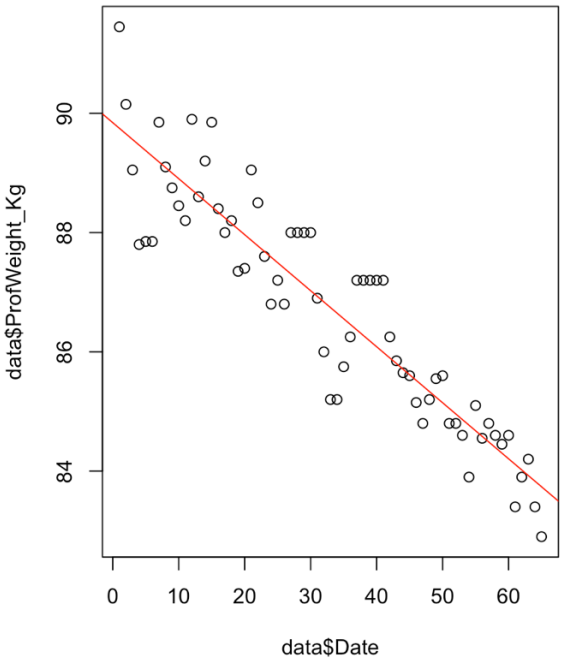
(4) There is a partial score. Even if you can't resolve it completely, I hope you can try it as far as you can.
(부분 점수가 있으므로, 완벽하게 해결하지 못하더라도 할 수 있는 만큼 시도해 보길 바랍니다.)

[Q1 – Q4] is based on "*DataQ1.xlsx*" data.

[Q1] This is a record of weight loss during the summer vacation of 2021. **Let's perform predictive modeling of professor and student weight changes, respectively**. What is the best model established from professor and student data, respectively? Also, through inferred statistical comparison, **which of the two had the faster weight loss effect**? (*Must be interpreted through **comparison of estimated statistics**) **(10 Points).**

(Kor: 2021년 여름방학 동안의 체중 감량 기록에 관한 자료 입니다. 교수와 학생의 체중 변화에 대한 예측 모델링을 각각 수행해 보십시오. 교수 데이터와 학생 데이터로부터 최적의 모델은 무엇입니까? 또한, 추론된 통계적 비교를 통해, 둘 중 어느 것이 더 빠른 체중 감량 효과를 보였나요? (*추정 통계와의 비교를 통해 해석되야 함)

[Established model for Professor]



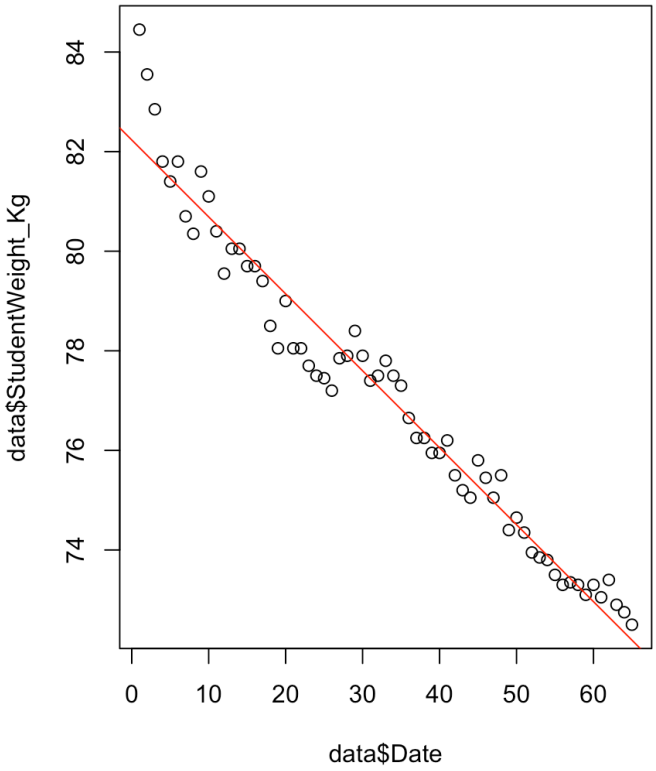

```
Call:
lm(formula = ProfWeight_Kg ~ Date, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66740 -0.51135 -0.03096  0.49471  1.70077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.843173   0.192947  465.64   <2e-16 ***
Date        -0.093942   0.005083  -18.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7688 on 63 degrees of freedom
Multiple R-squared:  0.8443,    Adjusted R-squared:  0.8418
F-statistic: 341.6 on 1 and 63 DF,  p-value: < 2.2e-16
```

[Established model for Student]




```
Residuals:
     Min       1Q   Median       3Q      Max
-1.24798 -0.25746 -0.06134  0.32422  2.37056

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.233966   0.155752  527.98   <2e-16 ***
Date        -0.154526   0.004103  -37.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6206 on 63 degrees of freedom
Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9568
F-statistic:  1418 on 1 and 63 DF,  p-value: < 2.2e-16
```

[Interpretation of estimated models]

단순선형회귀 모델을 통해 교수와 학생에 대한 모델을 만들었다. 교수의 모델에서 b1, b0는 각각 89.843173, -0.093942이고, 학생의 b1, b0는 각각 82.233966, -0.154526이다. 시간이 지날수록 기울기의 절댓값이 큰 학생에게서 더 빠른 체중감량 효과를 보일 것으로 추론할 수 있다. 이는 plot한 그래프를 통해서도 확인할 수 있다.

[Q2] If they continued to diet from that time until 2021-11-22, **what is their predicted weight**? **(10 Points)**

(Kor: 만약 그들이 2021-11-22까지 다이어트를 계속했다면 그들의 예상 체중은 얼마나 될까요?)

다이어트 시작: 2021-06-28, 다이어트 끝 2021-11-22 -> 148일이 되는 날 끝

[Predicted Professor's weight in 2021-11-22]

```
  PredictedWeight Date
1        75.93971  148
```

[Predicted Student's weight in 2021-11-22]

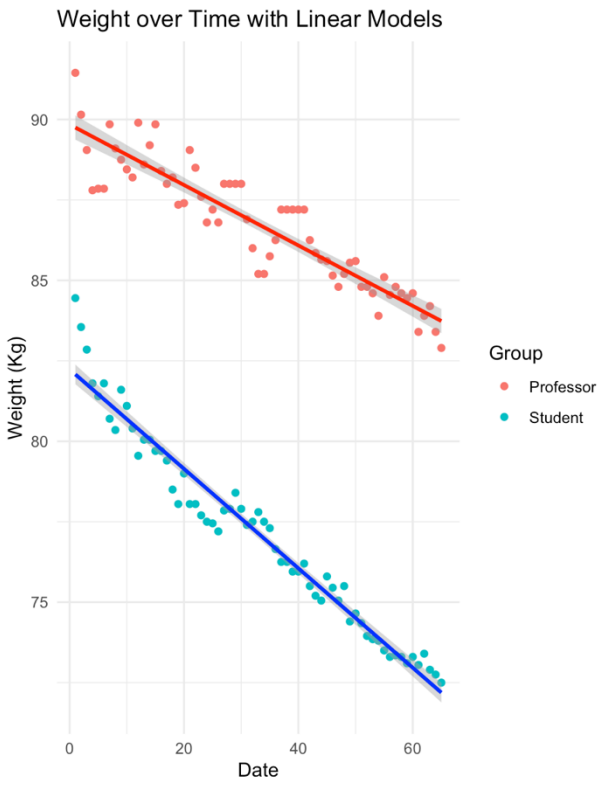
```
  PredictedWeight Date
1        59.36415  148
```

[Q3] Let's Visualize the given data through Scatter plots and by representing the best models estimated by professor and students, respectively, in different colors. **(10 Points)**

(*Hint: https://rafalab.github.io/dsbook/regression.html)

(Kor: 주어진 데이터를 산점도(Scatter Plot)를 통해 시각화하고, 교수와 학생이 각각 추정한 최고의 모델을 서로 다른 색상으로 표현하십시오.)

[Scatter plot with fitted models]



[Q4] According to the given model, if they had dieted until today (2023-12-18), what would each of their expected weights be as predicted by the model? Through these results, clearly interpret the limitations of the model you have built. **(10 Points)**

(Kor: 주어진 모델에 따르면, 그들이 만약 오늘 (2023-12-18) 까지 다이어트를 수행했다면 모델로부터 예측된 예상 체중은 각각 몇인가요? 본 결과를 통해 본인이 구축한 모형의 한계점에 대해서 명확하게 해석하십시오.)

다이어트 시작: 2021-06-28, 다이어트 끝 2023-12-18 -> 904일이 되는 날 끝

[Predicted Professor's weight in 2023-12-18]

```
  PredictedWeight Date
1        4.919327  904
```

[Predicted Student's weight in 2023-12-18]

```
  PredictedWeight Date
1       -57.45734  904
```

[Interpretation of the model] **within 1 page
(*Tip: Based on these results, consider the potential pitfalls of simple linear regression.)

이 문제를 해결하기 위해서 단순선형회귀 모델을 적용하였다. 이 모델에서 사용되는 변수는 날짜 뿐이다. 따라서 날짜를 제외한 다른 요인들은 반영되지 않는다는 한계점이 있다. 실제로, 사람의 몸무게는 감소량은 선형관계를 보이지 않는다. 사람의 몸무게가 계속해서 줄어들지 않기 때문이다. 하지만, 이 모델에서는 단순히 날짜에 따라 체중을 감소시키고 있으므로, 사람의 몸무게가 위와 같이 극단적으로 작게 나오거나 음수가 나오는 것이다.

(Q5) Load "*DataQ5_DNA_Database.txt*" and "*DataQ5_DNA_Query.txt*" files (tab-seperete files). As you can see from the link (https://edition.cnn.com/2019/10/04/asia/south-korean-serial-murder-confessed-intl-hnk-scli/index.html) and (https://en.wikipedia.org/wiki/Lee_Choon-jae) presented in 2019, the informatics technology we have learned is being applied in various fields.

(Kor: "DataQ5_DNA_Database.txt" 및 "DataQ5_DNA_Query.txt" (탭으로 구분된 형식) 파일을 로드하십시오. 화성 연쇄 살인사건 (상단 웹 링크 참조)에서 봤듯이, 우리가 배운 정보학 기술은 다양한 분야에 적용되고 있습니다.)

DataQ5_DNA_Database.csv: the information of 2,178 DNA criminals living in Seoul is databased.

DataQ5_DNA_Query.csv: this file contains the targeted criminal's DNA information.

(Kor: DataQ5_DNA_Database.csv: 서울에 거주하는 2,178명의 DNA범죄자들의 정보에 대한 데이터베이스입니다.

DataQ5_DNA_Query.csv: 우리가 찾고자 하는 범죄자의 DNA 정보를 담고 있는 파일입니다.)

Please apply the skills we learned this semester to identify who the serial killer is. All the given data has already been pre-processed (All values are real number). DNA information is characterized by very little information that changes over several centuries, meaning that having the same value across all features is more likely to be a target.

(Kor: 이번 학기에 배운 기술을 적용하여 연쇄 살인범이 누구인지 신원을 밝혀보십시오. DNA 정보는 몇 세기가 지나도 변화하는 정보가 극히 적다는 특징이 있으며, 이는 모든 특징에 걸쳐 동일한 값을 가질수록 표적이 될 가능성이 높음을 의미합니다.)

[Q5-1: What analysis methodology did you use? Why did you use it?] (10 Points)
(Kor: 문제해결을 위해 어떤 방법론을 사용했나요? 왜 사용했나요?)

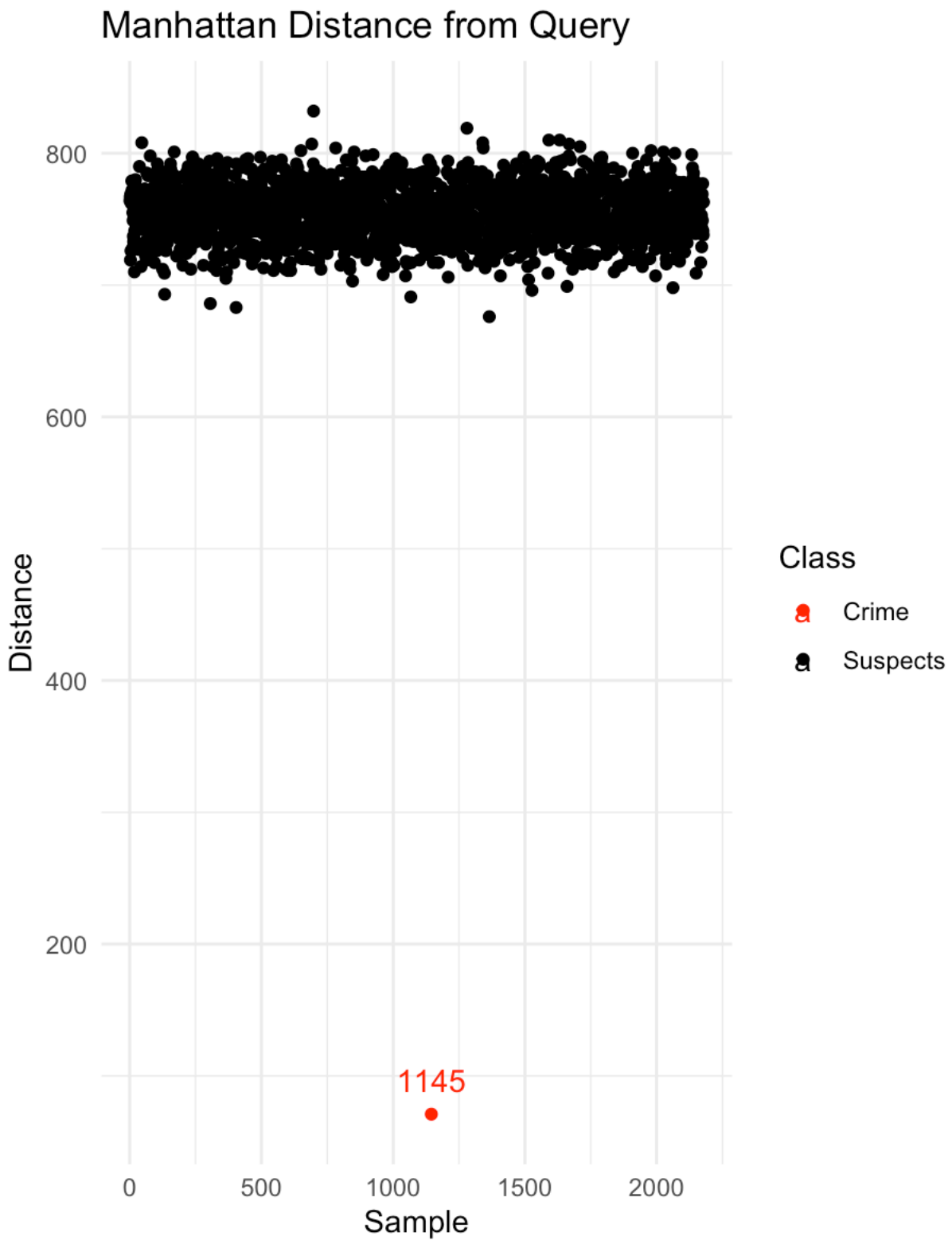
주어진 db와 쿼리를 살펴보면, 클래스가 존재하지 않으므로, 지도학습을 사용할 수 없다는 것을 알 수 있다. 문제를 해결하기 위해서는 쿼리가 데이터베이스 안에 있는 데이터와 가장 유사한 것을 찾아야 한다. 따라서, 쿼리와 데이터베이스 사이의 거리행렬을 만들고, 최소거리가 되는 데이터를 연쇄 살인범이라고 추정할 수 있다.

[Q5-2: Summary of Results and Conclusion] (10 Points)
(Kor: 본인이 수행한 모델링의 개요를 설명하고 결론을 내리십시오)

데이터를 살펴보면 유전자의 개수를 나타내고 이는 0, 1, 2의 형태로 표현된다. Manhattan distance를 사용하여 최소거리를 찾는 방법을 진행해보면 db의 1145번 index의 범죄자가 가장 유사한 것으로 확인할 수 있다. 따라서, db의 1145번째 범죄자를 조사한다면 이 사람이 연쇄 살인범일 가능성이 가장 높다.

[Q5-3: Visualization evidence to support your results] (10 Points)

(Kor: 결과를 증명할 수 있는 훌륭한 시각화 결과를 보이십시오)



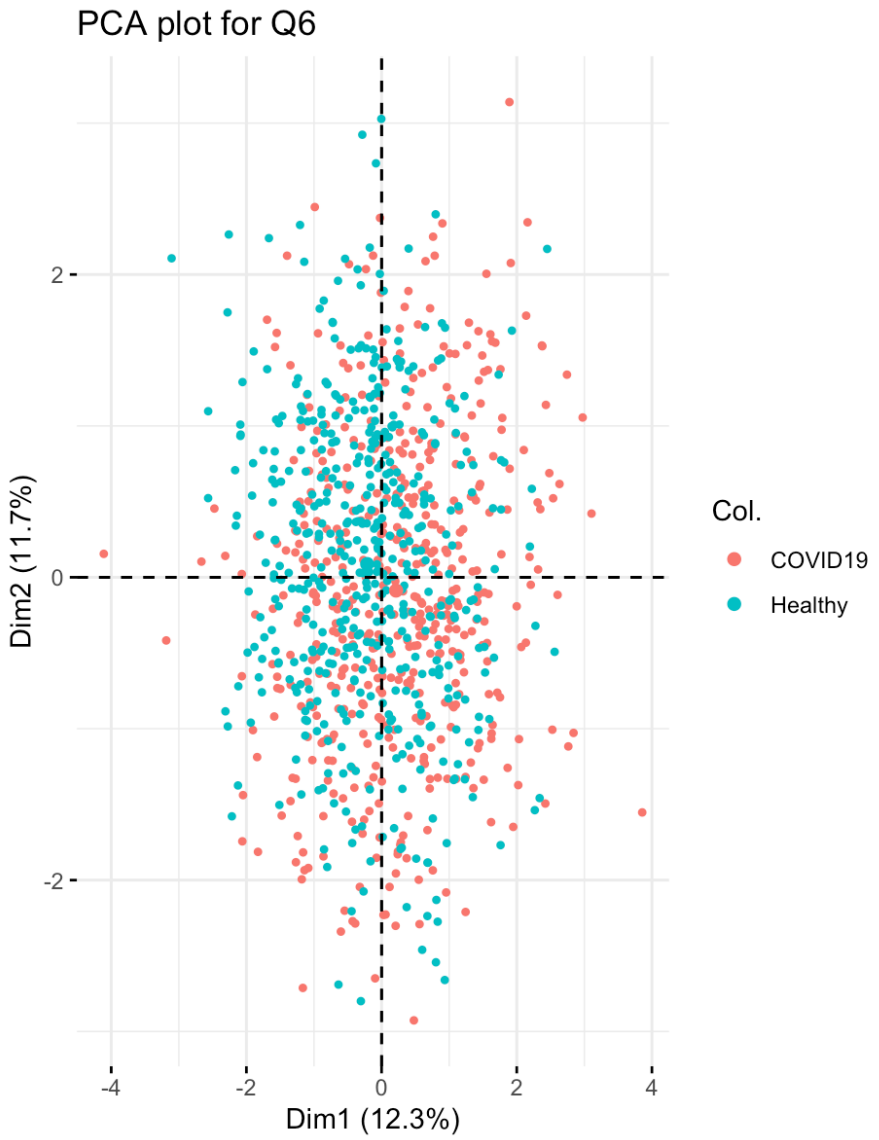
(Q6) Load "*DataQ6.txt*" file (tab-separate file). It is data collected about **10 feature values collected from a total of 999 people** and whether they are positive for COVID-19. Based on this data, construct the best model for predicting COVID-19 based on 10 features.

(Kor: "DataQ6.txt" 자료를 불러오십시오. 이 자료는 999명의 사람에 대해서 10가지 특징값과 COVID-19 양성 여부에 대한 정보가 수집되어 있습니다. 이 자료를 바탕으로 COVID-19를 예측할 수 있는 최적의 모형을 설계해보십시오.)

[Q6-1: Before modeling, **perform a visualization of 999 individuals at once in the given data to determine if COVID-19 patients can be distinguished from normal people based on given data**. (**10 Points**)

(***Hint: Since the number of features are 10 dimensions, we have to process it through the method we learned, and it is better to identify the disease information in different colors.)

(Kor: 모델링전 주어진 자료의 정보를 바탕으로 999명에 대해 동시에 시각화 하여, 일반인과 구분되는 특징이 있을지에 대해서 조사해보십시오. 힌트: 10차원의 자료이므로, 우리가 배운 별도의 작업을 선행해야 하며, 시각화 당시 질병정보를 다른 색으로 표현하는 것이 좋습니다.)



주어진 데이터를 PCA 한 결과, 환자와 정상인 사이의 차이점을 확인할 수 없었다.

[Q6-2, Q6-3]

아래 문제를 해결하기 위해 지도학습 모델 8가지[lda, rpart, rf, svmLinear, nnet, knn, gbm, glm]을 사용했다. 각각의 모델을 통해 정확도와 민감도를 구한 결과는 아래와 같다.

```
[1] "lda"
  parameter  Accuracy Sensitivity AccuracySD SensitivitySD
1      none 0.5906061       0.584 0.03046723     0.1023284
[1] "rpart"
          cp  Accuracy Sensitivity AccuracySD SensitivitySD
1 0.06412826 0.8428384       0.740 0.03803806    0.08219219
2 0.30460922 0.7528283       0.534 0.08079343    0.18620180
3 0.35871743 0.5595051       0.718 0.07667898    0.36483786
[1] "rf"
  mtry  Accuracy Sensitivity AccuracySD SensitivitySD
1    2 0.9188990       0.898 0.04337088    0.06285786
2    5 0.9198990       0.898 0.03806620    0.05769652
3    9 0.9048889       0.884 0.03141911    0.05316641
[1] "svmLinear"
  C Accuracy Sensitivity AccuracySD SensitivitySD
1 1 0.590596       0.592 0.03686607    0.06941021
[1] "nnet"
  size decay  Accuracy Sensitivity AccuracySD SensitivitySD
1    1 0e+00 0.5935758       0.468 0.07466591     0.1360392
2    1 1e-04 0.5985657       0.518 0.05865047     0.1944679
3    1 1e-01 0.6326364       0.586 0.05789933     0.2299855
4    3 0e+00 0.6767071       0.566 0.06584864     0.1616718
5    3 1e-04 0.6155253       0.490 0.06207240     0.1609002
6    3 1e-01 0.6626768       0.572 0.04230774     0.1222747
7    5 0e+00 0.6506465       0.634 0.03284360     0.0848790
8    5 1e-04 0.6415354       0.608 0.09552621     0.1764338
9    5 1e-01 0.7127980       0.628 0.06571162     0.1132156
[1] "knn"
  k  Accuracy Sensitivity AccuracySD SensitivitySD
1 5 0.7847677       0.718 0.03789540    0.06425643
2 7 0.7877273       0.708 0.04253503    0.06941021
3 9 0.7907778       0.700 0.03152307    0.04714045
[1] "gbm"
  shrinkage interaction.depth n.minobsinnode n.trees  Accuracy Sensitivity AccuracySD SensitivitySD
1       0.1                 1             10      50 0.8848586       0.822 0.03578098    0.06762642
4       0.1                 2             10      50 0.9078788       0.872 0.02754633    0.05593647
7       0.1                 3             10      50 0.9098788       0.884 0.03443843    0.04788876
2       0.1                 1             10     100 0.9078990       0.866 0.02939592    0.05420127
5       0.1                 2             10     100 0.9239091       0.896 0.02418953    0.04402020
8       0.1                 3             10     100 0.9189192       0.902 0.03754845    0.04565572
3       0.1                 1             10     150 0.9199091       0.892 0.02628672    0.04341019
6       0.1                 2             10     150 0.9229091       0.904 0.02365112    0.04402020
9       0.1                 3             10     150 0.9178990       0.904 0.02666100    0.04087923
[1] "glm"
  parameter  Accuracy Sensitivity AccuracySD SensitivitySD
1      none 0.6035657       0.586 0.03644224    0.05081557
```

[Q6-2: Using at least 8 different models, **find the best model in terms of accuracy] (10 Points)**
(Kor: 최소 8개의 다른 모델링 방법을 통해 정확도면에서 최고의 모델을 찾아보세요.)

```
[1] "Best Accuracy Model: gbm with Accuracy: 0.923909090909091"
```

[Q6-3: Using at least 8 different models, **find the best model in terms of sensitivity] (10 Points)**
(Kor: 최소 8개의 다른 모델링 방법을 통해 민감도면에서 최고의 모델을 찾아보세요.)

```
[1] "Best Sensitivity Model: gbm with Sensitivity: 0.904"
```

-At the end- Well done for one semester!