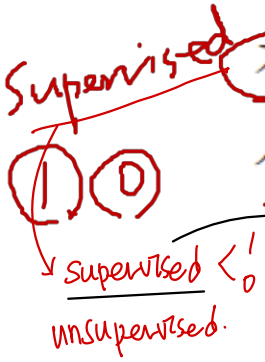
경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

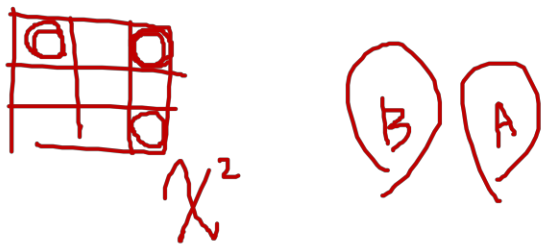
유 진 호

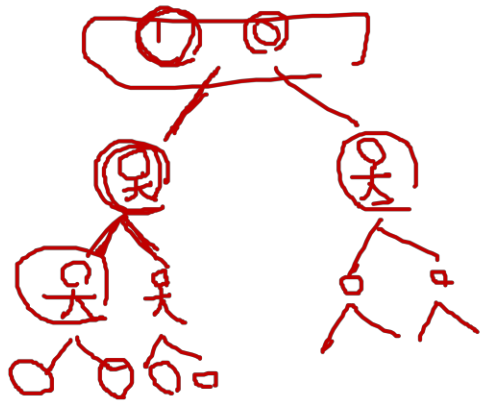
jhyoo@smu.ac.kr

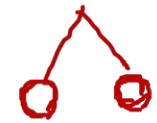
의사결정나무(Decision Tree)는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소그룹으로 분류하거나 예측할 때 수행하는 분석방법이다. Decision Tree는 시각적이고 명시적인 방법으로 의사결정 과정과 결정된 의사를 보여주는데 사용된다. Decision Tree 분석방법은 지도 분류 학습에서 가장 유용하게 사용되고 있는 기법 중 하나로서, 노드(집단)내에서는 동질성이 커지고, 노드(집단)간에는 이질성이 가장 커지도록 개체를 분류한다.



목표변수가 이산형인 경우 상위노드에서 가지분할을 수행할 때, 분류 기준변수와 분류 기준값의 선택방법으로 카이제곱 통계량의 p-값, 지니 지수(Gini index), 엔트로피 지수(Entropy index) 등이 사용된다.



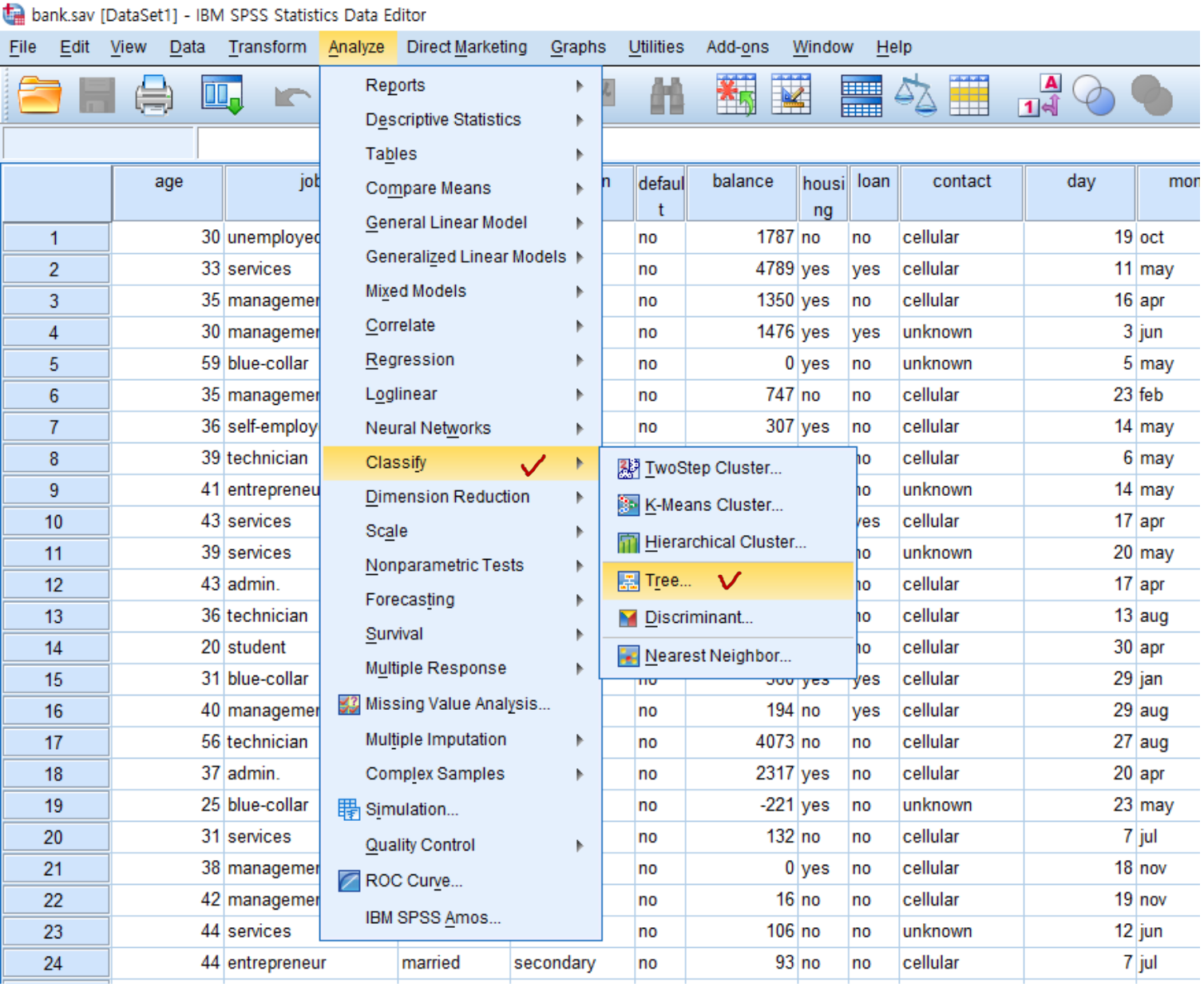




CHAID(Chi-square Automatic Interaction Detection) 알고리즘은 최적의 분할, 즉 최적의 예측변수를 선택하는데 있어 카이제곱 검정 또는 F-검정을 이용하여 데이터를 분리하는 방법을 사용한다. 범주형 변수에 대해서는 카이제곱 검정을 사용하고, 연속형 변수에 대해서는 F-검정을 사용한다.

$$F = \frac{\text{Between}}{\text{within}}$$

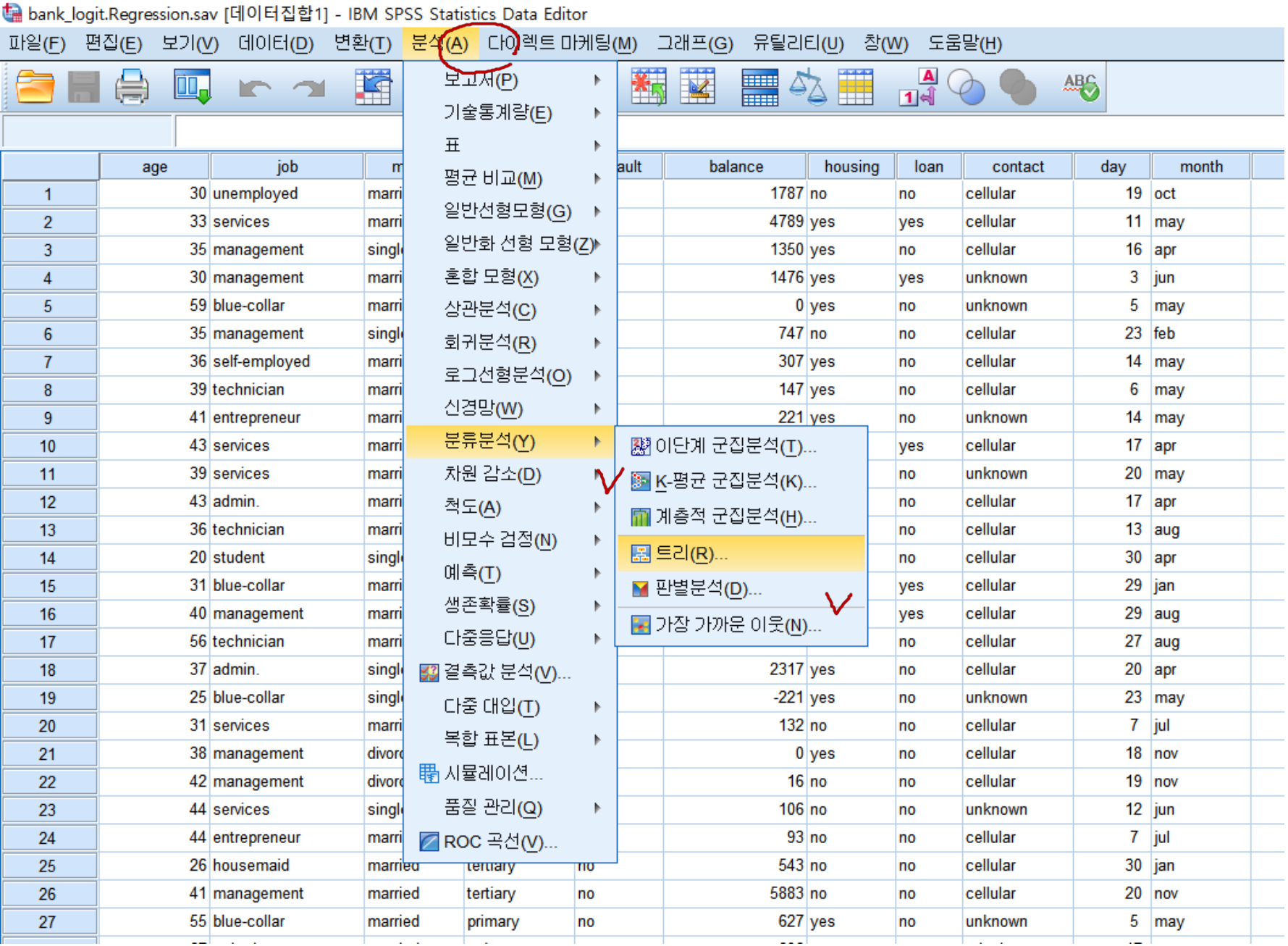
CART(Classification and Regression Trees) 알고리즘은 지니 지수(Gini Index) 또는 분산의 감소량을 사용하여 나무의 가지를 이진(Binary) 분리한다. 범주형 변수에 대해서는 지니 지수를 사용하고, 연속형 변수에 대해서는 분산의 감소량을 사용한다.



의미있는 이산형 변수,
의미있는 연속형 변수들을 가지고
Decision Tree 분석
.분류표 제공(예측력)

6개로 실습

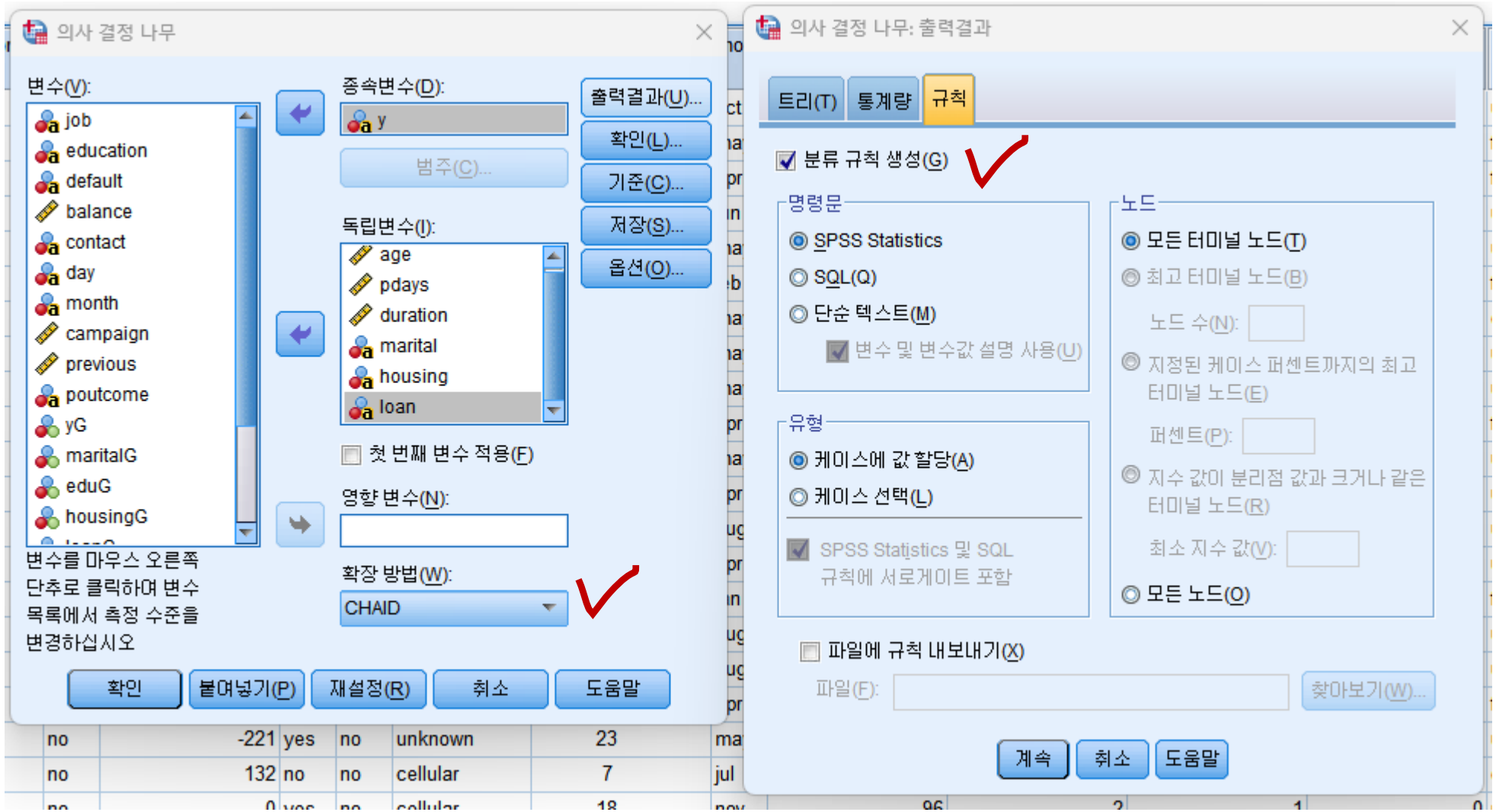
age
pdays
duration
marital -> maritalG
Housing -> HousingG
Loan -> LoanG



	age	job	marital	education	default	balance	housing	loan	contact	day	month
1	30	unemployed	marri			1787	no	no	cellular	19	oct
2	33	services	marri			4789	yes	yes	cellular	11	may
3	35	management	single			1350	yes	no	cellular	16	apr
4	30	management	marri			1476	yes	yes	unknown	3	jun
5	59	blue-collar	marri			0	yes	no	unknown	5	may
6	35	management	single			747	no	no	cellular	23	feb
7	36	self-employed	marri			307	yes	no	cellular	14	may
8	39	technician	marri			147	yes	no	cellular	6	may
9	41	entrepreneur	marri			221	yes	no	unknown	14	may
10	43	services	marri					yes	cellular	17	apr
11	39	services	marri					no	unknown	20	may
12	43	admin.	marri					no	cellular	17	apr
13	36	technician	marri					no	cellular	13	aug
14	20	student	single					no	cellular	30	apr
15	31	blue-collar	marri					yes	cellular	29	jan
16	40	management	marri					yes	cellular	29	aug
17	56	technician	marri					no	cellular	27	aug
18	37	admin.	single			2317	yes	no	cellular	20	apr
19	25	blue-collar	single			-221	yes	no	unknown	23	may
20	31	services	marri			132	no	no	cellular	7	jul
21	38	management	divorc			0	yes	no	cellular	18	nov
22	42	management	divorc			16	no	no	cellular	19	nov
23	44	services	single			106	no	no	unknown	12	jun
24	44	entrepreneur	marri			93	no	no	cellular	7	jul
25	26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan
26	41	management	married	tertiary	no	5883	no	no	cellular	20	nov
27	55	blue-collar	married	primary	no	627	yes	no	unknown	5	may

6개로 실습

age
pdays
duration
marital -> maritalG
Housing -> HousingG
Loan -> LoanG

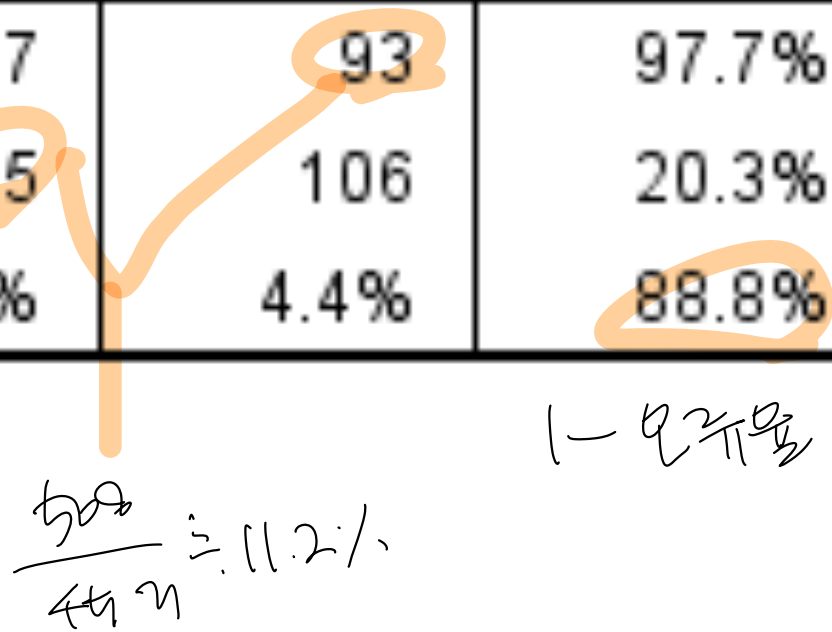

의사 결정 나무
변수(V):
job
education
default
balance
contact
day
month
campaign
previous
poutcome
yG
maritalG
eduG
housingG
변수를 마우스 오른쪽 단추로 클릭하여 변수 목록에서 측정 수준을 변경하십시오
종속변수(D):
y
범주(C)...
독립변수(I):
age
pdays
duration
marital
housing
loan
첫 번째 변수 적용(F)
영향 변수(N):
확장 방법(W):
CHAID
출력결과(U)...
확인(L)...
기준(C)...
저장(S)...
옵션(O)...
확인
붙여넣기(P)
재설정(R)
취소
도움말
의사 결정 나무: 출력결과
트리(T)
통계량
규칙
분류 규칙 생성(G)
명령문
SPSS Statistics
SQL(Q)
단순 텍스트(M)
변수 및 변수값 설명 사용(U)
유형
케이스에 값 할당(A)
케이스 선택(L)
SPSS Statistics 및 SQL 규칙에 서로게이트 포함
노드
모든 터미널 노드(T)
최고 터미널 노드(B)
노드 수(N):
지정된 케이스 퍼센트까지의 최고 터미널 노드(E)
퍼센트(P):
지수 값이 분리점 값과 크거나 같은 터미널 노드(R)
최소 지수 값(V):
모든 노드(O)
파일에 규칙 내보내기(X)
파일(F):
찾아보기(W)...
계속
취소
도움말
no -221 yes no unknown 23
no 132 no no cellular 7

분류

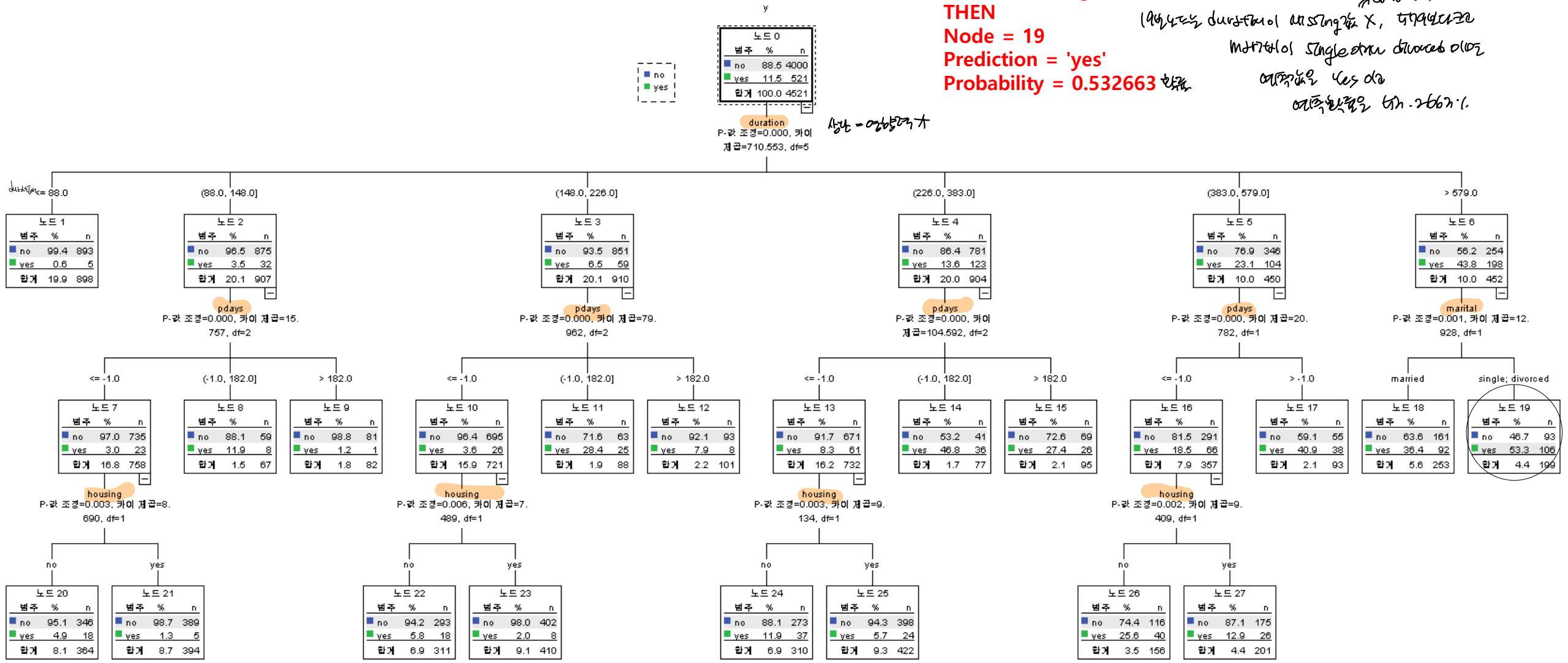
감시됨	예측		
	no	yes	정확도(%)
no	3907	93	97.7%
yes	415	106	20.3%
전체 퍼센트	95.6%	4.4%	88.8%

성장방법: CHAID

종속변수: y



/* Node 19 */.
IF (duration NOT MISSING AND (duration > 579)) AND (marital = "single" OR marital = "divorced")
THEN
Node = 19
Prediction = 'yes'
Probability = 0.532663



Decision Tree 이용

415 + 93 = 508

분류

감시됨	예측		
	.00	1.00	정확도(%)
.00	3907	93	97.7%
1.00	415	106	20.3%
전체 퍼센트	95.6%	4.4%	88.8%

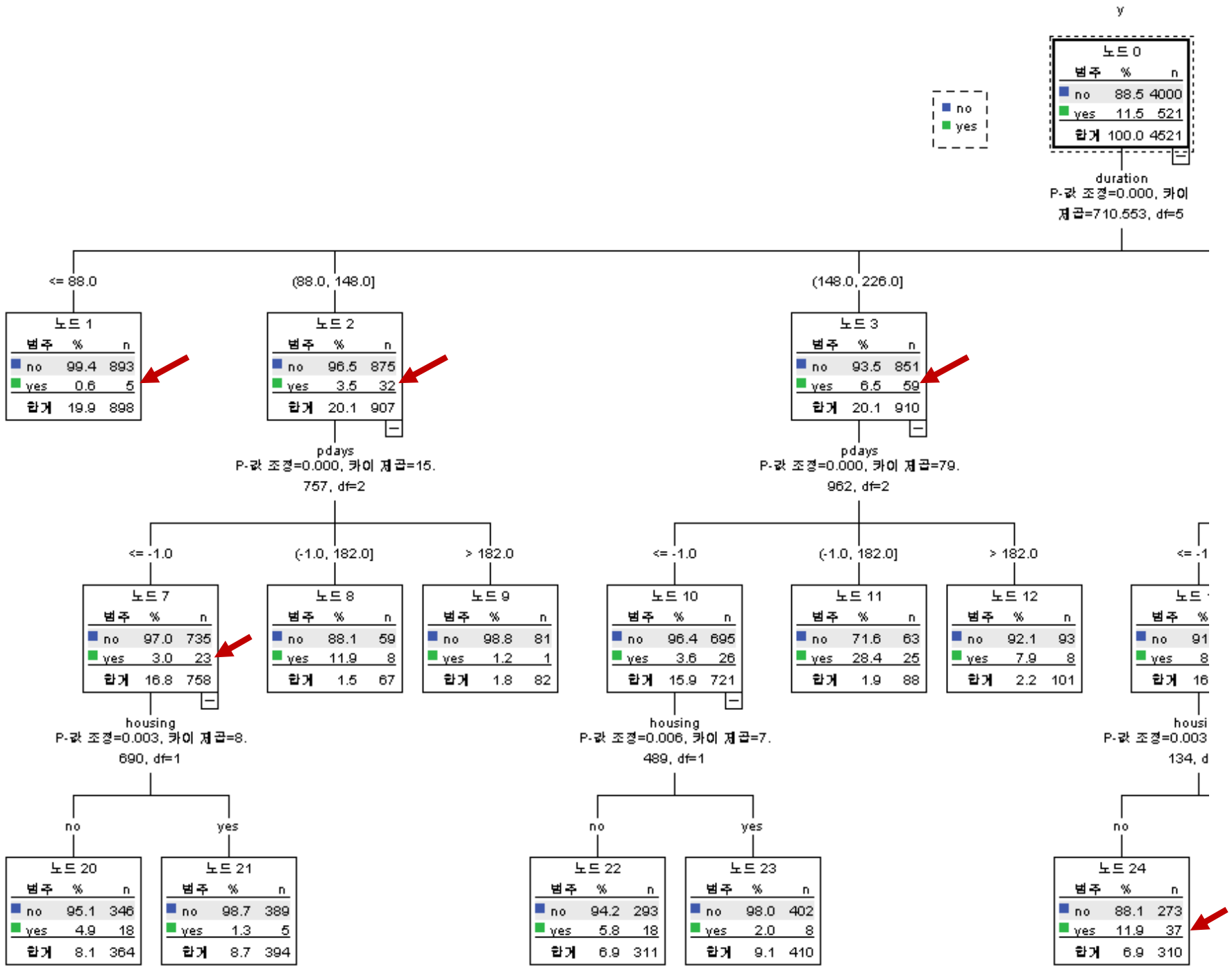
성장방법: CHAID
종속변수: YG

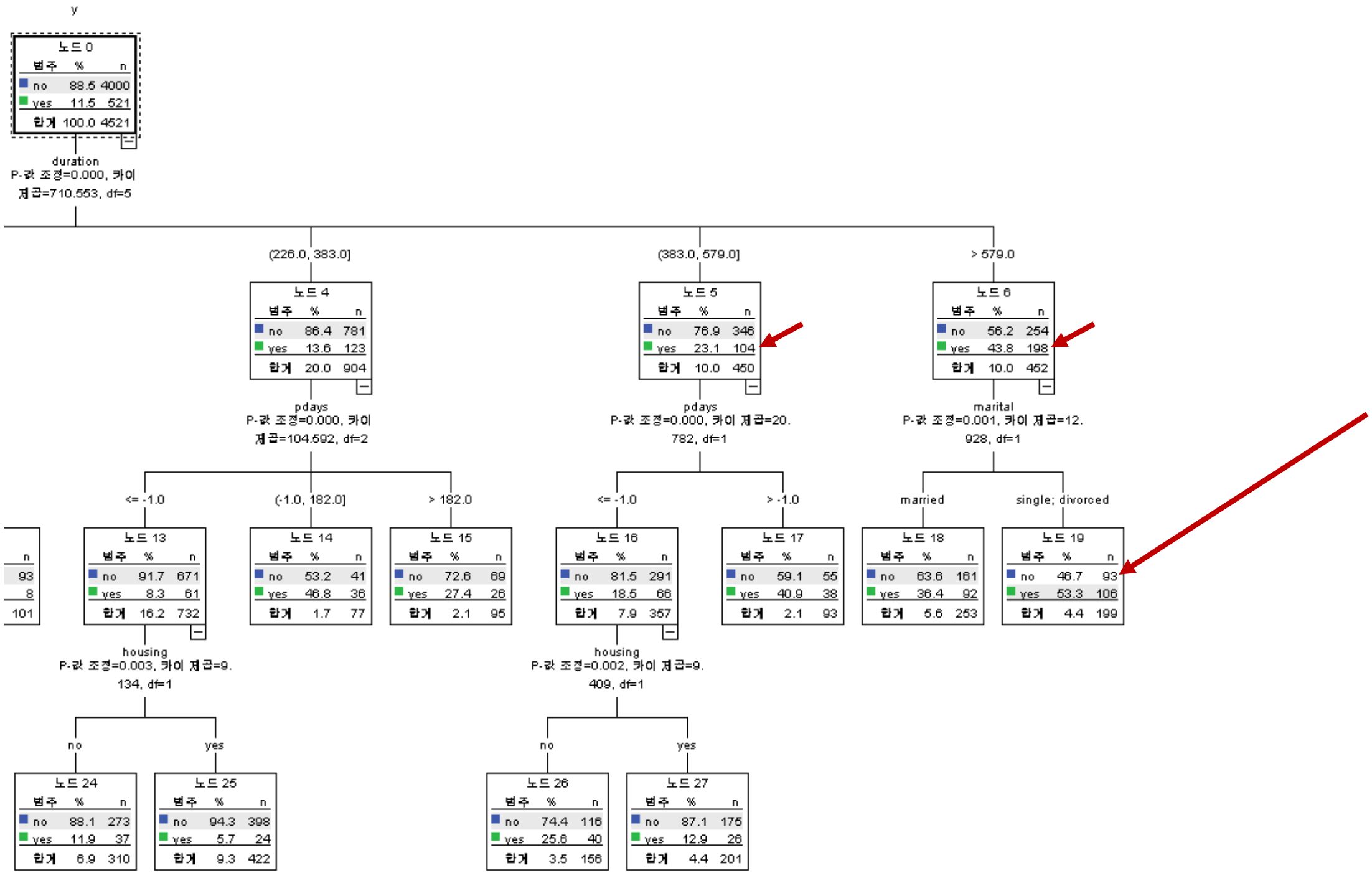
415,
93,

어디서 나오나?

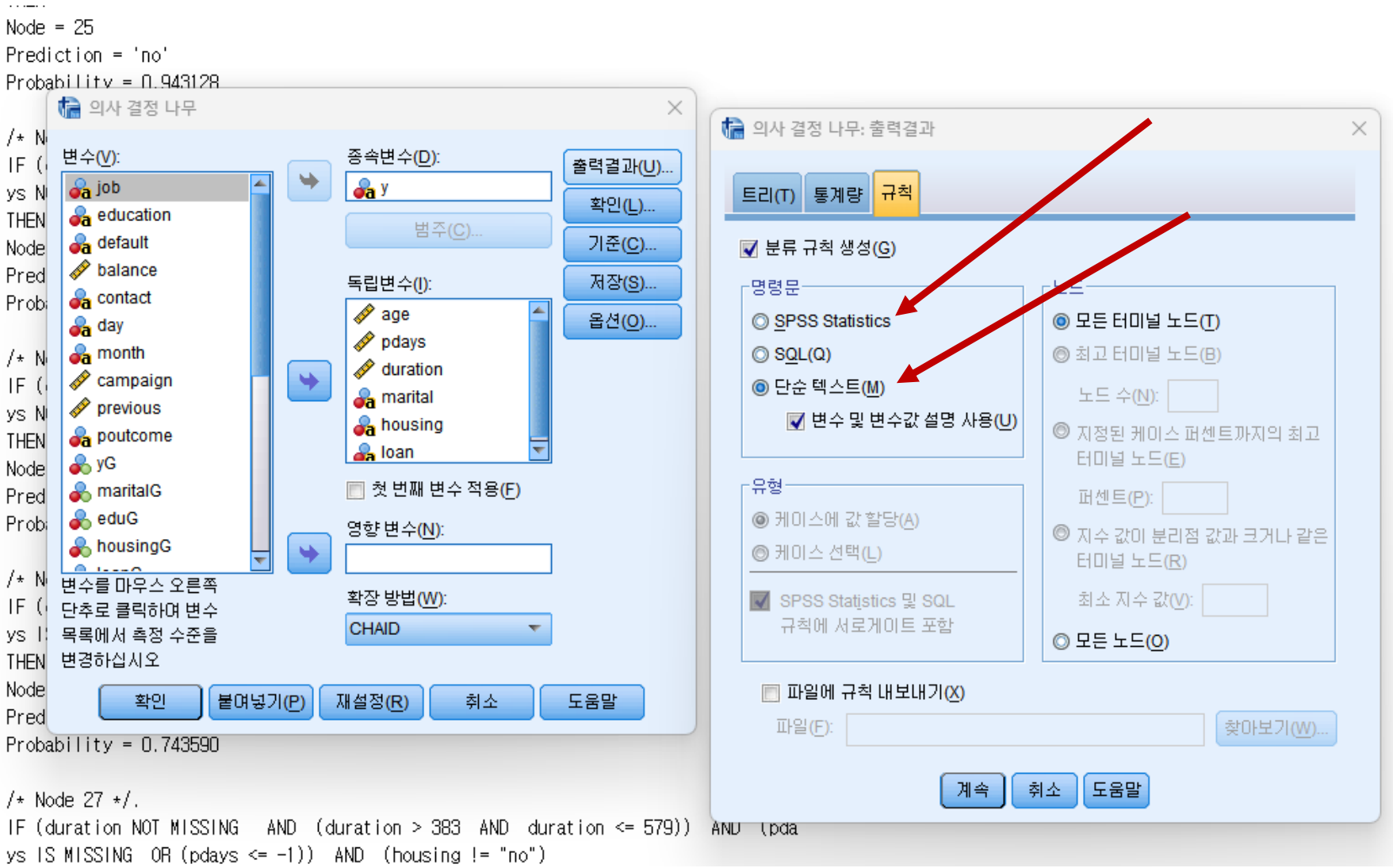
No로예측 but Yes

Yes로 예측 but No.


y
no
yes
노드 0
범주 % n
no 88.5 4000
yes 11.5 521
합계 100.0 4521
duration
P-값 조정=0.000, 카이 제곱=710.553, df=5
<= 88.0
(88.0, 148.0]
(148.0, 226.0]
노드 1
범주 % n
no 99.4 893
yes 0.6 5
합계 19.9 898
노드 2
범주 % n
no 96.5 875
yes 3.5 32
합계 20.1 907
노드 3
범주 % n
no 93.5 851
yes 6.5 59
합계 20.1 910
pdays
P-값 조정=0.000, 카이 제곱=15.757, df=2
pdays
P-값 조정=0.000, 카이 제곱=79.962, df=2
<= -1.0
(-1.0, 182.0]
> 182.0
노드 7
범주 % n
no 97.0 735
yes 3.0 23
합계 16.8 758
노드 8
범주 % n
no 88.1 59
yes 11.9 8
합계 1.5 67
노드 9
범주 % n
no 98.8 81
yes 1.2 1
합계 1.8 82
노드 10
범주 % n
no 96.4 695
yes 3.6 26
합계 15.9 721
노드 11
범주 % n
no 71.6 63
yes 28.4 25
합계 1.9 88
노드 12
범주 % n
no 92.1 93
yes 7.9 8
합계 2.2 101
housing
P-값 조정=0.003, 카이 제곱=8.690, df=1
housing
P-값 조정=0.006, 카이 제곱=7.489, df=1
no
yes
노드 20
범주 % n
no 95.1 346
yes 4.9 18
합계 8.1 364
노드 21
범주 % n
no 98.7 389
yes 1.3 5
합계 8.7 394
노드 22
범주 % n
no 94.2 293
yes 5.8 18
합계 6.9 311
노드 23
범주 % n
no 98.0 402
yes 2.0 8
합계 9.1 410
노드 24
범주 % n
no 88.1 273
yes 11.9 37
합계 6.9 310


y
노드 0
범주 % n
no 88.5 4000
yes 11.5 521
합계 100.0 4521
duration
P-값 조정=0.000, 카이 제곱=710.553, df=5
(226.0, 383.0]
(383.0, 579.0]
> 579.0
노드 4
범주 % n
no 86.4 781
yes 13.6 123
합계 20.0 904
노드 5
범주 % n
no 76.9 346
yes 23.1 104
합계 10.0 450
노드 6
범주 % n
no 56.2 254
yes 43.8 198
합계 10.0 452
pdays
P-값 조정=0.000, 카이 제곱=104.592, df=2
pdays
P-값 조정=0.000, 카이 제곱=20.782, df=1
marital
P-값 조정=0.001, 카이 제곱=12.928, df=1
<= -1.0
(-1.0, 182.0]
> 182.0
<= -1.0
> -1.0
married
single; divorced
노드 13
범주 % n
no 91.7 671
yes 8.3 61
합계 16.2 732
노드 14
범주 % n
no 53.2 41
yes 46.8 36
합계 1.7 77
노드 15
범주 % n
no 72.6 69
yes 27.4 26
합계 2.1 95
노드 16
범주 % n
no 81.5 291
yes 18.5 66
합계 7.9 357
노드 17
범주 % n
no 59.1 55
yes 40.9 38
합계 2.1 93
노드 18
범주 % n
no 63.6 161
yes 36.4 92
합계 5.6 253
노드 19
범주 % n
no 46.7 93
yes 53.3 106
합계 4.4 199
n
93
8
101
housing
P-값 조정=0.003, 카이 제곱=9.134, df=1
housing
P-값 조정=0.002, 카이 제곱=9.409, df=1
no
yes
no
yes
노드 24
범주 % n
no 88.1 273
yes 11.9 37
합계 6.9 310
노드 25
범주 % n
no 94.3 398
yes 5.7 24
합계 9.3 422
노드 26
범주 % n
no 74.4 116
yes 25.6 40
합계 3.5 156
노드 27
범주 % n
no 87.1 175
yes 12.9 26
합계 4.4 201

Rules 생성 및 적용
.SPSS Statistics(적용)
.Simple Text(이해)



숙제

교재처럼 6개 변수를 예측변수로 활용하여 의사결정나무 분석을 실시하고,

그 결과를 Excel파일(*.xls)로 export(내보내기)하여 통계 결과에 대한 해석을 달기

1) 가장 큰 영향을 주는 변수는?

2) 정확도 해석하기.

3) simple text로 수행 후 => 노드1번, 18번, 19번, 20번, 21번 해석하기

4) 오류난 것 합계 계산하여 <508개>와 일치하는지 확인하기