> 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 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

>

> #=================================================================================================================

>

> # install.packages("ipred");

> # install.packages('rpart.plot');

> library(ipred);

> library(rpart);

> library(caret);

> library(rpart.plot);

>

>

> baggFit = bagging(is\_dead~., data=covid\_train, nbag = 100, control=rpart.control(minsplit = 10),coob=T);

>

> # plot(baggFit); text(baggFit);

> print(baggFit);

Bagging classification trees with 100 bootstrap replications

Call: bagging.data.frame(formula = is\_dead ~ ., data = covid\_train,

nbag = 100, control = rpart.control(minsplit = 10), coob = T)

Out-of-bag estimate of misclassification error: 0.3128

>

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

> summary(prediction);

1 2

2721 3199

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

Confusion Matrix and Statistics

Reference

Prediction 1 2

1 1704 1017

2 1256 1943

Accuracy : 0.616

95% CI : (0.6035, 0.6285)

No Information Rate : 0.5

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.2321

Mcnemar's Test P-Value : 5.975e-07

Sensitivity : 0.5757

Specificity : 0.6564

Pos Pred Value : 0.6262

Neg Pred Value : 0.6074

Prevalence : 0.5000

Detection Rate : 0.2878

Detection Prevalence : 0.4596

Balanced Accuracy : 0.6160

'Positive' Class : 1

>

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

> comparison=as.data.frame(comparison);

> # print(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 : 3647"

> print(paste("사망여부 예측 정확도 : " ,nrow(predictCorrect)/nrow(covid\_test))); # 61%

[1] "사망여부 예측 정확도 : 0.616047297297297"

>

> #=================================================================================================================

>

> baggFit\_dead = bagging(day\_cnt~., data=covid\_dead\_train, nbag = 100, control=rpart.control(minsplit = 10),coob=T);

>

> # plot(baggFit\_dead); text(baggFit\_dead);

> print(baggFit\_dead);

Bagging regression trees with 100 bootstrap replications

Call: bagging.data.frame(formula = day\_cnt ~ ., data = covid\_dead\_train,

nbag = 100, control = rpart.control(minsplit = 10), coob = T)

Out-of-bag estimate of root mean squared error: 7.4414

>

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

> summary(prediction\_dead);

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

10.49 10.96 11.06 10.95 11.06 11.06

> # print(prediction\_dead);

> # prediction\_dead;

>

> 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(comparison\_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) : 207"

> print(paste("투병일수 예측 정확도(",deadPredictCorrectCreteria,"일) : ", nrow(deadPredictCorrect) / nrow(covid\_dead\_test)));

[1] "투병일수 예측 정확도( 5 일) : 0.492857142857143"

>

> # 5일 : 49.2%, 7일 : 69.7%, 10일 : 92.1%