> setwd('C:/Rdirectory/data\_mining/data\_mining\_covid');

> # install.packages("tree");

> library(tree);

> covid\_train = read.csv("covid\_train.csv", header=T);

> covid\_train$sex = as.factor(covid\_train$sex);

> covid\_train$patient\_type = as.factor(covid\_train$patient\_type);

> covid\_train$pneumonia = as.factor(covid\_train$pneumonia);

> covid\_train$age = as.numeric(covid\_train$age);

> covid\_train$diabetes = as.factor(covid\_train$diabetes);

> covid\_train$copd = as.factor(covid\_train$copd);

> covid\_train$asthma = as.factor(covid\_train$asthma);

> covid\_train$inmsupr = as.factor(covid\_train$inmsupr);

> covid\_train$hypertension = as.factor(covid\_train$hypertension);

> covid\_train$other\_disease = as.factor(covid\_train$other\_disease);

> covid\_train$cardiovascular = as.factor(covid\_train$cardiovascular);

> covid\_train$obesity = as.factor(covid\_train$obesity);

> covid\_train$renal\_chronic = as.factor(covid\_train$renal\_chronic);

> covid\_train$tobacco = as.factor(covid\_train$tobacco);

> covid\_train$contact\_other\_covid = as.factor(covid\_train$contact\_other\_covid);

> covid\_train$is\_dead = as.factor(covid\_train$is\_dead);

> str(covid\_train);

'data.frame': 32356 obs. of 16 variables:

$ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...

$ pneumonia : Factor w/ 2 levels "1","2": 2 2 1 1 1 2 1 1 1 2 ...

$ age : num 25 52 51 67 59 52 54 78 80 40 ...

$ diabetes : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 2 1 2 ...

$ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : Factor w/ 2 levels "1","2": 2 2 2 1 2 1 2 1 1 2 ...

$ other\_disease : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2 2 2 2 ...

$ obesity : Factor w/ 2 levels "1","2": 2 1 2 1 2 2 2 1 1 2 ...

$ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : Factor w/ 2 levels "1","2": 2 1 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 1 1 1 1 2 2 1 1 2 ...

$ is\_dead : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 2 2 ...

> head(covid\_train);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco

1 1 2 2 25 2 2 2 2 2 2 2 2 2 2

2 1 2 2 52 2 2 2 2 2 2 2 1 2 1

3 1 2 1 51 2 2 2 2 2 2 2 2 2 2

4 1 2 1 67 1 2 2 2 1 2 2 1 2 2

5 1 2 1 59 1 2 2 2 2 2 2 2 2 2

6 1 2 2 52 1 2 2 2 1 2 1 2 2 2

contact\_other\_covid is\_dead

1 1 2

2 1 2

3 1 2

4 1 2

5 1 2

6 2 2

> covid\_test = read.csv("covid\_test.csv", header=T);

> covid\_test$sex = as.factor(covid\_test$sex);

> covid\_test$patient\_type = as.factor(covid\_test$patient\_type);

> covid\_test$pneumonia = as.factor(covid\_test$pneumonia);

> covid\_test$age = as.numeric(covid\_test$age);

> covid\_test$diabetes = as.factor(covid\_test$diabetes);

> covid\_test$copd = as.factor(covid\_test$copd);

> covid\_test$asthma = as.factor(covid\_test$asthma);

> covid\_test$inmsupr = as.factor(covid\_test$inmsupr);

> covid\_test$hypertension = as.factor(covid\_test$hypertension);

> covid\_test$other\_disease = as.factor(covid\_test$other\_disease);

> covid\_test$cardiovascular = as.factor(covid\_test$cardiovascular);

> covid\_test$obesity = as.factor(covid\_test$obesity);

> covid\_test$renal\_chronic = as.factor(covid\_test$renal\_chronic);

> covid\_test$tobacco = as.factor(covid\_test$tobacco);

> covid\_test$contact\_other\_covid = as.factor(covid\_test$contact\_other\_covid);

> covid\_test$is\_dead = as.factor(covid\_test$is\_dead);

> str(covid\_test);

'data.frame': 5920 obs. of 16 variables:

$ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...

$ pneumonia : Factor w/ 2 levels "1","2": 2 2 1 1 1 1 2 2 2 1 ...

$ age : num 52 36 0 85 75 19 46 60 62 39 ...

$ diabetes : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 1 2 ...

$ copd : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...

$ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : Factor w/ 2 levels "1","2": 2 2 2 1 1 2 2 2 2 1 ...

$ other\_disease : Factor w/ 2 levels "1","2": 1 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...

$ obesity : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...

$ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...

$ is\_dead : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 2 2 2 2 ...

> head(covid\_test);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco

1 1 2 2 52 2 2 2 2 2 1 2 2 2 2

2 1 2 2 36 2 2 2 2 2 2 2 2 2 2

3 1 2 1 0 2 2 2 2 2 2 2 2 2 2

4 1 2 1 85 2 1 2 2 1 2 1 1 2 2

5 1 2 1 75 2 2 2 2 1 2 2 2 2 2

6 1 2 1 19 2 2 2 2 2 2 2 2 2 2

contact\_other\_covid is\_dead

1 2 2

2 2 2

3 2 1

4 2 2

5 2 2

6 2 2

> covid\_dead\_train = read.csv("covid\_dead\_train.csv", header=T);

> covid\_dead\_train$sex = as.factor(covid\_dead\_train$sex);

> covid\_dead\_train$patient\_type = as.factor(covid\_dead\_train$patient\_type);

> covid\_dead\_train$pneumonia = as.factor(covid\_dead\_train$pneumonia);

> covid\_dead\_train$age = as.numeric(covid\_dead\_train$age);

> covid\_dead\_train$diabetes = as.factor(covid\_dead\_train$diabetes);

> covid\_dead\_train$copd = as.factor(covid\_dead\_train$copd);

> covid\_dead\_train$asthma = as.factor(covid\_dead\_train$asthma);

> covid\_dead\_train$inmsupr = as.factor(covid\_dead\_train$inmsupr);

> covid\_dead\_train$hypertension = as.factor(covid\_dead\_train$hypertension);

> covid\_dead\_train$other\_disease = as.factor(covid\_dead\_train$other\_disease);

> covid\_dead\_train$cardiovascular = as.factor(covid\_dead\_train$cardiovascular);

> covid\_dead\_train$obesity = as.factor(covid\_dead\_train$obesity);

> covid\_dead\_train$renal\_chronic = as.factor(covid\_dead\_train$renal\_chronic);

> covid\_dead\_train$tobacco = as.factor(covid\_dead\_train$tobacco);

> covid\_dead\_train$contact\_other\_covid = as.factor(covid\_dead\_train$contact\_other\_covid);

> covid\_dead\_train$day\_cnt = as.numeric(covid\_dead\_train$day\_cnt);

> str(covid\_dead\_train);

'data.frame': 3600 obs. of 16 variables:

$ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...

$ pneumonia : Factor w/ 2 levels "1","2": 1 1 2 1 1 1 1 1 1 2 ...

$ age : num 78 65 58 78 92 70 71 65 56 72 ...

$ diabetes : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 1 2 2 ...

$ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...

$ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : Factor w/ 2 levels "1","2": 1 2 2 2 1 2 1 1 2 1 ...

$ other\_disease : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 2 2 2 2 ...

$ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ obesity : Factor w/ 2 levels "1","2": 1 1 2 1 2 1 1 1 2 2 ...

$ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 1 ...

$ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...

$ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 2 1 1 2 2 1 2 2 2 ...

$ day\_cnt : num 10 7 1 6 3 16 21 14 30 20 ...

> head(covid\_dead\_train);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco

1 1 2 1 78 2 2 2 2 1 2 2 1 2 2

2 1 2 1 65 2 2 2 2 2 2 2 1 2 2

3 1 2 2 58 2 2 2 2 2 1 2 2 2 2

4 1 2 1 78 2 2 2 2 2 2 2 1 2 2

5 1 2 1 92 2 2 2 2 1 2 2 2 2 2

6 1 2 1 70 2 2 2 2 2 2 2 1 2 2

contact\_other\_covid day\_cnt

1 1 10

2 2 7

3 1 1

4 1 6

5 2 3

6 2 16

> covid\_dead\_test = read.csv("covid\_dead\_test.csv", header=T);

> covid\_dead\_test$sex = as.factor(covid\_dead\_test$sex);

> covid\_dead\_test$patient\_type = as.factor(covid\_dead\_test$patient\_type);

> covid\_dead\_test$pneumonia = as.factor(covid\_dead\_test$pneumonia);

> covid\_dead\_test$age = as.numeric(covid\_dead\_test$age);

> covid\_dead\_test$diabetes = as.factor(covid\_dead\_test$diabetes);

> covid\_dead\_test$copd = as.factor(covid\_dead\_test$copd);

> covid\_dead\_test$asthma = as.factor(covid\_dead\_test$asthma);

> covid\_dead\_test$inmsupr = as.factor(covid\_dead\_test$inmsupr);

> covid\_dead\_test$hypertension = as.factor(covid\_dead\_test$hypertension);

> covid\_dead\_test$other\_disease = as.factor(covid\_dead\_test$other\_disease);

> covid\_dead\_test$cardiovascular = as.factor(covid\_dead\_test$cardiovascular);

> covid\_dead\_test$obesity = as.factor(covid\_dead\_test$obesity);

> covid\_dead\_test$renal\_chronic = as.factor(covid\_dead\_test$renal\_chronic);

> covid\_dead\_test$tobacco = as.factor(covid\_dead\_test$tobacco);

> covid\_dead\_test$contact\_other\_covid = as.factor(covid\_dead\_test$contact\_other\_covid);

> covid\_dead\_test$day\_cnt = as.numeric(covid\_dead\_test$day\_cnt);

> str(covid\_dead\_test);

'data.frame': 420 obs. of 16 variables:

$ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...

$ pneumonia : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

$ age : num 66 68 49 61 47 56 68 72 78 64 ...

$ diabetes : Factor w/ 2 levels "1","2": 1 2 2 1 1 2 1 1 1 1 ...

$ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 2 2 ...

$ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 1 2 2 2 2 2 ...

$ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : Factor w/ 2 levels "1","2": 1 2 2 1 2 2 1 1 2 2 ...

$ other\_disease : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ obesity : Factor w/ 2 levels "1","2": 1 2 1 2 2 2 1 1 1 2 ...

$ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 2 2 2 2 2 2 2 2 2 ...

$ day\_cnt : num 3 5 16 4 18 10 9 6 6 8 ...

> head(covid\_dead\_test);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco

1 1 2 1 66 1 2 2 2 1 2 2 1 2 2

2 1 2 1 68 2 2 2 2 2 2 2 2 2 2

3 1 2 1 49 2 2 2 2 2 2 2 1 2 2

4 1 2 1 61 1 2 2 2 1 2 2 2 2 2

5 1 2 1 47 1 2 1 2 2 2 2 2 2 2

6 1 2 1 56 2 2 2 2 2 2 2 2 2 2

contact\_other\_covid day\_cnt

1 1 3

2 2 5

3 2 16

4 2 4

5 2 18

6 2 10

> tree1 = tree (is\_dead ~ ., data=covid\_train, split = c("deviance"), na.action = na.pass, control = tree.control(nobs = nrow(covid\_train), minsize = 100, mindev = 0.005))

> plot(tree1);text(tree1);

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

> # summary(tree1);

> print(tree1);

node), split, n, deviance, yval, (yprob)

\* denotes terminal node

1) root 32356 44850 1 ( 0.50000 0.50000 )

2) pneumonia: 1 24382 32880 1 ( 0.59671 0.40329 )

4) age < 53.0052 9144 12580 2 ( 0.44838 0.55162 )

8) age < 36.0006 3019 3836 2 ( 0.33157 0.66843 ) \*

9) age > 36.0006 6125 8490 1 ( 0.50596 0.49404 ) \*

5) age > 53.0052 15238 18970 1 ( 0.68572 0.31428 ) \*

3) pneumonia: 2 7974 8074 2 ( 0.20429 0.79571 )

6) age < 46.0175 3898 2135 2 ( 0.07799 0.92201 ) \*

7) age > 46.0175 4076 5141 2 ( 0.32507 0.67493 ) \*

> prediction = predict(tree1, covid\_test[], type="class")

> summary(prediction)

1 2

3339 2581

> comparison=cbind(covid\_test,prediction)

> comparison=as.data.frame(comparison)

> print(paste("test 건수 : ",nrow(covid\_test)))

[1] "test 건수 : 5920"

> predictCorrect = comparison[comparison$is\_dead == comparison$prediction,];

> print(paste("사망여부 예측성공 건수 : ", nrow(predictCorrect)))

[1] "사망여부 예측성공 건수 : 3657"

> print(paste("사망여부 예측 정확도 : " ,nrow(predictCorrect)/nrow(covid\_test))) # mindev 0.005: 61.7%, 0.03 : 60.8%

[1] "사망여부 예측 정확도 : 0.617736486486486"

> tree11 = tree (day\_cnt ~ ., data = covid\_dead\_train,split = c("deviance"), na.action = na.pass, control = tree.control(nobs = nrow(covid\_dead\_train), minsize = 10, mindev = 0.001))

> plot(tree11);text(tree11);

텍스트, 도표, 평면도, 기술 도면이(가) 표시된 사진

자동 생성된 설명

> # summary(tree1);

> print(tree1);

node), split, n, deviance, yval, (yprob)

\* denotes terminal node

1) root 32356 44850 1 ( 0.50000 0.50000 )

2) pneumonia: 1 24382 32880 1 ( 0.59671 0.40329 )

4) age < 53.0052 9144 12580 2 ( 0.44838 0.55162 )

8) age < 36.0006 3019 3836 2 ( 0.33157 0.66843 ) \*

9) age > 36.0006 6125 8490 1 ( 0.50596 0.49404 ) \*

5) age > 53.0052 15238 18970 1 ( 0.68572 0.31428 ) \*

3) pneumonia: 2 7974 8074 2 ( 0.20429 0.79571 )

6) age < 46.0175 3898 2135 2 ( 0.07799 0.92201 ) \*

7) age > 46.0175 4076 5141 2 ( 0.32507 0.67493 ) \*

> prediction = predict(tree1, covid\_test[], type="class")

> summary(prediction)

1 2

3339 2581

> comparison=cbind(covid\_test,prediction)

> comparison=as.data.frame(comparison)

> print(paste("test 건수 : ",nrow(covid\_test)))

[1] "test 건수 : 5920"

> predictCorrect = comparison[comparison$is\_dead == comparison$prediction,];

> print(paste("사망여부 예측성공 건수 : ", nrow(predictCorrect)))

[1] "사망여부 예측성공 건수 : 3657"

> print(paste("사망여부 예측 정확도 : " ,nrow(predictCorrect)/nrow(covid\_test))) # mindev 0.005: 61.7%, 0.03 : 60.8%

[1] "사망여부 예측 정확도 : 0.617736486486486"

> tree11 = tree (day\_cnt ~ ., data = covid\_dead\_train,split = c("deviance"), na.action = na.pass, control = tree.control(nobs = nrow(covid\_dead\_train), minsize = 10, mindev = 0.001))

> plot(tree11);text(tree11);

> summary(tree11);

Regression tree:

tree(formula = day\_cnt ~ ., data = covid\_dead\_train, na.action = na.pass,

control = tree.control(nobs = nrow(covid\_dead\_train), minsize = 10,

mindev = 0.001), split = c("deviance"))

Variables actually used in tree construction:

[1] "age" "renal\_chronic" "cardiovascular" "hypertension" "other\_disease" "tobacco" "pneumonia" "diabetes"

Number of terminal nodes: 19

Residual mean deviance: 53.34 = 191000 / 3581

Distribution of residuals:

Min. 1st Qu. Median Mean 3rd Qu. Max.

-18.800 -4.944 -1.698 0.000 3.302 42.630

> print(tree11);

node), split, n, deviance, yval

\* denotes terminal node

1) root 3600 199700.0 10.940

2) age < 79.5 3182 182500.0 11.160

4) renal\_chronic: 1 226 11150.0 9.549

8) age < 24.5 5 14.0 2.000 \*

9) age > 24.5 221 10850.0 9.719

18) age < 31.5 5 930.8 22.800 \*

19) age > 31.5 216 9043.0 9.417

38) age < 61.5 107 2763.0 8.196 \*

39) age > 61.5 109 5964.0 10.610

78) cardiovascular: 1 21 547.8 7.762 \*

79) cardiovascular: 2 88 5204.0 11.300

158) hypertension: 1 71 3138.0 10.460 \*

159) hypertension: 2 17 1813.0 14.760

318) age < 69.5 5 508.8 24.800 \*

319) age > 69.5 12 590.9 10.580 \*

5) renal\_chronic: 2 2956 170700.0 11.290

10) age < 47.5 594 32900.0 10.430

20) other\_disease: 1 29 3975.0 15.240

40) age < 33.5 17 2952.0 18.530 \*

41) age > 33.5 12 578.9 10.580 \*

21) other\_disease: 2 565 28220.0 10.180

42) age < 31.5 129 6266.0 8.682

84) tobacco: 1 6 261.3 14.330 \*

85) tobacco: 2 123 5804.0 8.407 \*

43) age > 31.5 436 21580.0 10.630

86) pneumonia: 1 394 20380.0 10.940 \*

87) pneumonia: 2 42 793.3 7.667 \*

11) age > 47.5 2362 137200.0 11.500

22) age < 69.5 1726 105600.0 11.800

44) diabetes: 1 757 44980.0 11.370 \*

45) diabetes: 2 969 60320.0 12.130 \*

23) age > 69.5 636 31110.0 10.700 \*

3) age > 79.5 418 15850.0 9.220

6) age < 94.5 403 15470.0 9.367

12) renal\_chronic: 1 24 437.3 6.167 \*

13) renal\_chronic: 2 379 14770.0 9.570 \*

7) age > 94.5 15 134.9 5.267 \*

> prediction\_dead = predict(tree11, covid\_dead\_test[], type="vector")

> summary(prediction\_dead)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 9.57 10.70 10.70 11.37 24.80

> comparison\_dead=cbind(covid\_dead\_test,prediction\_dead)

> comparison\_dead=as.data.frame(comparison\_dead)

> comparison\_dead$prediction\_dead = round(comparison\_dead$prediction\_dead)

> print(paste("test 건수 : ", nrow(covid\_dead\_test)));

[1] "test 건수 : 420"

> # 투병일수 예측성공 기준 설정

> deadPredictCorrectCreteria = 5;

> deadPredictCorrect = comparison\_dead[abs(comparison\_dead$day\_cnt-comparison\_dead$prediction\_dead)<=deadPredictCorrectCreteria, 0]

> print(paste("투병일수 예측성공 건수(",deadPredictCorrectCreteria,"일) : " , nrow(deadPredictCorrect)));

[1] "투병일수 예측성공 건수( 5 일) : 219"

> print(paste("투병일수 예측 정확도(",deadPredictCorrectCreteria,"일) : ", nrow(deadPredictCorrect) / nrow(covid\_dead\_test))) # 기준일수 5일 : 52.1%, 7일 : 72.6%, 10일: 90%

[1] "투병일수 예측 정확도( 5 일) : 0.521428571428571"