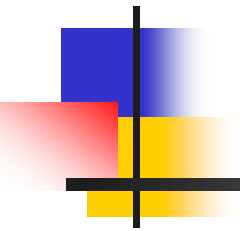


분류 및 예측 (1)
: 의사결정나무(Decision Tree)

R을 활용한 의사결정나무 실습

Decision Tree 실습 – 반품고객 예측

상황

국내 홈쇼핑 A사는 최근 소비자의 반품 횟수가 증가됨에 따라 마케팅 부서의 김팀장이 반품고객의 특성을 파악하고자 함.

데이터

홈쇼핑 A사 고객 500명에 대한 성별, 나이, 구매금액, 홈쇼핑 출연자, 반품 여부

분석 과정

① 데이터 준비 → ② 변수 지정 → ③ 훈련 · 테스트자료 분류 → ④ 의사결정 나무분석

Data: Hshopping.txt

No.	변수 이름		변수 설명	변수 유형
	SPSS용	SAS용		
1	ID	ID	고객 고유번호	수치형
2	성별	SEX	1=남자, 2=여자	범주형
3	나이	AGE	나이	수치형
4	구매금액	BUYM	1=10만 원 미만, 2=10~30만 원, 3=30만 원 이상	범주형
5	출연자	ACTOR	1=일반인, 2=유명인	범주형
6	반품 여부	RETURNSYN	0=반품 ×, 1=반품 ○	범주형

Decision Tree 실습 – 반품고객 예측

- Using "C50", "caret" & "ROCR" packages
- Related functions
  - createDataPartition() - caret package
  - C5.0() - C50 package
  - summary() - C50 package
  - C5imp() - C50 package
  - plot() - C50 package
  - predict() - C50 package
  - confusionMatrix() - caret package
  - prediction() - ROCR package
  - performance() - ROCR package
  - plot() - ROCR package

Decision Tree 실습 – 반품고객 예측

```
install.packages("caret")
install.packages("C50")
install.packages("ROCR")
library(caret)
library(C50)
library(ROCR)

cb <- read.delim("D:/Hshopping.txt", stringsAsFactors=FALSE)
head(cb)
```

```
  ID 성별 나이 구매금액 출연자 반품여부
1  1    1   33        2      2        0
2  2    2   21        3      2        1
3  3    1   45        1      1        0
4  4    1   50        2      1        0
5  5    1   21        3      1        1
6  6    1   22        3      1        1
```

```
str(cb)
```

```
data.frame': 500 obs. of  6 variables:
$ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
$ 성별    : int  1 2 1 1 1 1 2 2 2 2 ...
$ 나이    : int  33 21 45 50 21 22 27 26 28 24 ...
$ 구매금액: int  2 3 1 2 3 3 3 3 2 3 ...
$ 출연자  : int  2 2 1 1 1 1 2 2 2 2 ...
$ 반품여부: int  0 1 0 0 1 1 1 1 1 1 ...
```

범주형 값으로 변경해야 함.

```
cb$반품여부 <- factor(cb$반품여부)
```

Decision Tree 실습 – 반품고객 예측

```
set.seed(1)
inTrain <- createDataPartition(y=cb$반품여부, p=0.6, list=FALSE)
cb.train <- cb[inTrain,]
cb.test <- cb[-inTrain,]
dim(cb.train); dim(cb.test)
```

```
[1] 301   6
[1] 199   6
```

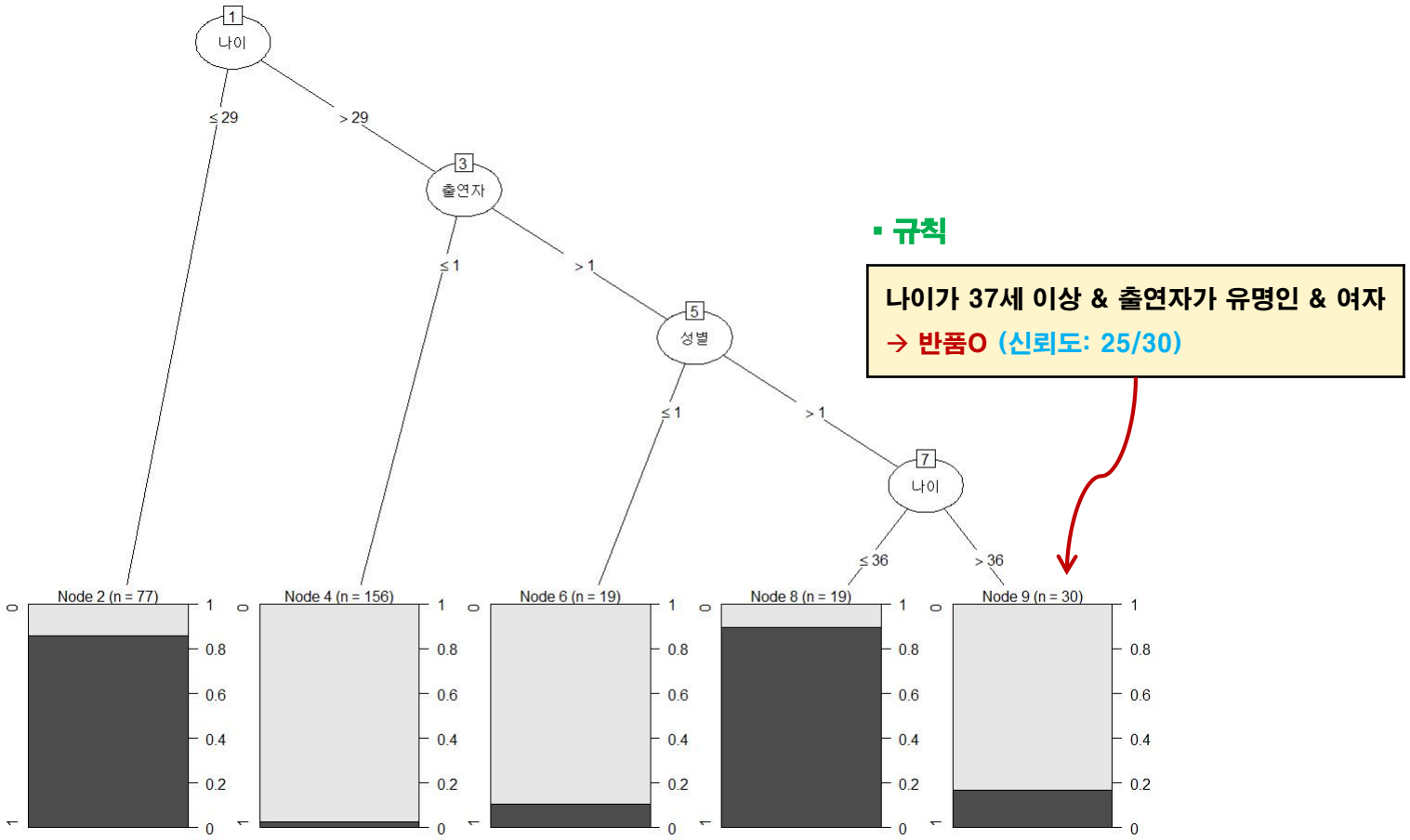
```
c5_options <- C5.0Control(winnow = FALSE, noGlobalPruning = FALSE)
c5_model <- C5.0(반품여부 ~ 성별+나이+구매금액+출연자, data=cb.train,
control=c5_options, rules=FALSE)
summary(c5_model)
```

```
Decision tree:

나이 <= 29: 1 (77/11)
나이 > 29:
:...출연자 <= 1: 0 (156/4)
    출연자 > 1:
    :...성별 <= 1: 0 (19/2)
        성별 > 1:
        :...나이 <= 36: 1 (19/2)
            나이 > 36: 0 (30/5)
```

```
  Decision Tree
----------------
Size      Errors

   5   24( 8.0%)   <<


 (a)   (b)    <-classified as
----  ----
 194    13    (a): class 0
  11    83    (b): class 1


Attribute usage:

100.00% 나이
 74.42% 출연자
 22.59% 성별
```

Decision Tree 실습 – 반품고객 예측

- `plot(c5_model)`



▪ **규칙**

나이가 37세 이상 & 출연자가 유명인 & 여자
→ 반품O (신뢰도: 25/30)

C5.0 Features (1/2)

Rule-Based Model

- Tree에 의해 생성되는 if-then statement의 set

  if X1 >= 1.7 and X2 >= 202.1 then Class = 1
  if X1 >= 1.7 and X2 < 202.1 then Class = 1
  if X1 < 1.7 then Class = 2

- 사용법: C5.0함수에서 rules 파라미터를 True로 지정

Boosting

- 여러 개의 분류모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법
  - 붓스트랩 표본을 구성하는 재표본 과정에서 각 자료에 동일한 확률을 부여하는 것이 아니라 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출
  - 붓스트랩 표본을 추출하여 분류기를 만든 후 그 분류결과를 이용하여 각 데이터가 추출될 확률을 조정하고 다음 붓스트랩 표본을 추출하는 과정을 반복
- 사용법: C5.0함수에서 trials 파라미터에 부스팅 반복횟수를 지정

C5.0 Features (2/2)

❖ Winnowing

- 입력 필드에 대해서 사전에 필드가 유용한지 측정한 다음 유용하지 않는 경우 배제하고 모델링
  - 입력필드가 많을 경우 유용
- 사용법: C5.0Control함수에서 winnow 파라미터를 True로 지정

❖ Pruning severity

- 지역적 가지치기의 강도를 조정
  - 이 값이 작으면 작을수록 가지치기 강도가 강해져서, Over-fitting의 가능성이 적어지지만, 대신 가지가 적게 되어 정확도가 떨어질 수 있음
- 사용법: C5.0Control함수에서 CF 파라미터를 0에서 1사이의 값으로 설정(default는 0.25)

❖ Global Pruning

- 전역적 가지치기 여부를 결정
  - 전역적 가지치기는 전체적으로 만들어진 Tree 구조에서 가지치기를 수행하는데 강도가 약한 sub-tree자체를 삭제
- 사용법: C5.0Control함수에서 noGlobalPruning 파라미터를 설정(default는 FALSE)

모형평가의 기본 개념

❖ 모형평가의 기준

- 일반화의 가능성
  - 같은 모집단 내의 다른 데이터에 적용하는 경우 얼마나 안정적인 결과를 제공해 주는가?
  - 확장하여 적용가능한지 여부
- 효율성
  - 모형이 얼마나 효과적으로 구축되었는가?
  - 얼마나 적은 입력변수로 모형을 구축했는가?
- 예측과 분류의 정확성
  - 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는가?
  - 아무리 안정적이고 효과적인 모형도 실제 문제에 적용했을 경우 빗나간 결과만을 양산한다면 아무런 의미가 없음

❖ 모형평가

- 예측을 위해 구축된 모형이 '임의의 모형(random model)' 보다 과연 우수한지, 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측력을 보유하고 있는지 등을 비교하고 분석하는 과정
- 성능이 좋은 모형을 찾기 위한 기준도 목표변수의 형태에 의해 다르게 고려되어야 함

모형 평가 방법 – Confusion Matrix (1/2)

재현율(Recall) or 민감도(Sensitivity)

• a/(a+b): 실제 정답인 true 중 얼마나 많은 true를 찾았는지에 대한 퍼센트(True positive rate)

		예측 결과	
		true	false
실제	true	a (TP)	b (FN)
	false	c (FP)	d (TN)

149	35
4	62

$$\frac{149}{149+35}$$

정밀도(Precision)

• a/(a+c): 모형이 true라고 판단한 것 중에서 실제 true인 것의 퍼센트(Positive predictive value)

		예측 결과	
		true	false
실제	true	a (TP)	b (FN)
	false	c (FP)	d (TN)

149	35
4	62

$$\frac{149}{149+4}$$

모형 평가 방법 – Confusion Matrix (2/2)

특이도(Specificity)

• d/(c+d): 실제 정답인 false 중 얼마나 많은 false를 찾았는지에 대한 퍼센트(True negative rate)

		예측 결과	
		true	false
실제	true	a (TP)	b (FN)
	false	c (FP)	d (TN)

149	35
4	62

$$\frac{62}{4+62}$$

정확도(Accuracy)

• (a+d)/(a+b+c+d): 전체 결과인 a, b, c, d 중에서 실제 정답과 같은 판단을 한 퍼센트

		예측 결과	
		true	false
실제	true	a (TP)	b (FN)
	false	c (FP)	d (TN)

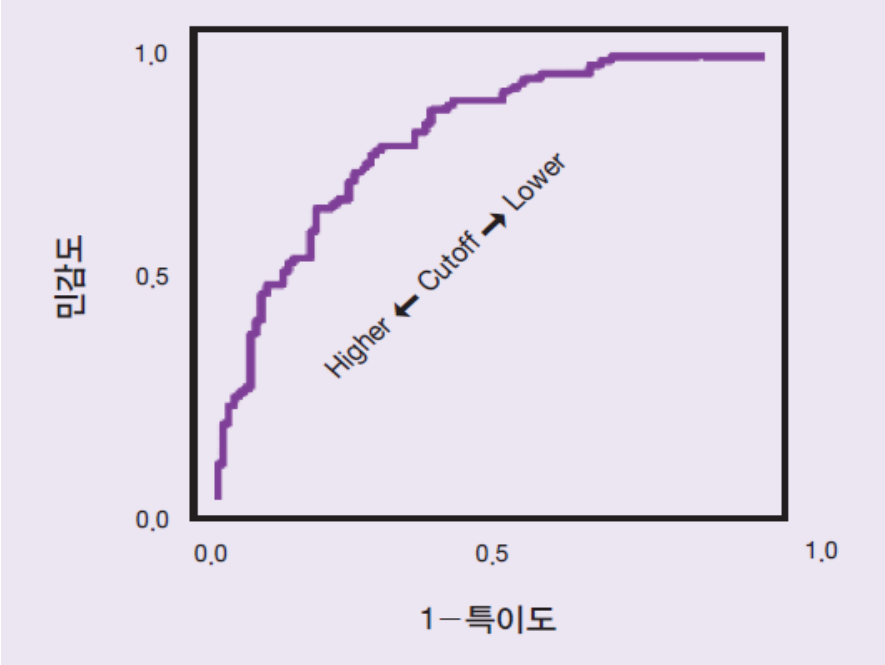
149	35
4	62

$$\frac{149+62}{149+35+4+62}$$

모형 평가 방법 – ROC curve

ROC curve & AUC

- 1-특이도(x축)와 민감도(y축)의 관계로 모형을 판단
- 모형 판단의 기준은 ROC-curve의 밑부분 면적(area under the ROC curve; AUC)이 넓을수록 좋은 모형으로 봄
  - AUC가 1이라면 완벽한 모형
  - 일반적으로 덜 정확한($0.5 < AUC \leq 0.7$), 정확한($0.7 < AUC \leq 0.9$), 매우 정확한($0.9 < AUC < 1$) 그리고 완벽한 모형($AUC = 1$)으로 분류할 수 있음



모형 평가 방법 – Lift chart (1/2)

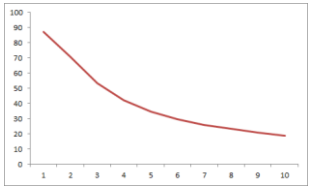
Response

- 각 등급에서 목표범주 1(true)의 비율을 나타냄

$$\frac{\text{해당 등급에서 목표변수의 특정 범주 빈도}}{\text{해당 등급에서 전체 빈도}} \times 100$$

등급	비누적				누적			
	빈도			반응률	빈도			반응률
	합계	Y=1	Y=0	Response	합계	Y=1	Y=0	Response
1	200	174	26	**174/200=87.0**	200	174	26	**174/200=87.0**
2	200	110	90	**110/200=55.0**	400	284	116	**284/400=71.0**
3	200	38	162	**38/200=19.0**	600	322	278	**322/600=53.6**
4	200	14	186	**14/200=7.0**	800	336	464	**336/800=42.0**
5	200	11	189	**11/200=5.5**	1000	347	653	**347/1000=34.7**
6	200	10	190	**10/200=5.0**	1200	357	843	**357/1200=29.7**
7	200	7	193	**7/200=3.5**	1400	364	1036	**364/1400=26.0**
8	200	10	190	**10/200=5.0**	1600	374	1226	**374/1600=23.3**
9	200	3	197	**3/200=1.5**	1800	377	1423	**377/1800=20.9**
10	200	4	196	**4/200=2.0**	2000	381	1619	**381/2000=19.0**

모형 평가 방법 – Lift chart (2/2)



Lift

- 전체 반응률에 비해 각 등급에서 반응률이 얼마나 높은지를 나타냄
- 상위 등급에서의 Lift가 매우 크고 하위 등급으로 갈수록 Lift가 감소하면 이는 모형의 예측력이 적절함을 의미함. 등급에 관계없이 Lift에 차이가 없다면 이는 모형의 예측력이 좋지 않음을 나타냄

$$\text{Lift} = \frac{\text{해당 등급에서 반응률(response)}}{\text{전체 반응률}}$$

등급	비누적				누적			
	빈도			반응률	빈도			반응률
	합계	Y=1	Y=0	Lift	합계	Y=1	Y=0	Lift
1	200	174	26	**870/190=4.57**	200	174	26	**870/190=4.57**
2	200	110	90	**550/190=2.89**	400	284	116	**710/190=3.73**
3	200	38	162	**190/190=1.00**	600	322	278	**536/190=2.82**
4	200	14	186	**70/190=0.36**	800	336	464	**420/190=2.21**
5	200	11	189	**55/190=0.28**	1000	347	653	**347/190=1.82**
6	200	10	190	**50/190=0.26**	1200	357	843	**297/190=1.56**
7	200	7	193	**35/190=0.18**	1400	364	1036	**260/190=1.36**
8	200	10	190	**50/190=0.26**	1600	374	1226	**233/190=1.23**
9	200	3	197	**15/190=0.07**	1800	377	1423	**209/190=1.10**
10	200	4	196	**20/190=0.10**	2000	381	1619	**190/190=1.00**
전체	전체 반응률 =381/2000=19%							

Decision Tree 실습 – 반품고객 예측

```
cb.test$c5_pred <- predict(c5_model, cb.test, type="class")
cb.test$c5_pred_prob <- predict(c5_model, cb.test, type="prob")
head(cb.test)
```

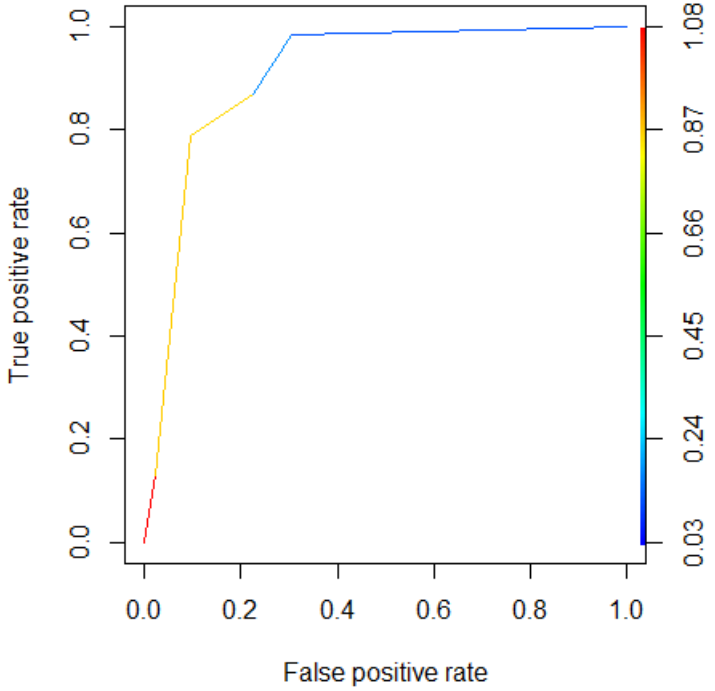
```
   ID 성별 나이 구매금액 출연자 반품여부 c5_pred c5_pred_prob.0 c5_pred_prob.1
1   1    1   33        2      2        0       0     0.88438538     0.11561462
2   2    2   21        3      2        1       1     0.14984241     0.85015759
3   3    1   45        1      1        0       0     0.97253317     0.02746683
5   5    1   21        3      1        1       1     0.14984241     0.85015759
8   8    2   26        3      2        1       1     0.14984241     0.85015759
12 12    1   64        1      1        0       0     0.97253317     0.02746683
```

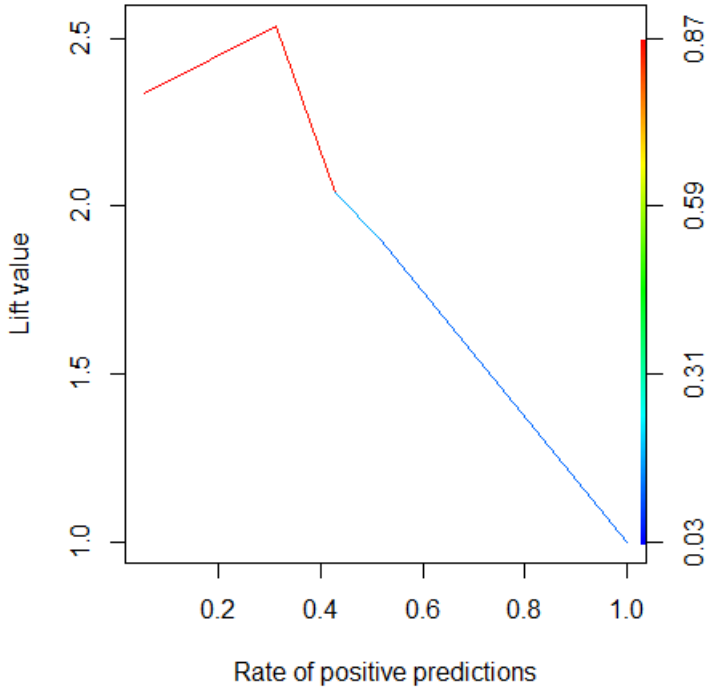
```
confusionMatrix(cb.test$c5_pred, cb.test$반품여부)
```

```
          Reference
Prediction   0   1
         0 124  13
         1  13  49

               Accuracy : 0.8693
                 95% CI : (0.8144, 0.9128)
    No Information Rate : 0.6884
    P-Value [Acc > NIR] : 2.375e-09

                  Kappa : 0.6954
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9051
            Specificity : 0.7903
         Pos Pred Value : 0.9051
         Neg Pred Value : 0.7903
             Prevalence : 0.6884
         Detection Rate : 0.6231
   Detection Prevalence : 0.6884
      Balanced Accuracy : 0.8477
```

Decision Tree 실습 – 반품고객 예측

```
c5_pred <- prediction(cb.test$c5_pred_prob[,2],cb.test$반품여부)
c5_model.perf1 <- performance(c5_pred, "tpr", "fpr") # ROC curve
c5_model.perf2 <- performance(c5_pred, "lift", "rpp") # Lift chart
plot(c5_model.perf1, colorize=TRUE); plot(c5_model.perf2, colorize=TRUE)
```






```
performance(c5_pred, "auc")@y.values[[1]]
```

```
[1] 0.9064045
```

개인과제 #1 – 10월29일 제출

과제내용

- 그룹과제#1에서 만들었던 Customer Signature를 이용하여 H백화점 고객의 성별을 예측하는 의사결정나무(C5.0) 분석을 수행하시오. C5.0의 다양한 옵션을 사용하여 여러 개의 의사결정나무 모형을 생성한 후 모형평가를 통해 최종모형을 선택하시오.
- 임의의 고객에 대한 구매정보(HDS_Transactions_MG.tab)는 알고 있지만 그 고객이 누구인지 모른다는 가정 하에 성별을 예측하는 것이므로, H백화점의 고객정보(HDS_Customers.tab) 중에서 성별 필드만 예측변수로 사용하고 나머지 필드는 독립변수로 사용하지 말아야 함.

제출방법

- 가상대학 과제관리를 통해 제출해야 함.
- 분석보고서(*.PPT 또는 *.PDF)와 분석코드(*.R)를 같이 제출할 것.
- 각 화일명은 본인의 이름으로 할 것.