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

> 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

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

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

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

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

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

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

tobacco contact\_other\_covid is\_dead

1 2 1 2

2 1 1 2

3 2 1 2

4 2 1 2

5 2 1 2

6 2 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

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

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

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

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

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

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

tobacco contact\_other\_covid is\_dead

1 2 2 2

2 2 2 2

3 2 2 1

4 2 2 2

5 2 2 2

6 2 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

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

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

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

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

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

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

tobacco contact\_other\_covid day\_cnt

1 2 1 10

2 2 2 7

3 2 1 1

4 2 1 6

5 2 2 3

6 2 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

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

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

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

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

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

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

tobacco contact\_other\_covid day\_cnt

1 2 1 3

2 2 2 5

3 2 2 16

4 2 2 4

5 2 2 18

6 2 2 10

> # install.packages("randomForest");

> library(randomForest);

randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.

다음의 패키지를 부착합니다: ‘randomForest’

The following object is masked from ‘package:ggplot2’:

margin

> rfFit = randomForest(is\_dead~., data=covid\_train, importance=TRUE, ntree=1000, mtry=2);

> print(rfFit);

Call:

randomForest(formula = is\_dead ~ ., data = covid\_train, importance = TRUE, ntree = 1000, mtry = 2)

Type of random forest: classification

Number of trees: 1000

No. of variables tried at each split: 2

OOB estimate of error rate: 30.7%

Confusion matrix:

1 2 class.error

1 12706 3472 0.2146124

2 6460 9718 0.3993077

> importance(rfFit);

1 2 MeanDecreaseAccuracy MeanDecreaseGini

sex 0.000000 0.0000000 0.00000 0.00000

patient\_type 0.000000 0.0000000 0.00000 0.00000

pneumonia 57.843627 27.8811153 46.69177 961.64769

age 37.723645 32.5475698 40.73734 824.18457

diabetes 8.501244 23.7662724 27.28457 170.82291

copd 1.907834 20.4681794 25.91702 28.08909

asthma 28.700336 -1.2267128 26.93381 13.78154

inmsupr 29.195466 2.1318838 28.97388 16.45565

hypertension 6.004084 18.8850496 24.18081 205.52341

other\_disease 26.123785 0.9169037 28.59837 17.18299

cardiovascular 6.413146 19.0260967 28.03315 26.70368

obesity 20.176996 8.2823386 18.75455 44.93269

renal\_chronic 13.200148 20.9371296 29.17336 40.91887

tobacco 27.894262 11.5112351 29.96805 15.23364

contact\_other\_covid 24.046483 14.0194243 22.71260 83.72687

> importance(rfFit, type=1);

MeanDecreaseAccuracy

sex 0.00000

patient\_type 0.00000

pneumonia 46.69177

age 40.73734

diabetes 27.28457

copd 25.91702

asthma 26.93381

inmsupr 28.97388

hypertension 24.18081

other\_disease 28.59837

cardiovascular 28.03315

obesity 18.75455

renal\_chronic 29.17336

tobacco 29.96805

contact\_other\_covid 22.71260

> varImpPlot(rfFit, type=1);

> varImpPlot(rfFit, type=2);

> prediction = predict(rfFit, newdata=covid\_test[], type="class");

> summary(prediction);

1 2

2994 2926

> confusionMatrix(prediction, covid\_test$is\_dead);

Confusion Matrix and Statistics

Reference

Prediction 1 2

1 1856 1138

2 1104 1822

Accuracy : 0.6213

95% CI : (0.6088, 0.6337)

No Information Rate : 0.5

P-Value [Acc > NIR] : <2e-16

Kappa : 0.2426

Mcnemar's Test P-Value : 0.4858

Sensitivity : 0.6270

Specificity : 0.6155

Pos Pred Value : 0.6199

Neg Pred Value : 0.6227

Prevalence : 0.5000

Detection Rate : 0.3135

Detection Prevalence : 0.5057

Balanced Accuracy : 0.6213

'Positive' Class : 1

> 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 : 3678"

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

[1] "사망여부 예측 정확도 : 0.621283783783784"

> rfFit\_dead = randomForest(day\_cnt~., data=covid\_dead\_train, importance=TRUE, ntree=1000, mtry=2);

> plot(rfFit\_dead);

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

> print(rfFit\_dead);

Call:

randomForest(formula = day\_cnt ~ ., data = covid\_dead\_train, importance = TRUE, ntree = 1000, mtry = 2)

Type of random forest: regression

Number of trees: 1000

No. of variables tried at each split: 2

Mean of squared residuals: 55.23043

% Var explained: 0.44

> importance(rfFit\_dead);

%IncMSE IncNodePurity

sex 0.0000000 0.0000

patient\_type 0.0000000 0.0000

pneumonia -1.1447263 661.7227

age 15.0517244 4503.0030

diabetes 5.2880215 612.0243

copd 4.0808580 524.8853

asthma -0.2276794 275.2071

inmsupr 5.2664634 413.5369

hypertension 5.5101403 544.8129

other\_disease -6.3606345 556.8313

cardiovascular -2.8214459 560.2403

obesity -1.3638791 503.6451

renal\_chronic 14.9105251 729.2328

tobacco -6.1033934 303.9541

contact\_other\_covid -1.2611167 555.3433

> importance(rfFit\_dead, type=1);

%IncMSE

sex 0.0000000

patient\_type 0.0000000

pneumonia -1.1447263

age 15.0517244

diabetes 5.2880215

copd 4.0808580

asthma -0.2276794

inmsupr 5.2664634

hypertension 5.5101403

other\_disease -6.3606345

cardiovascular -2.8214459

obesity -1.3638791

renal\_chronic 14.9105251

tobacco -6.1033934

contact\_other\_covid -1.2611167

> varImpPlot(rfFit\_dead, type=1);

> varImpPlot(rfFit\_dead, type=2);

> prediction\_dead = predict(rfFit\_dead, covid\_dead\_test[], type="response");

> summary(prediction\_dead);

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

7.989 10.347 10.906 10.773 11.298 13.049

> 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];