

빅데이터 분석을 위한  
**데이터마이닝 방법론**  
SAS Enterprise Miner 활용사례를 중심으로

## <<제3장>> 의사결정나무분석

### Chapter 3 Decision Tree Analysis

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현

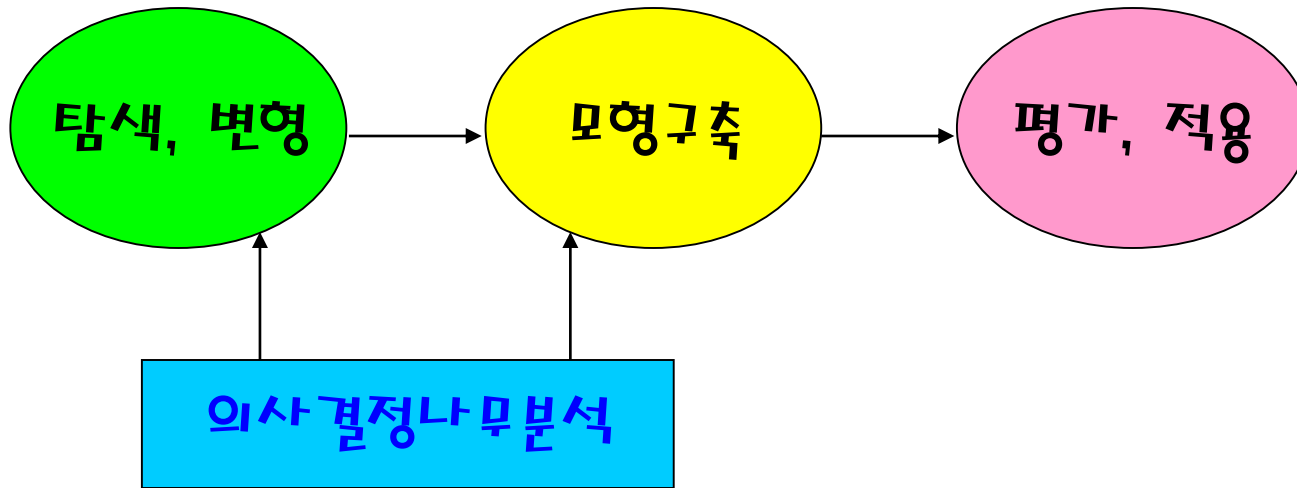
Update: 2014. 4. 1.

## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

# 데이터마이닝과 의사결정나무



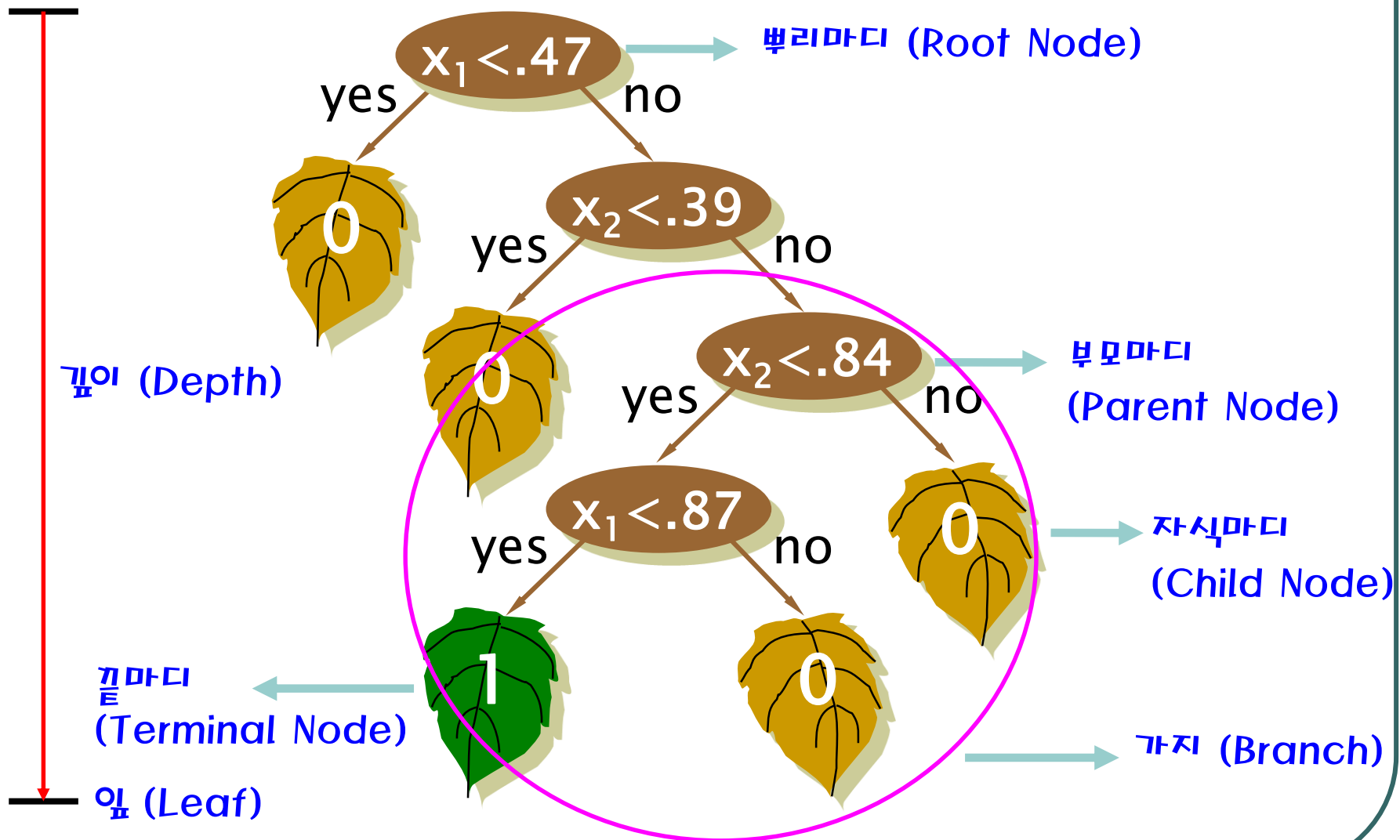
- 데이터 탐색

- ✓ 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수 또는 모델에 포함되어야 할 상호작용의 효과를 찾아내기 위해서 사용될 수 있다.

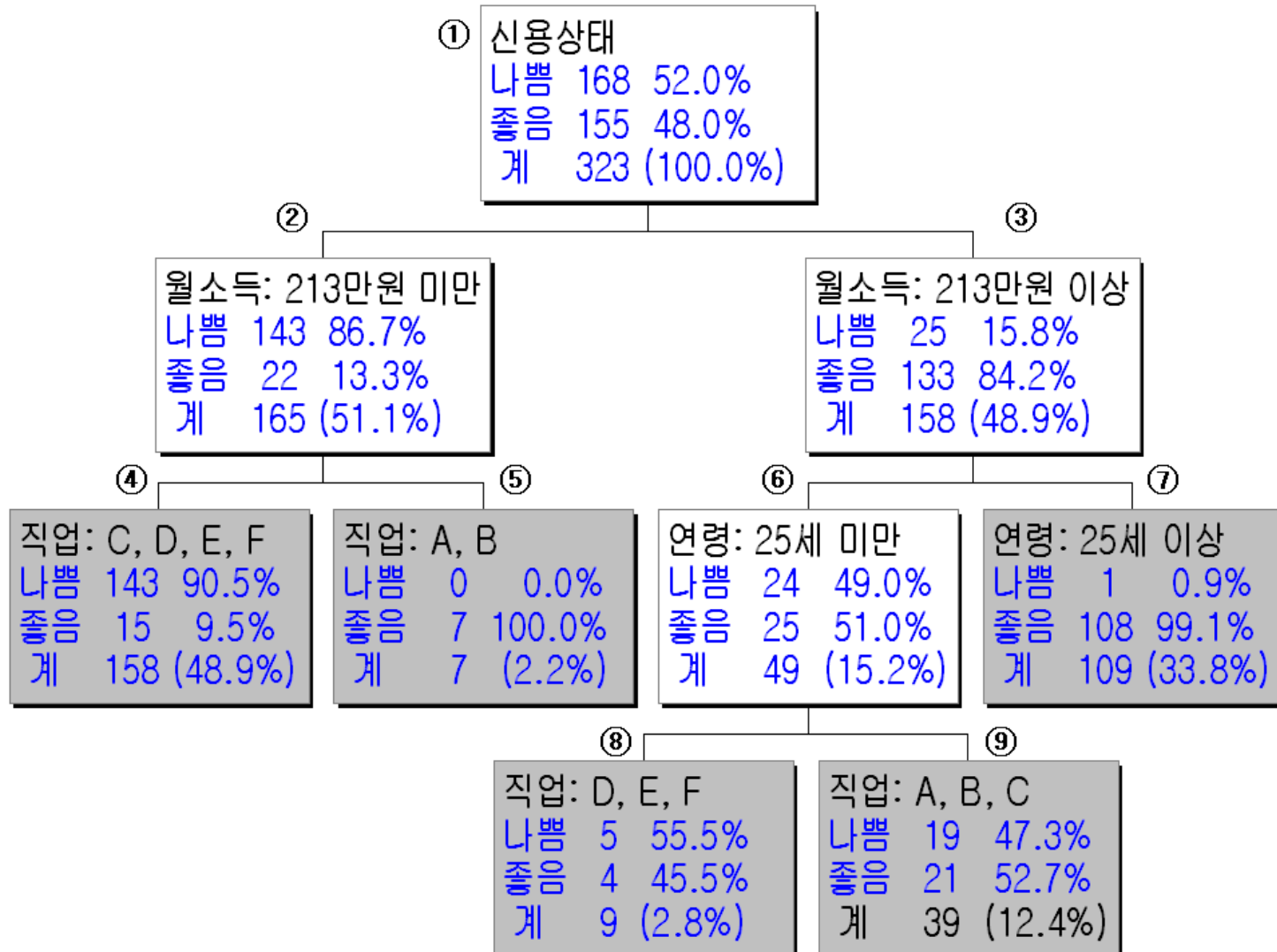
- 예측모형

- ✓ 의사결정나무 자체가 하나의 분류 또는 예측모형으로 사용될 수 있다.

### 3.1.1 의사결정나무의 구성요소



## <<사례>> 신용평가 문제



### 3.1.2 의사결정나무의 형성과정

---

- 의사결정나무의 형성

- ✓ 분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.

- 가지치기

- ✓ 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론 규칙(induction rule)을 가지고 있는 가지(branch)를 제거한다.

- 타당성 평가

- ✓ 이익도표(gains chart)나 위험도표(risk chart)와 같은 모형평가 도구 또는 검증용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.

- 해석 및 예측

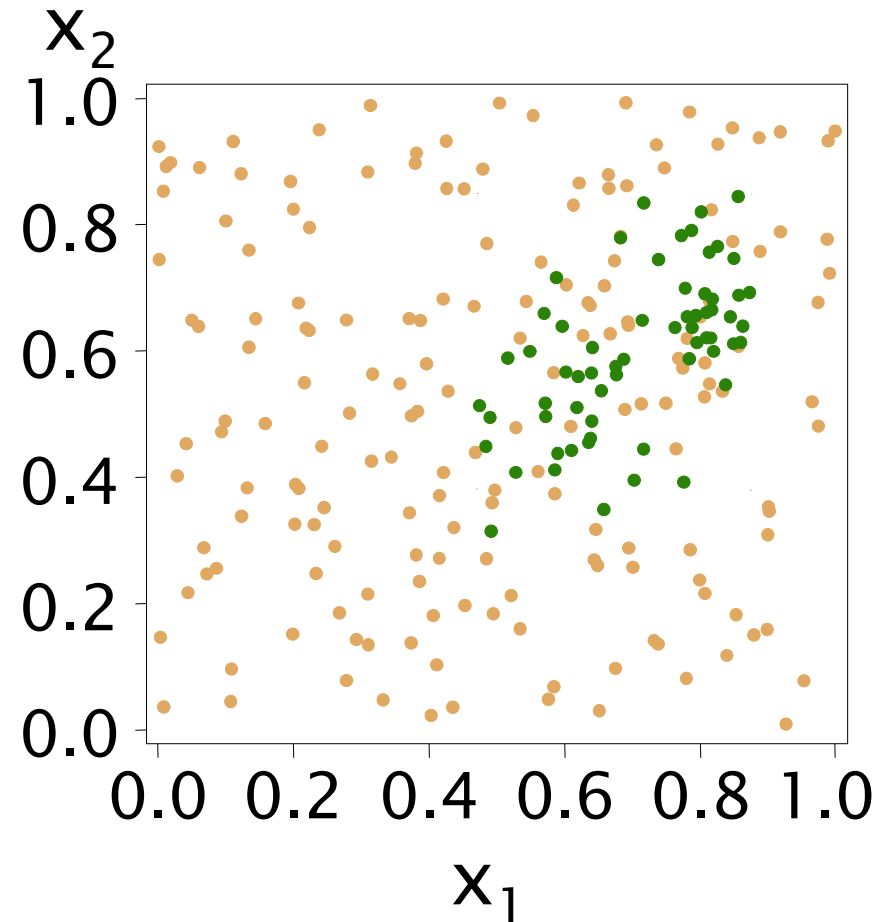
- ✓ 의사결정나무를 해석하고 예측모형을 구축한다.

## <<사례>> 최적 분리

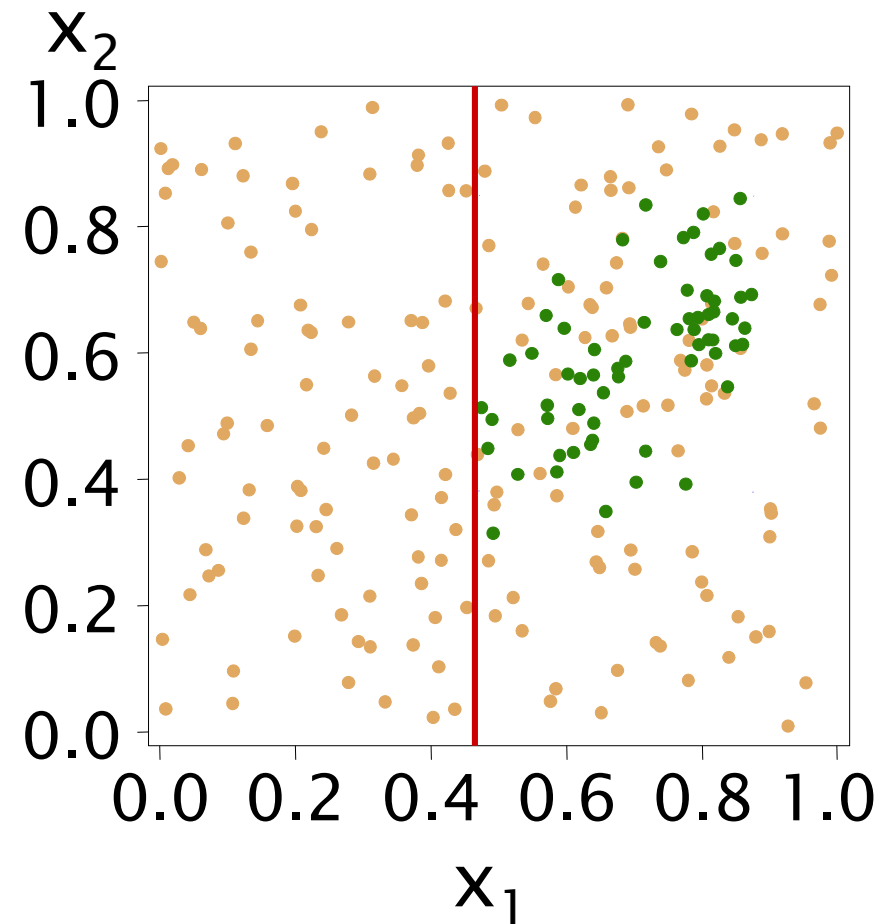
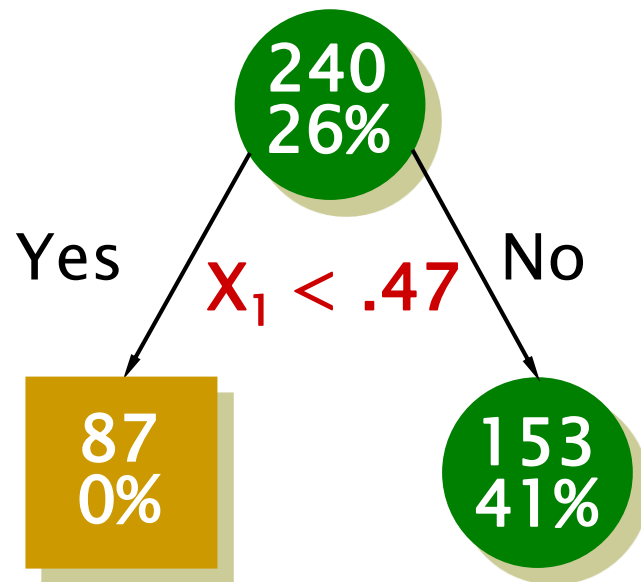
240  
26%

478 개의 분리조합이 존재

- $x_1$ 에 대해서 239개  
( $x_1 < .25$ ,  $x_1 < .26$ , etc.)
- $x_2$ 에 대해서 239개  
( $x_2 < .43$ ,  $x_2 < .86$ , etc.)

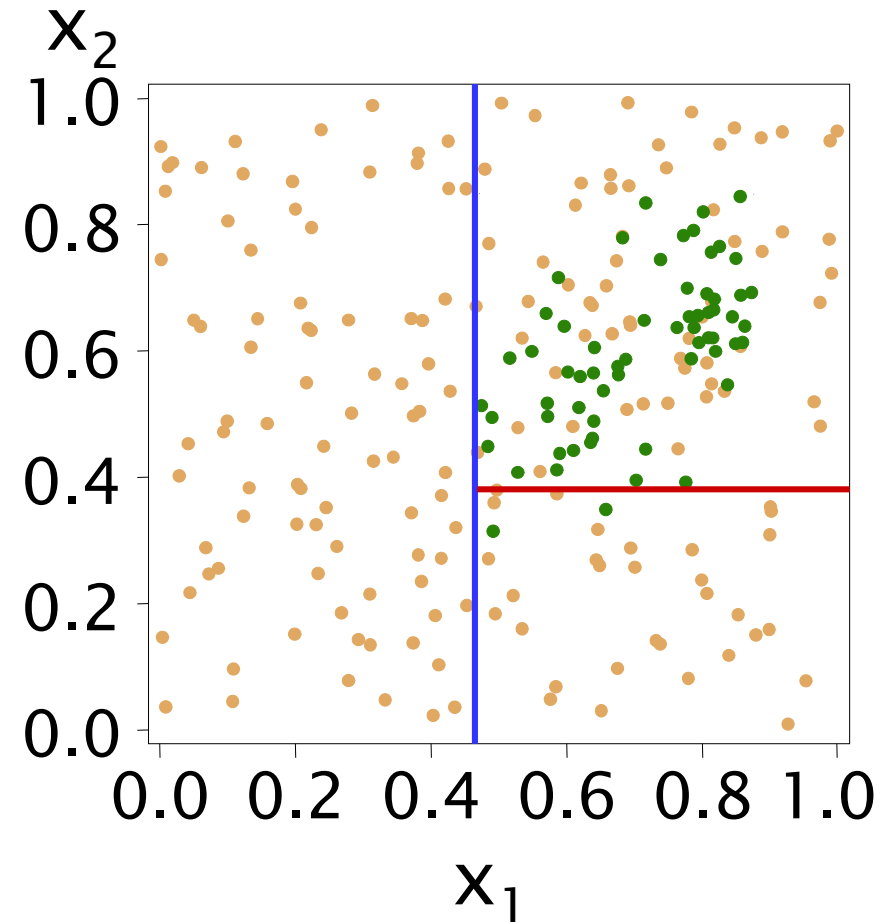
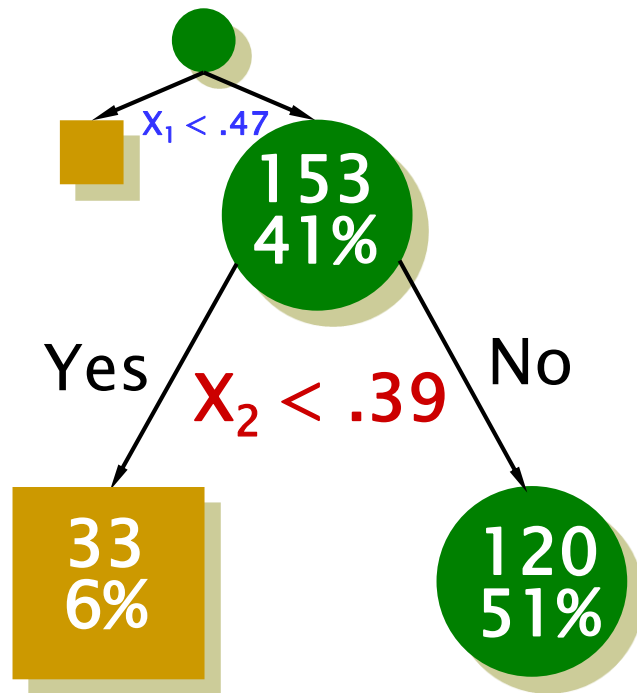


## ... <<사례>> 최적 분리

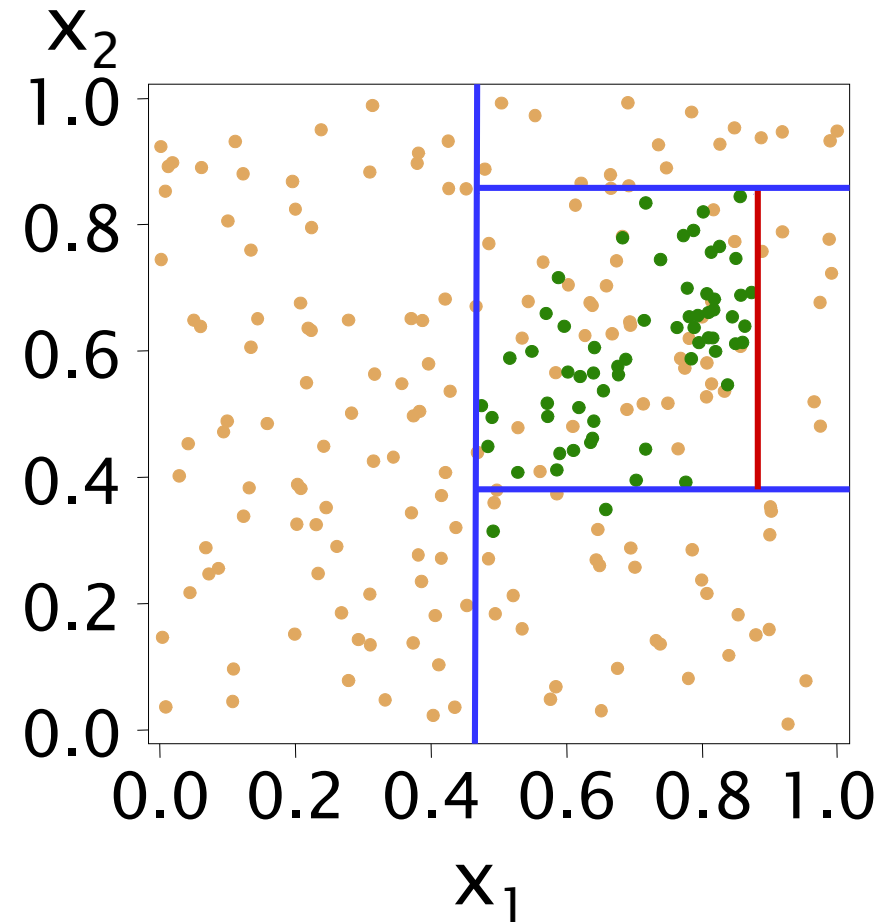
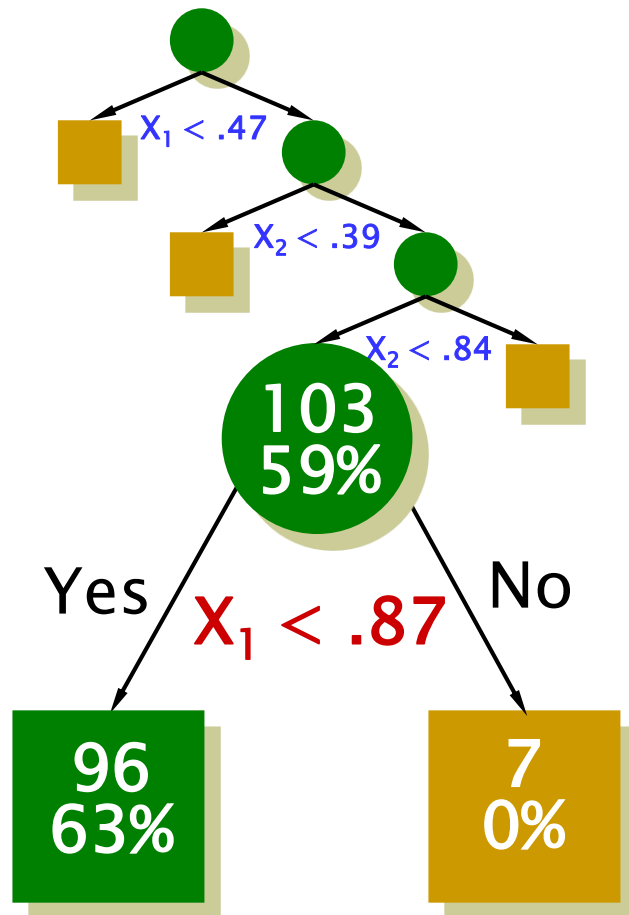




## ... <<사례>> 최적 분리



## ... <<사례>> 최적 분리



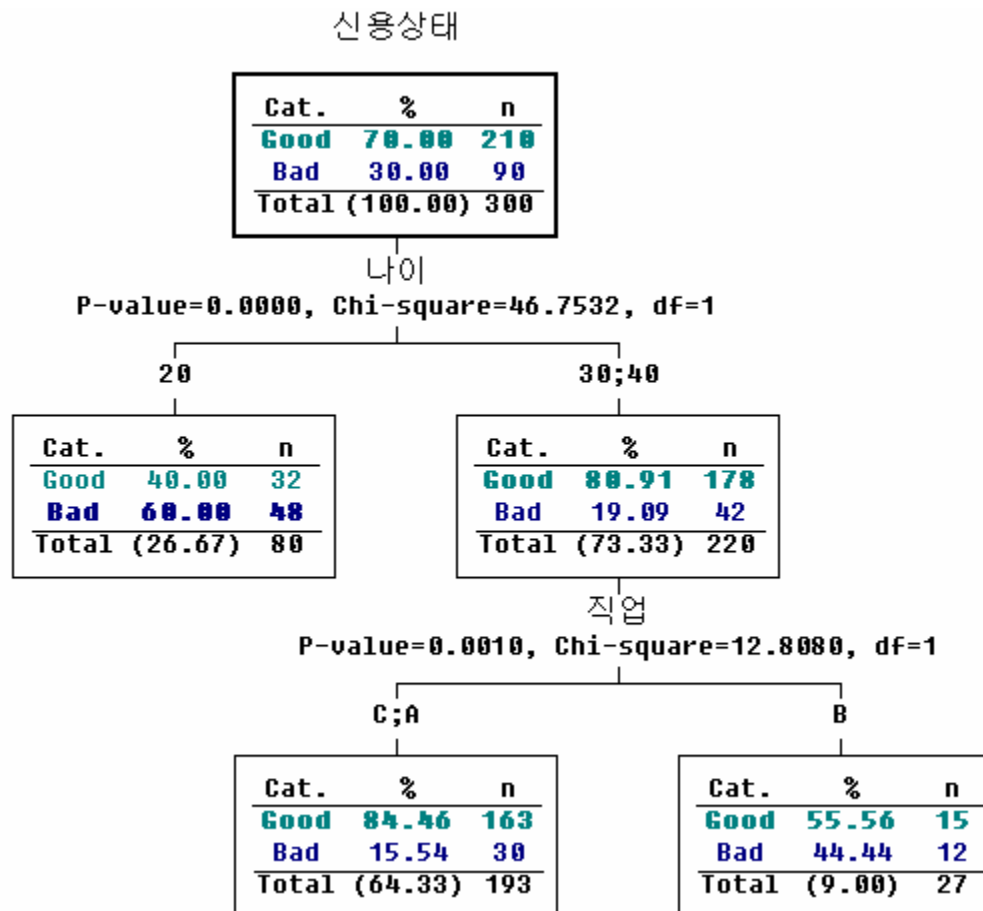
## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

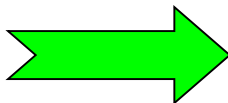
## 3.2.1 분류나무(Classification Tree)

- 목표변수: 이산형(범주형; 질적변수)

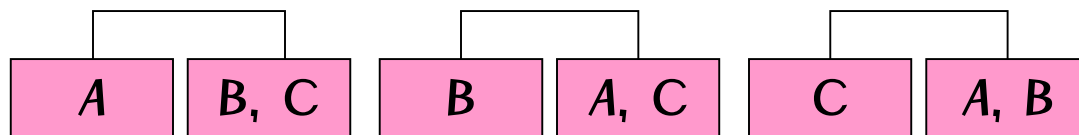
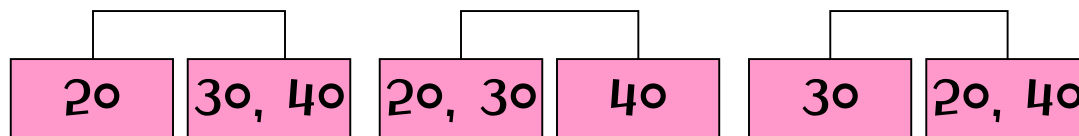


## 분류나무의 분리기준

ID	AGE	JOB	CREDIT
001	40	B	Good
002	20	A	Bad
003	20	C	Bad
004	30	A	Good
005	40	C	Good
...	...	...	...



Cat.	%	n
Good	70.00	210
Bad	30.00	90
Total (100.00)		300



300명 (Good: 210, Bad: 90)

AGE : 20대 미만, 30대, 40대 이상

JOB : A, B, C

Splitting Criteria

카이제곱 통계량의 p-값

지니 지수(Gini Index)

엔트로피 지수 (Entropy Index)

## 카이제곱 통계량의 $p$ -값

기대도수 ( $E_{ij}$ )

	Good	Bad	합계 ( $n_{i\cdot}$ )
20	56 (70.0)	24 (30.0)	80
30, 40	154 (70.0)	66 (30.0)	220
합계 ( $n_{\cdot j}$ )	210 (70.0)	90 (30.0)	330 (n)

실제도수 ( $O_{ij}$ )

	Good	Bad	합계 ( $n_{i\cdot}$ )
20	32 (40.0)	48 (60.0)	80
30, 40	178 (80.9)	42 (19.1)	220
합계 ( $n_{\cdot j}$ )	210 (70.0)	90 (30.0)	330 (n)

$$E_{ij} = n_{i\cdot} \cdot n_{\cdot j} / n$$

$$E_{11} = 80 \times 0.7 = 56$$

$$E_{12} = 80 \times 0.3 = 24$$

...

$$\begin{aligned} - \chi^2 &= \sum (E_{ij} - O_{ij})^2 / E_{ij} \\ &= (56-32)^2/56 + (24-48)^2/24 \\ &\quad + (154-178)^2/154 + (66-42)^2/66 = 46.75 \end{aligned}$$

$$- df = (r-1) \times (c-1) = (2-1) \times (2-1) = 1$$

$$- p\text{-value} = 0.00001$$

## ... 카이제곱 통계량의 $p$ -값

기대도수 ( $E_{ij}$ )

	Good	Bad	합계 ( $n_{i\cdot}$ )
20, 30	134 (70.0)	58 (30.0)	192
40	76 (70.0)	32 (30.0)	108
합계 ( $n_{\cdot j}$ )	210 (70.0)	90 (30.0)	330 (n)

실제도수 ( $O_{ij}$ )

	Good	Bad	합계 ( $n_{i\cdot}$ )
20, 30	124 (64.6)	68 (35.4)	192
40	86 (79.6)	22 (20.4)	108
합계 ( $n_{\cdot j}$ )	210 (70.0)	90 (30.0)	330 (n)

$$E_{ij} = n_{i\cdot} \cdot n_{\cdot j} / n$$

$$E_{11} = 192 \times 0.7 = 134$$

$$E_{12} = 192 \times 0.3 = 58$$

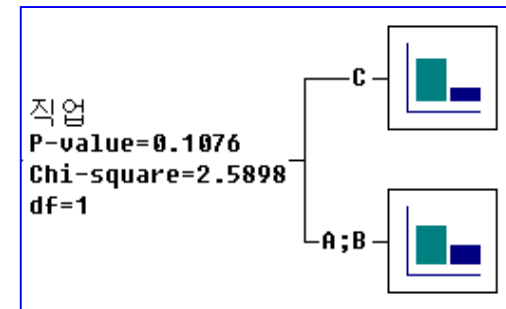
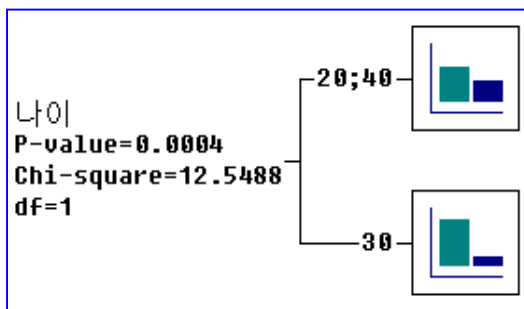
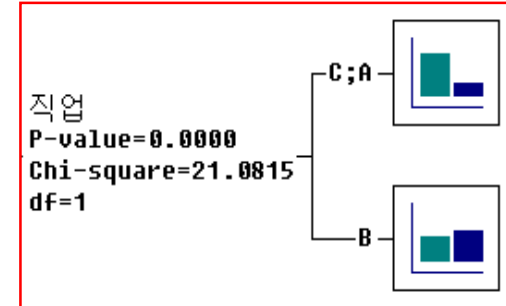
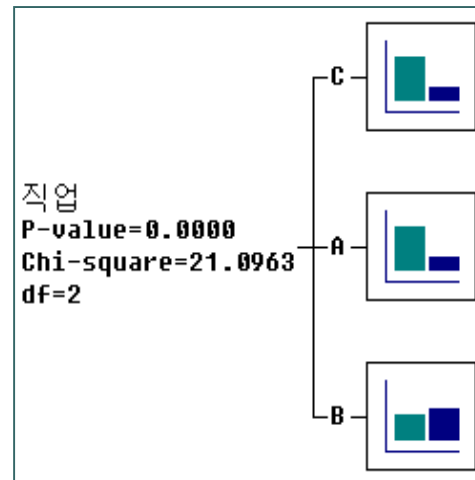
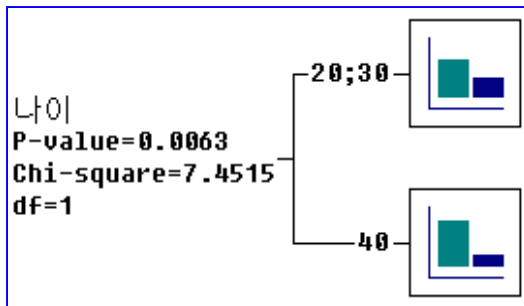
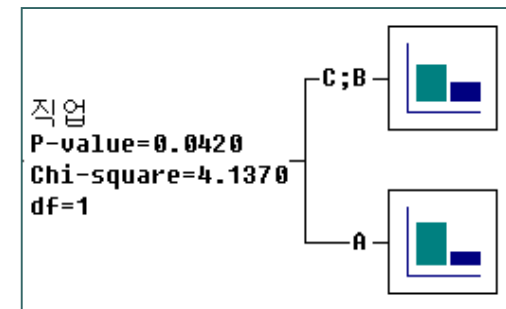
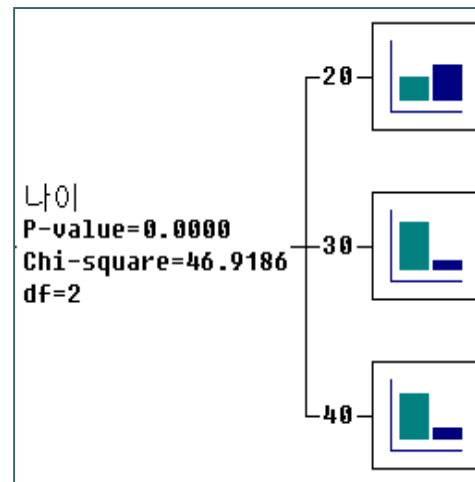
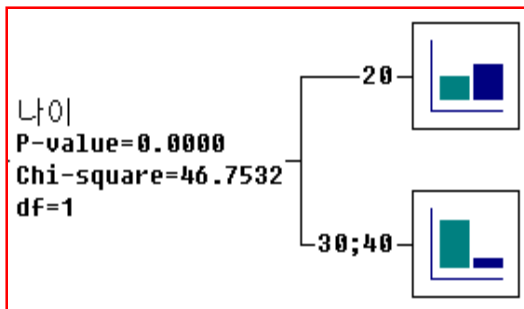
...

$$\begin{aligned} - \chi^2 &= \sum (E_{ij} - O_{ij})^2 / E_{ij} \\ &= (134-124)^2/134 + (58-68)^2/58 \\ &\quad + (76-86)^2/76 + (32-22)^2/32 = 7.45 \end{aligned}$$

$$- df = (r-1) \times (c-1) = (2-1) \times (2-1) = 1$$

$$- p\text{-value} = 0.0063$$

## ... 카이제곱 통계량의 $p$ -값





## 지니 지수(Gini index)

$$1 - \sum_{j=1}^r p_j^2 = 2 \sum_{j < k} p_j p_k$$

high diversity, low purity



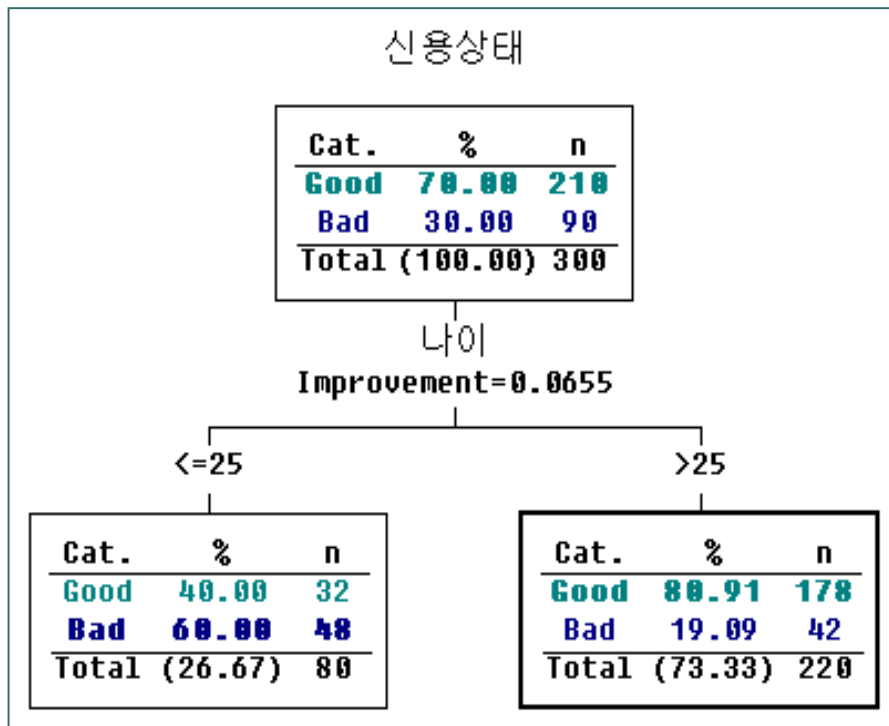
$$\text{Pr(interspecific encounter)} = 1 - 2(3/8)^2 - 2(1/8)^2 = .69$$

low diversity, high purity



$$\text{Pr(interspecific encounter)} = 1 - (6/7)^2 - (1/7)^2 = .24$$

## ... 지니 지수(Gini index)



$$G = 1 - (210/300)^2 - (90/300)^2 = .42$$

$$G = 1 - \sum (n_j/n_0)^2$$

$$\Delta G = G - G_L \times (n_L/n) - G_R \times (n_R/n)$$

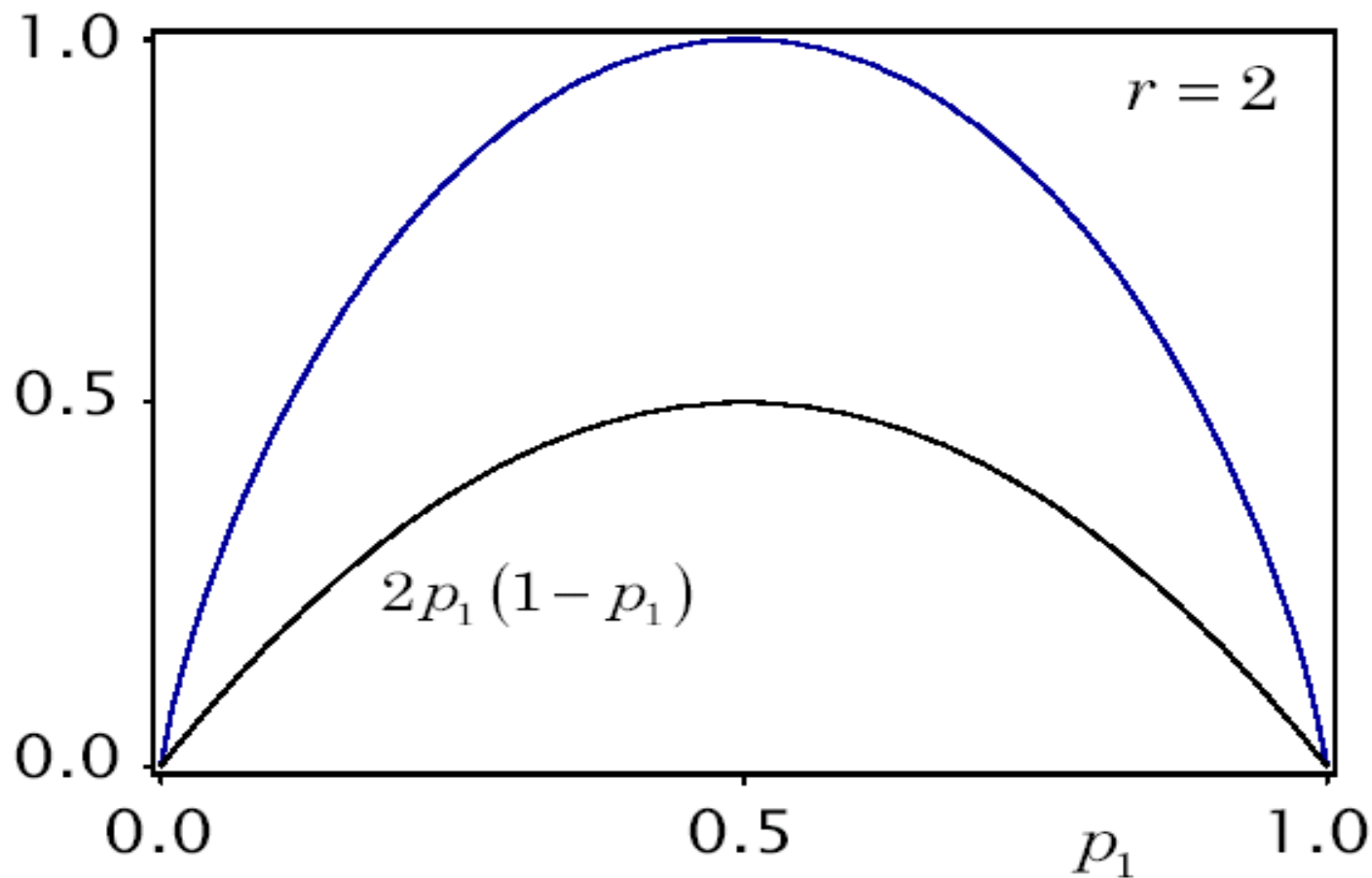
$$G_L = 1 - (32/80)^2 - (48/80)^2 = .48$$

$$G_R = 1 - (178/220)^2 - (42/220)^2 = .31$$

$$\text{Improvement} = 0.42 - 0.48 \times (80/300) - 0.31 \times (220/300) = 0.065$$

## 엔트로피 지수(Entropy index)

$$H(p_1, p_2, \dots, p_r) = -\sum_{i=1}^r p_i \log_2(p_i)$$



## 분류나무의 분리기준

---

- 카이제곱 통계량의  $p$ -값 : CHAID, Kass(1980)

- 지니 지수 (Gini Index) : CART, BFOS(1984)

$$\begin{aligned}\sum \sum P(j)P(j) &= \sum P(j)(1 - P(j)) = 1 - \sum P(j)^2 \\ &= 1 - \sum (n_j/n_0)^2\end{aligned}$$

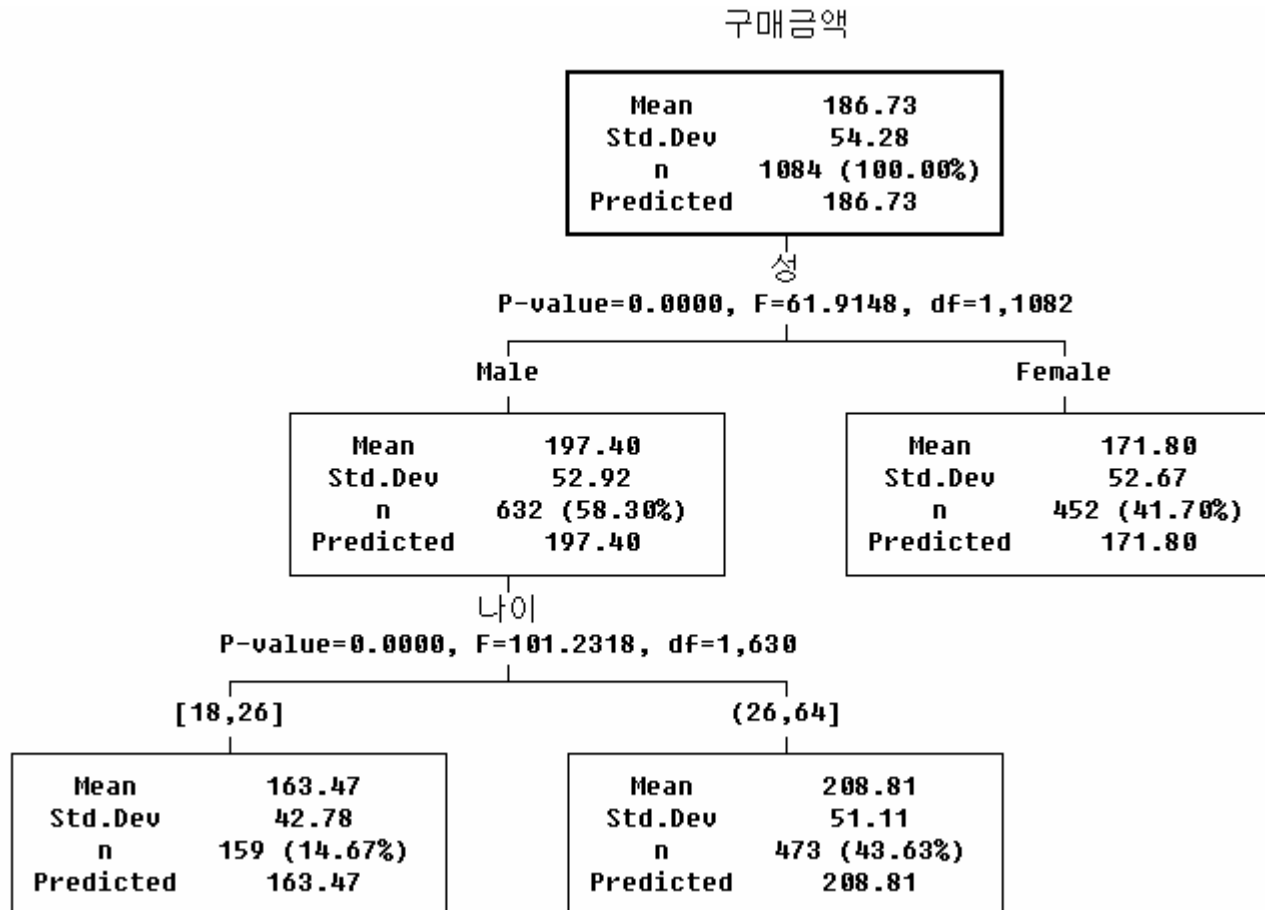
- 엔트로피 지수 (Entropy index) : C4.5, Quinlan(1993)

$$-\sum P(j) \log_2 P(j)$$

- ✓ 카이제곱 통계량이 지니 지수나 엔트로피 지수에 비해서 보다 단순한 형태의 나무구조를 가지게 하는 경향이 있음.

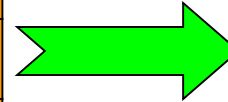
### 3.2.2 회귀나무(Regression Tree)

- 목표변수: 연속형(구간형; 양적변수)



## 회귀나무의 분리기준

ID	SEX	AGE	EDUC	SALES
001	F	25	12	122
002	M	47	12	161
003	F	49	6	214
004	F	36	12	207
005	M	29	15	183
...	...	...	...	...



구매금액

Mean	186.73
Std.Dev	54.28
n	1084 (100.00%)
Predicted	186.73

1084명

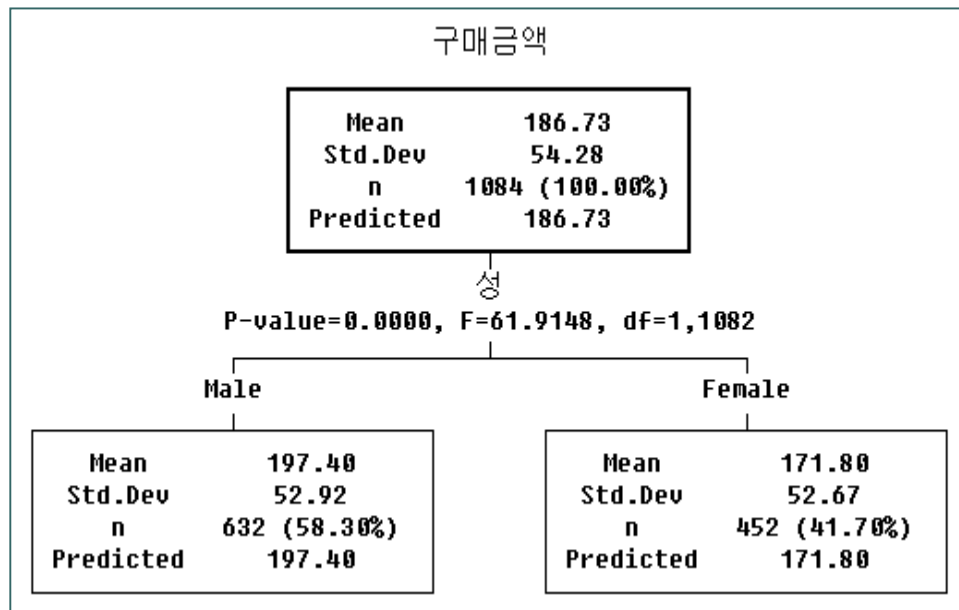
SEX : M, F

AGE : 18~64

EDUC : 6~18

Splitting Criteria { F-통계량  
분산의 감소량

## F-통계량의 p-값



$$\bar{y} = 186.73, \quad n = 1084$$

$$\bar{y}_1 = 197.40, \quad n_1 = 632$$

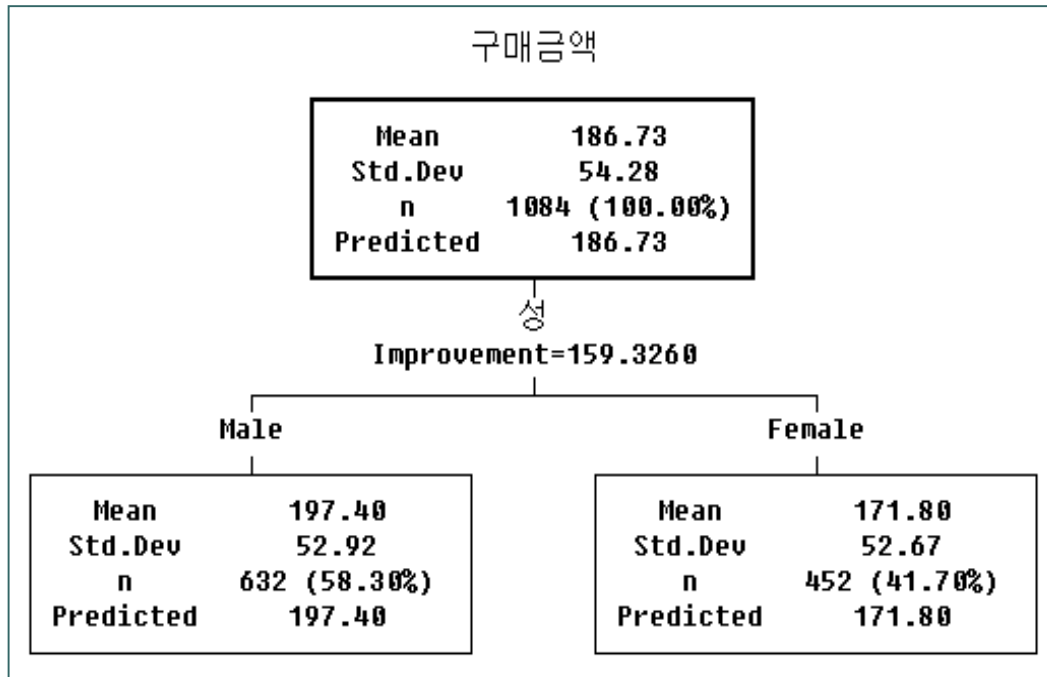
$$\bar{y}_2 = 171.80, \quad n_2 = 452$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

전체제곱합(TSS) = 처리제곱합(SST) + 오차제곱합(SSE)

$$F = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(n-k)}$$

## 분산의 감소량(Variance reduction)



$$V = \sum (y_j - \bar{y})^2 / n = 54.28^2$$

$$V_L = 52.92^2$$

$$V_R = 52.67^2$$

$$\Delta V = V - V_L \times (n_L/n) - V_R \times (n_R/n)$$

$$= 54.28^2 - 52.92^2 \times (632/1084) - 52.67^2 \times (452/1084)$$

$$= 159.33$$



## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

### 3.3.1 의사결정나무분석의 장점

- 해석의 용이성

- ✓ 나무구조에 의해서 모형이 표현되기 때문에 모형을 사용자가 쉽게 이해할 수 있다.
- ✓ 새로운 개체에 대한 분류 또는 예측을 하기 위해서 뿌리마디로부터 끝마디까지를 단순히 따라가면 되기 때문에, 새로운 자료에 모형을 적합시키기가 매우 쉽다.
- ✓ 나무구조로부터 어떤 입력변수가 목표변수를 설명하기 위해서 더 중요한지를 쉽게 파악할 수 있다.

- 상호작용 효과의 해석

- ✓ 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 알 수 있다.
- ✓ 의사결정나무는 유용한 입력변수나 상호작용(interaction)의 효과 또는 비선형성(nonlinearity)을 자동적으로 찾아내는 알고리즘이라고 할 수 있다.

- 비모수적 모형

- ✓ 의사결정나무는 선형성(linearity)이나 정규성(normality) 또는 등분산성(equal variance) 등의 가정을 필요로 하지 않는 비모수적인(nonparametric) 방법이다.
- ✓ 의사결정나무에서는 순서형 또는 연속형 변수는 단지 순위(rank)만 분석에 영향을 주기 때문에 이상치(outlier)에 민감하지 않다는 장점을 가지고 있다.

### 3.3.2 의사결정나무분석의 단점

- 비연속성

- ✓ 의사결정나무에서는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근처에서는 예측오류가 클 가능성이 있다.
- ✓ 최근에는 이러한 단점을 극복하기 위하여, 앞서 논의한 장점을 해치지 않고 모수적 모형이나 신경망 등을 의사결정나무와 결합하는 방법들이 연구되고 있다.

- 선형성 또는 주효과의 결여

- ✓ 회귀모형에서는 회귀계수나 오즈비(odds ratio)를 이용하여 결과에 대한 유용한 해석을 얻을 수 있다. 즉, 선형모형(linear model)에서 주효과(main effect)는 다른 예측변수와 관련시키지 않고서도 각 변수의 영향력을 해석할 수 있다는 장점을 가지고 있는데 의사결정나무에서는 선형(linear) 또는 주효과(main effect) 모형에서와 같은 결과를 얻을 수 없다는 한계점이 있다.

- 불안정성

- ✓ 분석용 자료(training data)에만 의존하는 의사결정나무는 새로운 자료의 예측에서는 불안정(unstable)할 가능성이 높다. 이와 같은 현상은 분석용 자료의 크기가 너무 작은 경우와 너무 많은 가지를 가지는 의사결정나무를 얻는 경우에 빈번히 발생한다.
- ✓ 따라서 검증용 자료(test data)에 의한 교차타당성(cross validation) 평가나 가지치기에 의해서 안정성 있는 의사결정나무를 얻는 것이 바람직하다.

## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

### 3.4.1 분석흐름도 작성과 변수 탐색

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter3\_1

데이터 분할  
노드의  
속성 패널

속성	
일반	
노드 ID	Part
가져온 데이터	...
내보낸 데이터	...
노드	...
분석	
변수	...
출력 유형	데이터
분할 방법(Partitioning M기분	
난수초기값	12345
데이터셋 할당	
분석용(Training)	70.0
평가용(Validation)	30.0
검증용(Test)	0.0
리포트	
Interval 타겟	예
Class 타겟	예
상태	
생성 시간	14, 2, 9 오후 2:43
실행 ID	2ffe32ea-f0d1-4e4e-8ee
최근 오류	
최근 상태	완료
최근 실행 시간	14, 2, 9 오후 2:46
일반	
실행 완료	

그림 탐색(Graph...)

HMEQ

등거량 탐색(...)

멀티플롯(Multi Plot)

데이터 분할(Data...)

결측값 처리(Impute)

의사결정트리(Decision Tree...)

회귀(Regression)

신경망(Neural Network)

모델비교(Model Comparison)

리포트 생성(Reporter)

왼쪽에서 오른쪽으로  
위에서 아래로

타이머그램  
탐색 툴바

hckang(으)로서의 hckang hckang-pc에 연결

# 변수 편집 메뉴

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter3\_1

그림 탐색(Graph...) HMEQ 통계량 탐색(...) 데이터 분할(Data...) 의사결정트리(Decision Tree...) 신경망(Neural network) 보고서 생성(Reporter)

**데이터 노드의 속성 패널**

속성	
<b>일반</b>	
노드 ID	lds
가져온 데이터	...
내보낸 데이터	...
노드	...
<b>분석</b>	
출력 유형	뷰
역할	Raw
재실행	아니요
요약	아니요
Map 변수 제거(Drop Me)	
<b>칼럼</b>	
변수	...
의사결정(Decisions)	...
메타데이터 새로 고침(R)	...
관리자(Advisor)	기본
고급 옵션(Advanced Op)	...
<b>데이터</b>	
데이터 선택	데이터 소스
표본	기본
표본추출 옵션	...
<b>데이터 소스</b>	
데이터 소스	HMEQ
<b>일반</b>	
실행 완료	

변수 편집...

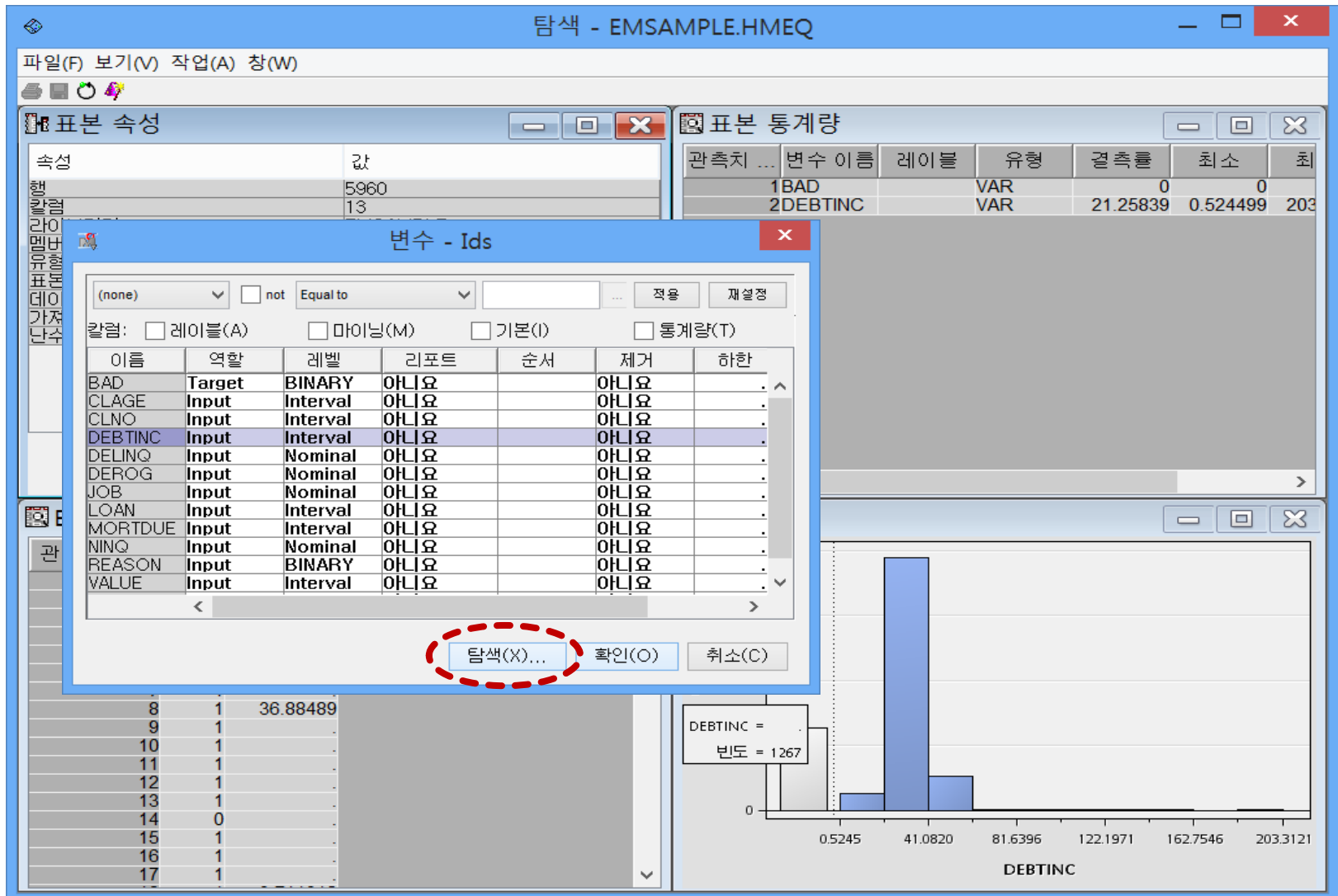
- 업데이트
- 실행
- 모델 패키지 생성...
- 결과...
- 경로를 SAS 프로그램으로 내보내기
- 자르기
- 복사(C)
- 삭제
- 이름 바꾸기
- 모두 선택
- 노드 선택
- 노드 연결
- 노드 연결 해제

리포트 생성(Reporter)

다이어그램 | 로그

hckang(으)로서의 hckang hckang-pc에 연결

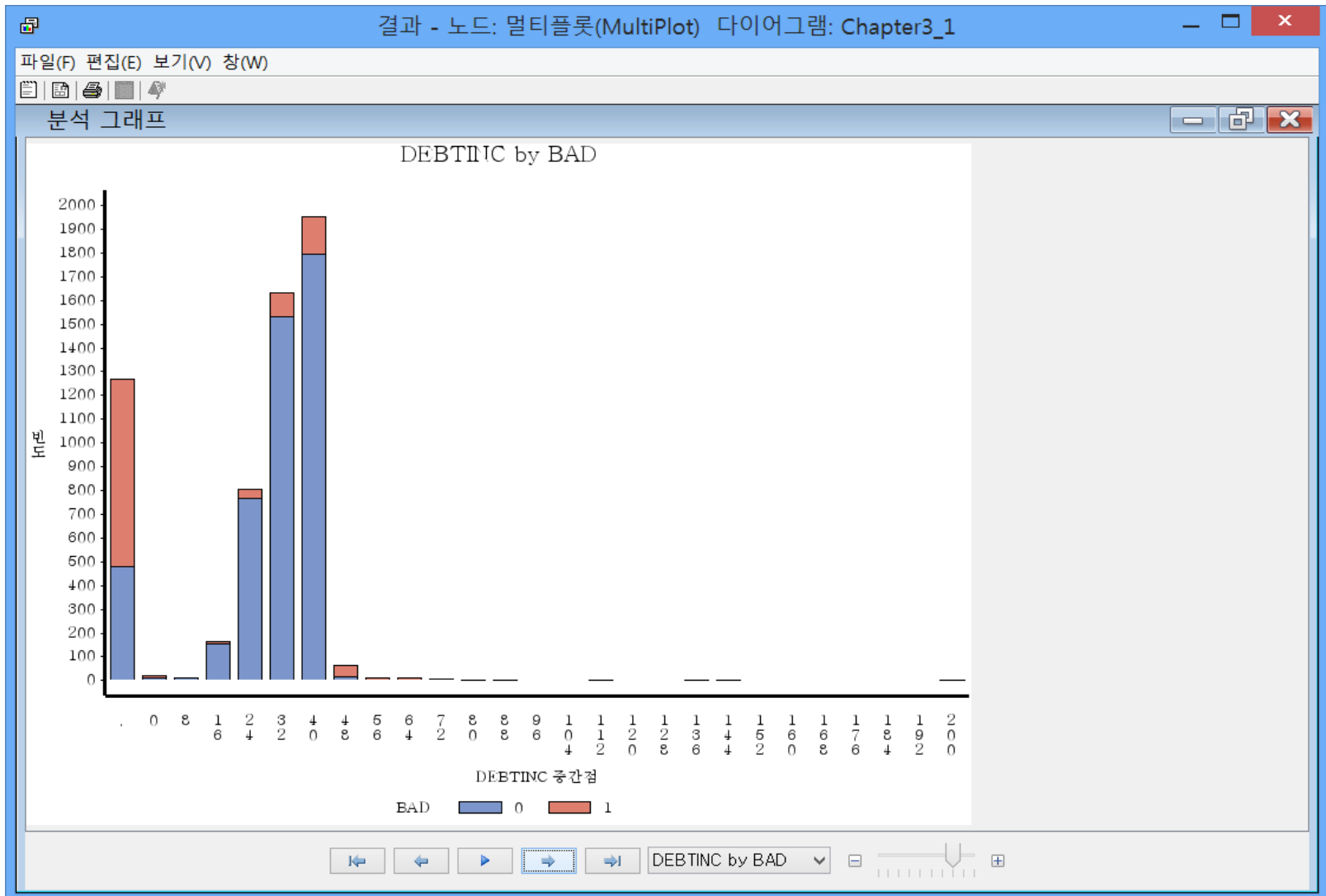
# 변수들의 분포에 대한 탐색



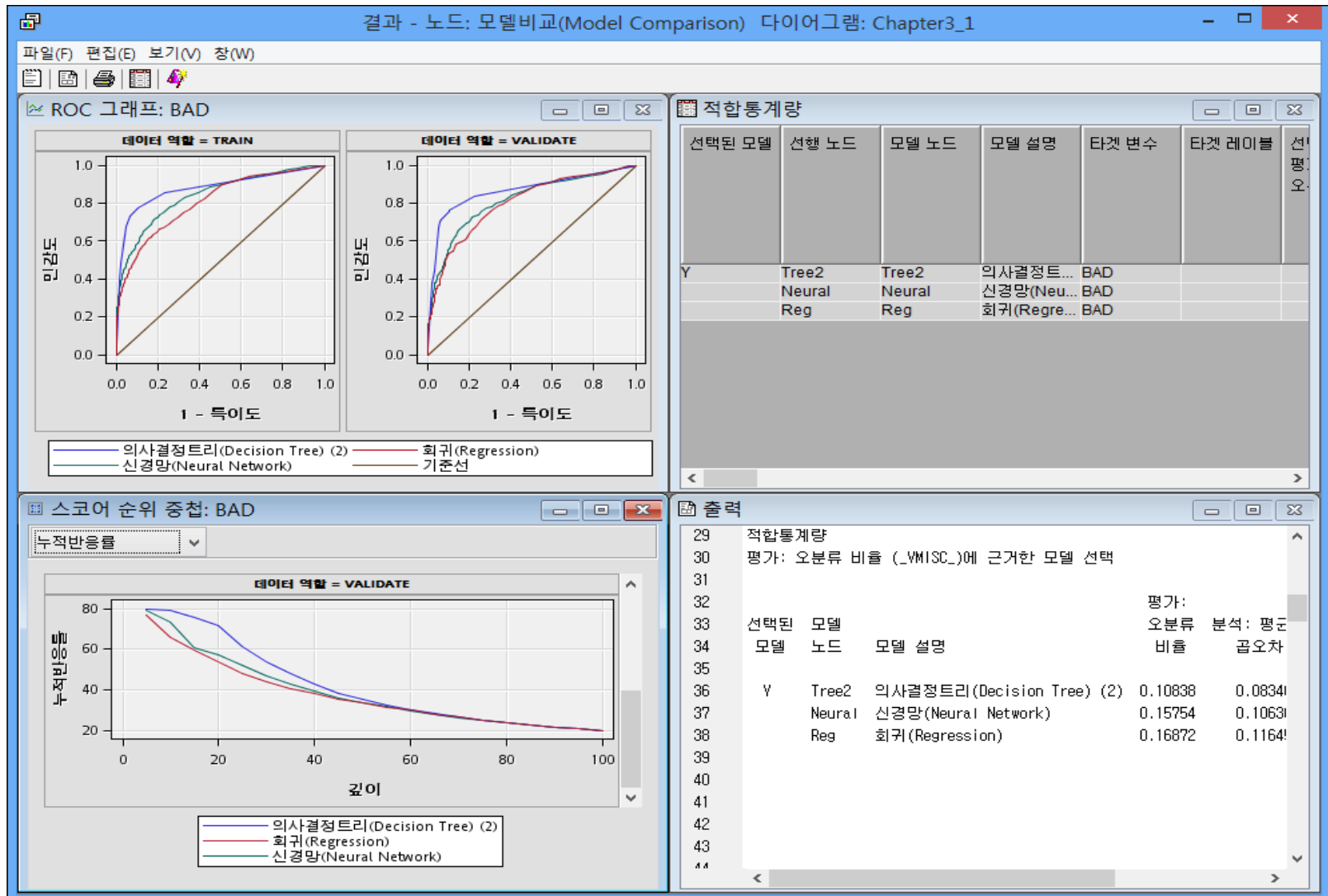




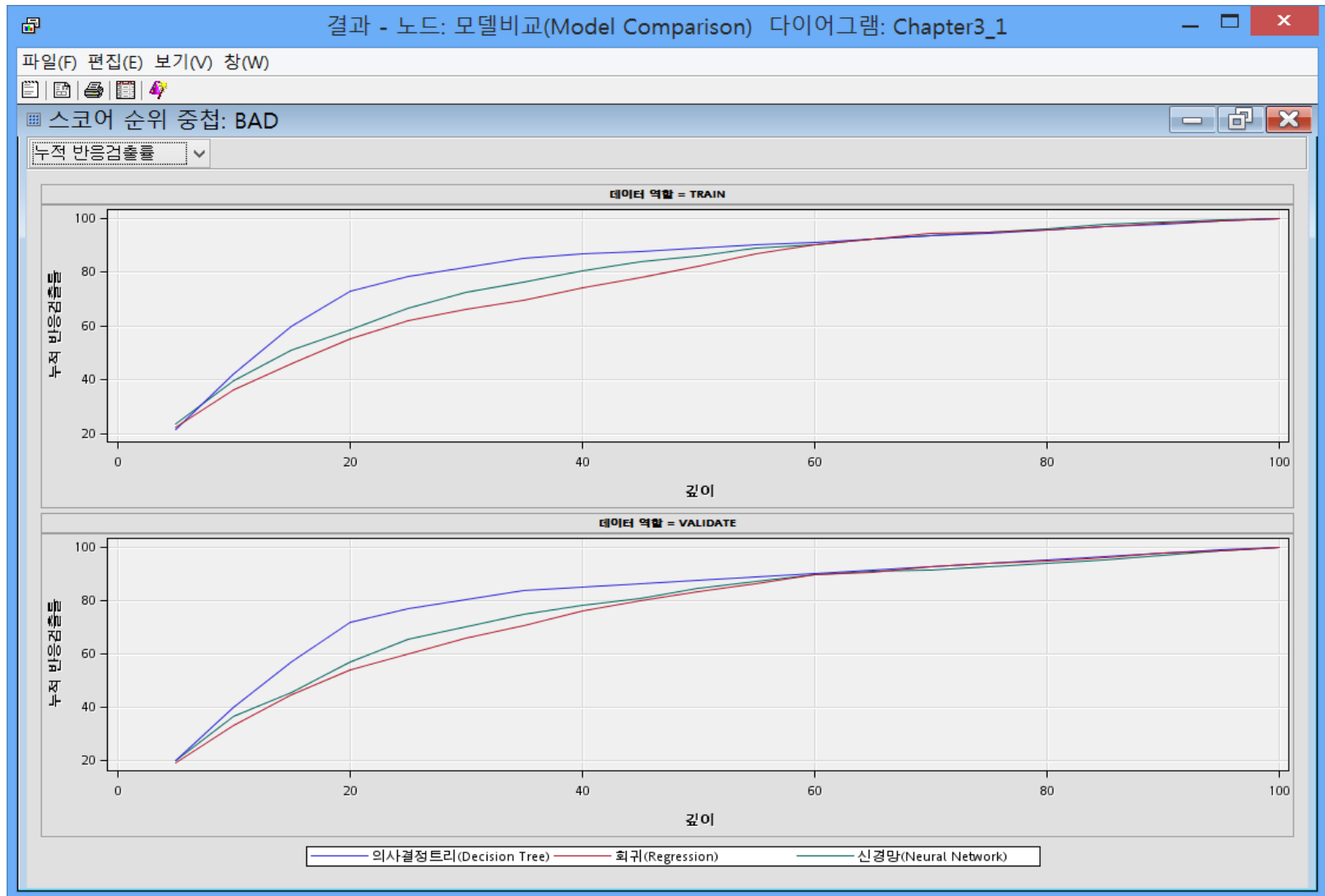
# 멀티플롯(MultiPlot) 노드 - 결과



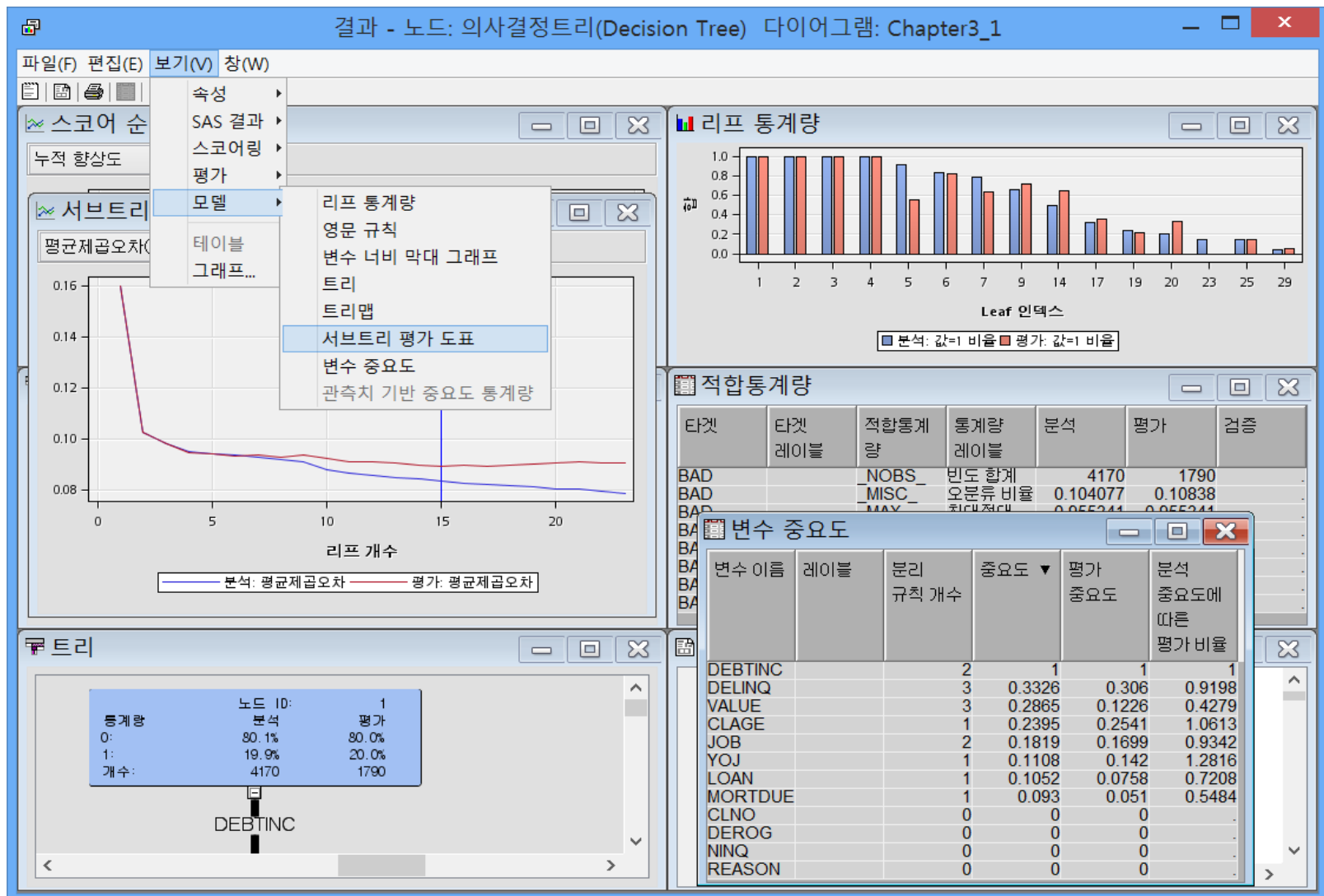
## 3.4.2 모형 평가와 결과 보기



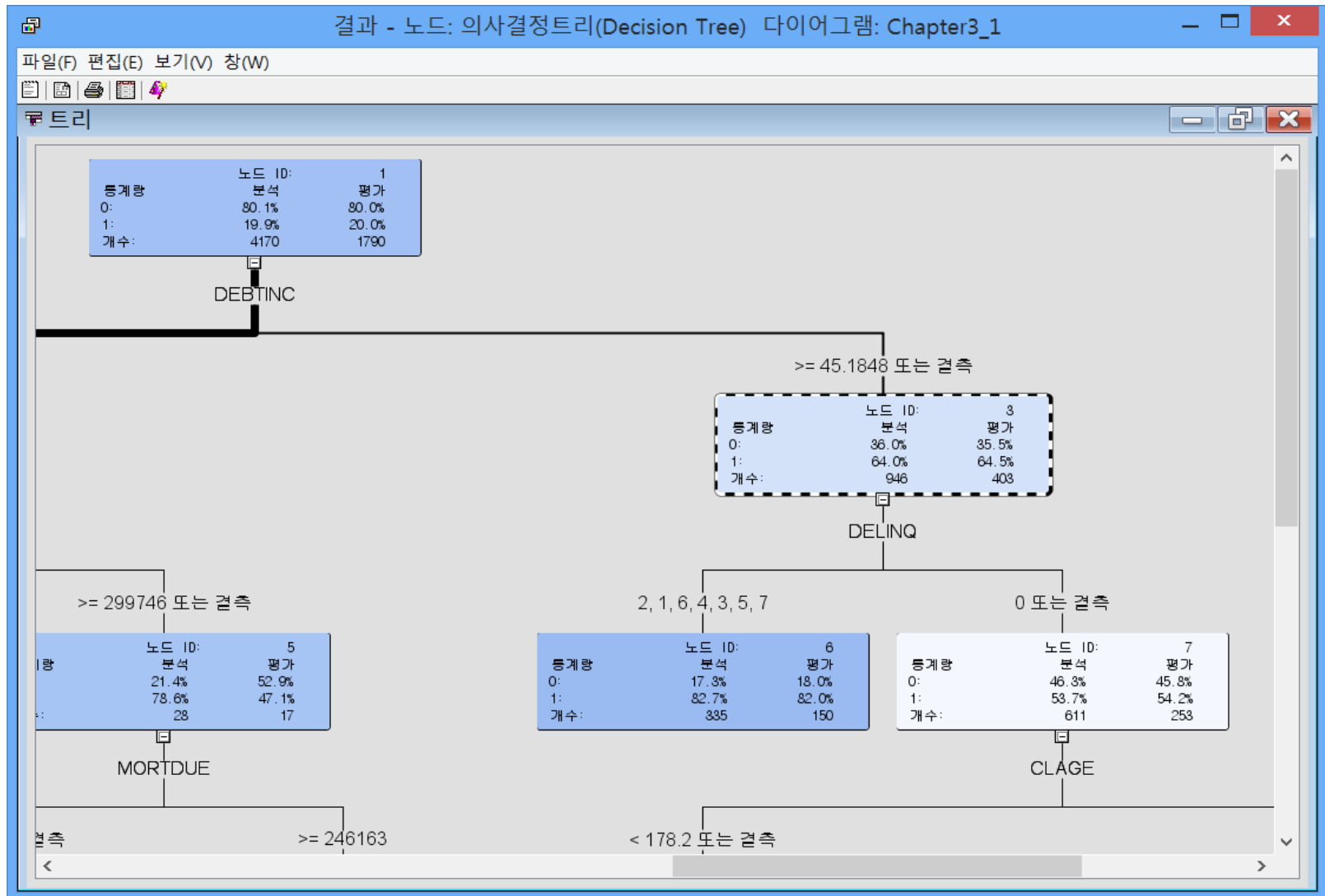
# 모델비교(Model Comparison) 노드 - 결과



# 의사결정트리(Decision Tree) 노드 - 결과



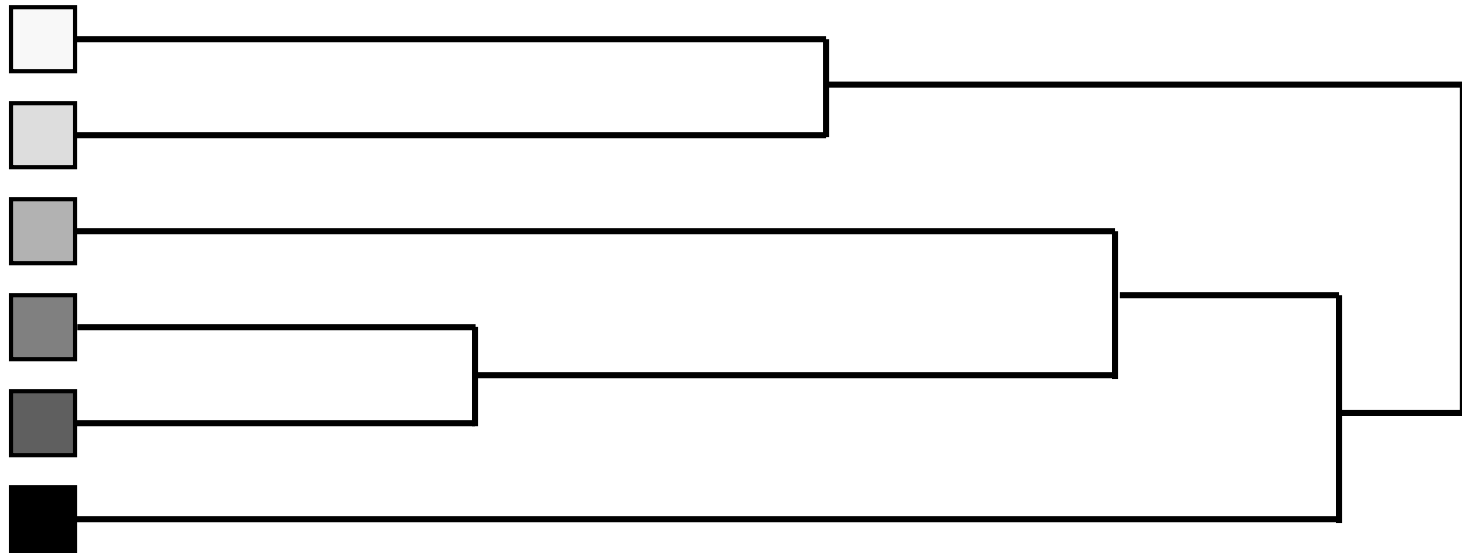
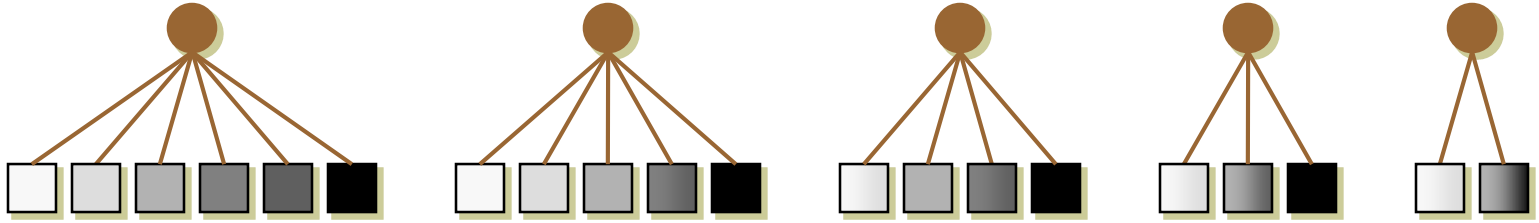
# 의사결정트리(Decision Tree) 노드 - 결과: 트리



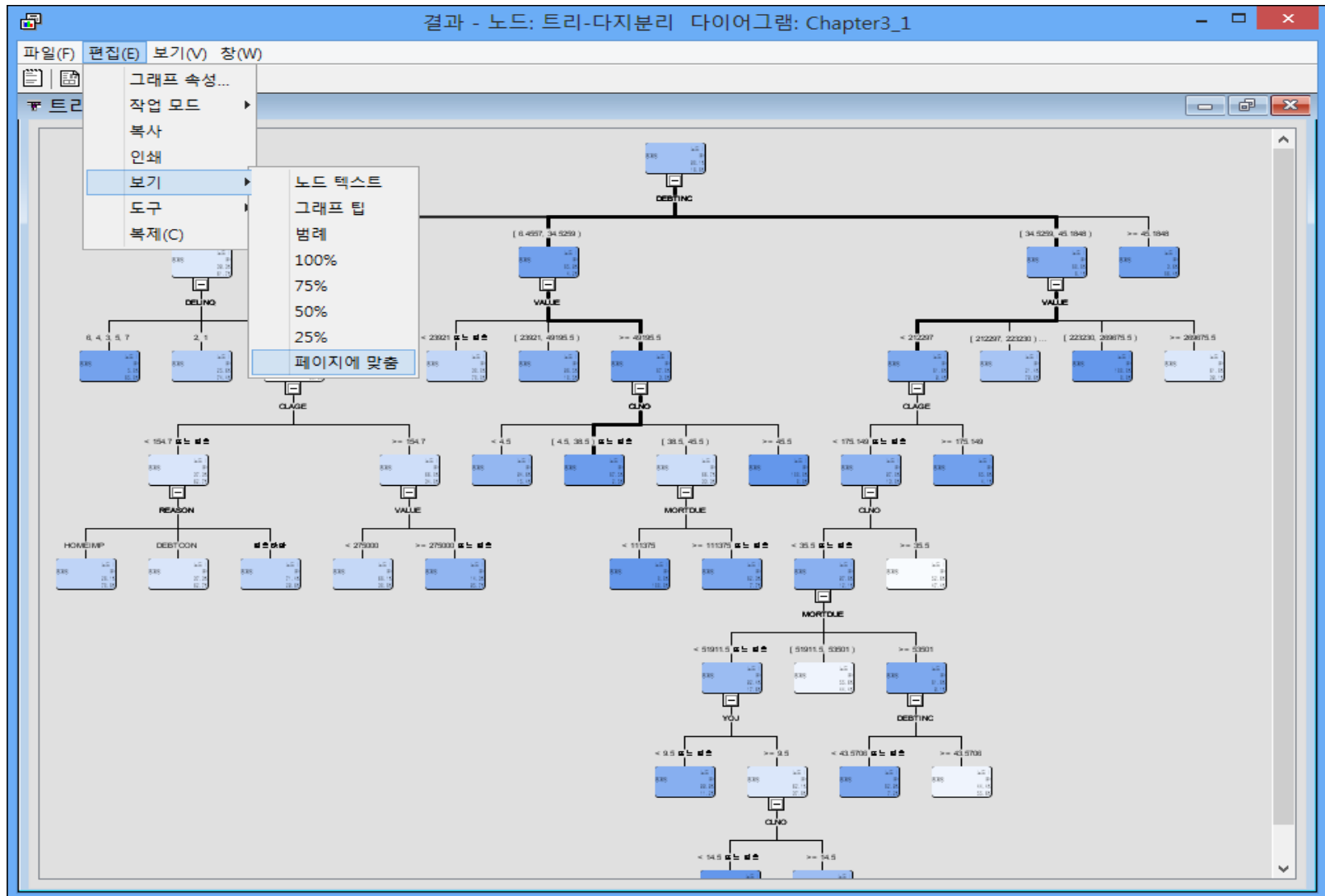
### 3.4.3 의사결정나무의 수정

The screenshot displays the Enterprise Miner - DM Project interface. On the left, a tree view shows the project structure with 'Chapter3\_1' selected. A red dashed circle highlights the '의사결정트리' (Decision Tree) node in the tree view, with a callout bubble containing the text '의사결정트리 노드의 속성 패널' (Decision Tree Node's Property Panel). Below the tree view, the '속성' (Properties) panel is visible, showing various settings for the decision tree, including '분리 규칙' (Splitting Rule) and '노드' (Nodes). A red dashed circle highlights the '최대 가지' (Maximum Branch) property, which is set to 4. In the main workspace, a flowchart diagram shows the data processing workflow. A context menu is open over the '의사결정트리' node, with the '이름 바꾸기' (Rename) option selected. A red arrow points from this option to a '이름 바꾸기' (Rename) dialog box. The dialog box has a text field containing '트리-다지분리' and buttons for '확인(O)' (OK) and '취소(C)' (Cancel).

## 다지분리(Multiway Splits)

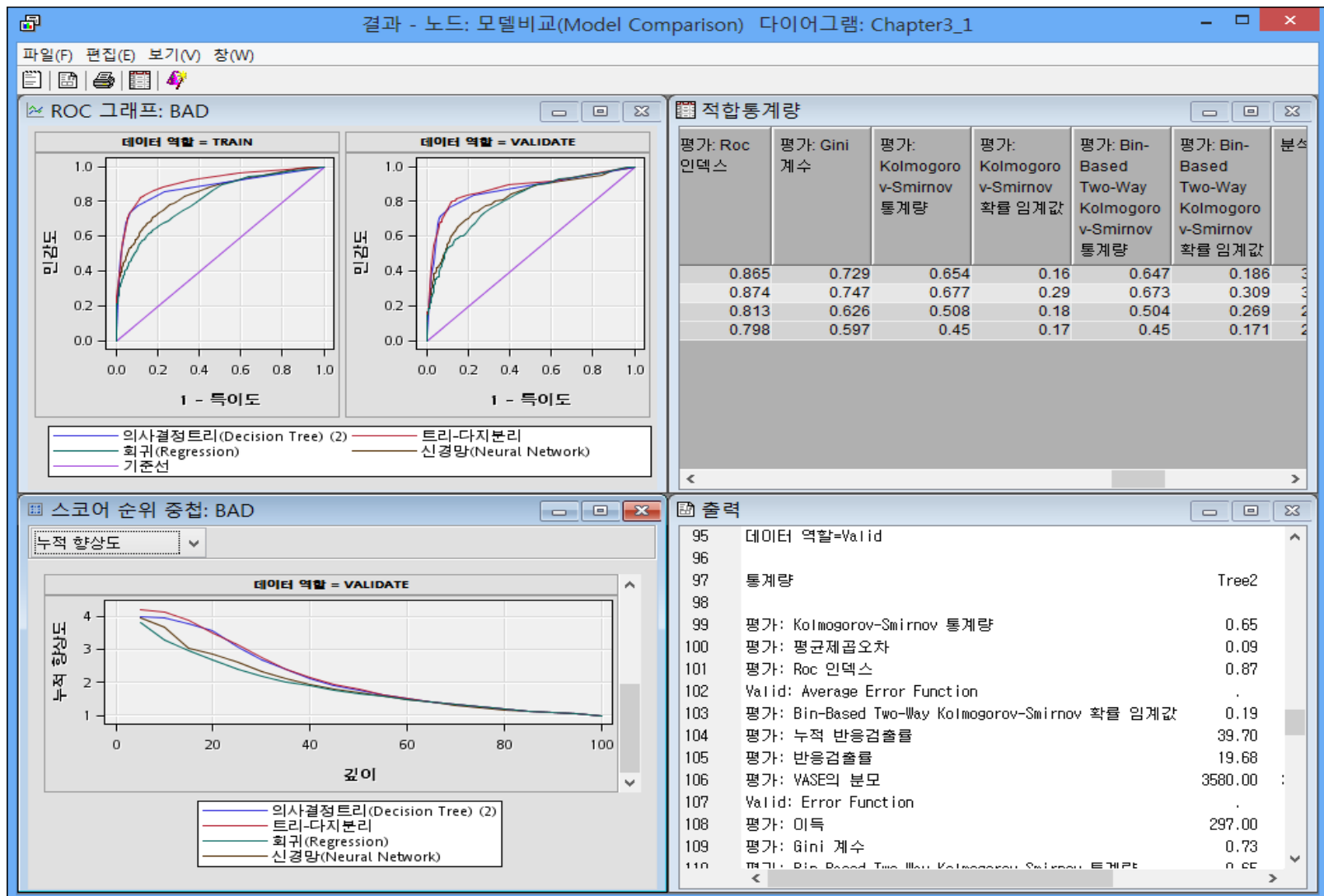


# 의사결정트리 노드 - 결과: 트리(다지분리 적용)

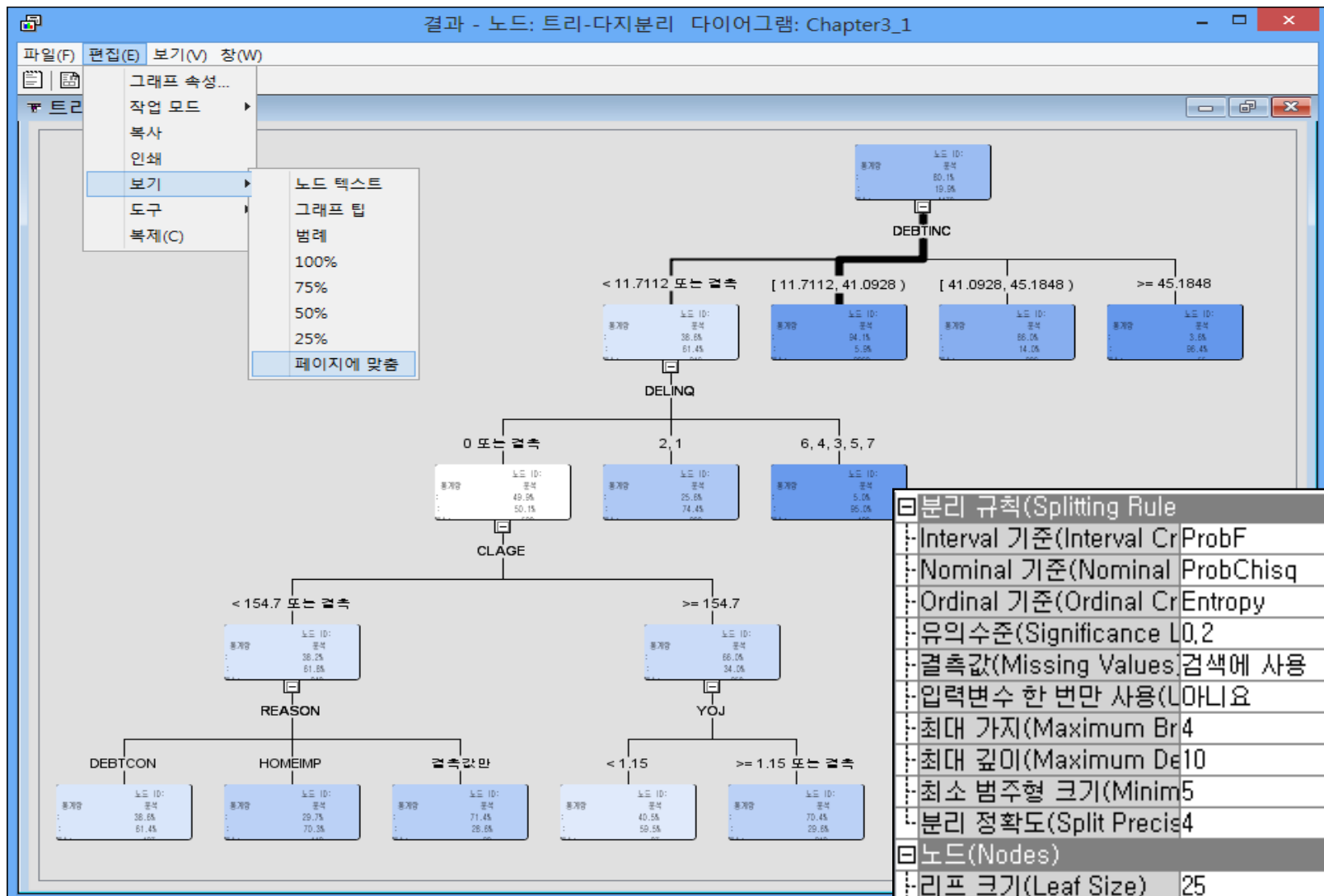




# 모델 비교(Model Comparison) 노드 - 결과

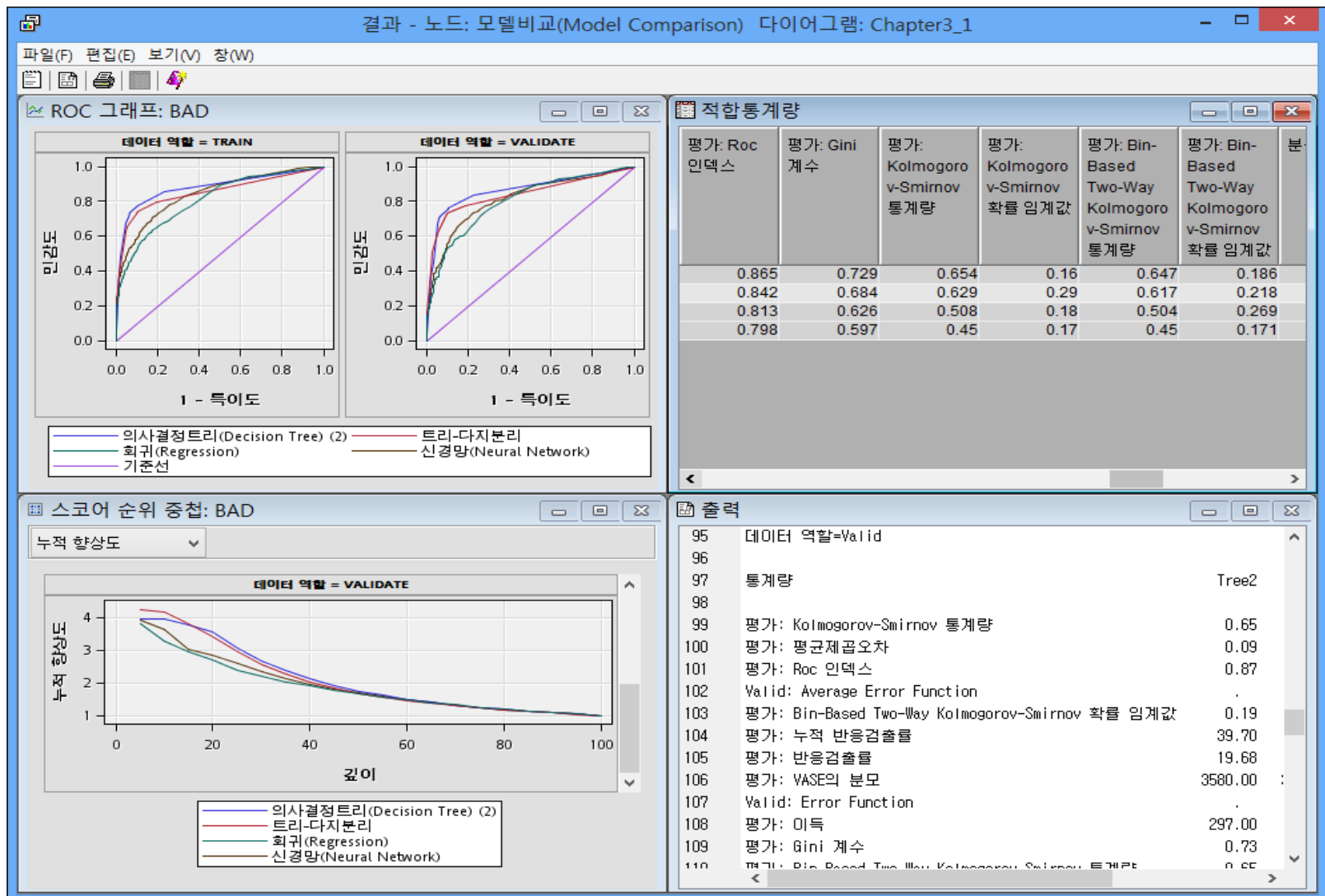


# 정지규칙의 설정(사전 가지치기)



분리 규칙(Splitting Rule)	
Interval 기준(Interval Cr	ProbF
Nominal 기준(Nominal	ProbChisq
Ordinal 기준(Ordinal Cr	Entropy
유의수준(Significance L	0.2
결측값(Missing Values)	검색에 사용
입력변수 한 번만 사용(L	아니요
최대 가지(Maximum Br	4
최대 깊이(Maximum De	10
최소 범주형 크기(Minim	5
분리 정확도(Split Preci	4
노드(Nodes)	
리프 크기(Leaf Size)	25
규칙 개수(Number of Ru	5
대체 규칙 수(Number of	0
분리 크기(Split Size)	100

# 모델 비교(Model Comparison) 노드 - 결과



## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

# 분석사례 - 2를 위한 다이어그램

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

Chapter3\_2

DM Project

- 데이터 소스
- 다이어그램
  - Chapter2
  - Chapter3\_1
  - Chapter3\_2
  - Chapter4\_1
  - Chapter4\_2
  - Chapter4\_3
  - Chapter4\_4
  - Chapter4\_5
  - Chapter4\_6
  - Chapter4\_7
  - Chapter4\_8
  - Chapter4\_9
  - Chapter4\_10
  - Chapter4\_11
  - Chapter4\_12
  - Chapter4\_13
  - Chapter4\_14
  - Chapter4\_15
  - Chapter4\_16
  - Chapter4\_17
  - Chapter4\_18
  - Chapter4\_19
  - Chapter4\_20
  - Chapter4\_21
  - Chapter4\_22
  - Chapter4\_23
  - Chapter4\_24
  - Chapter4\_25
  - Chapter4\_26
  - Chapter4\_27
  - Chapter4\_28
  - Chapter4\_29
  - Chapter4\_30
  - Chapter4\_31
  - Chapter4\_32
  - Chapter4\_33
  - Chapter4\_34
  - Chapter4\_35
  - Chapter4\_36
  - Chapter4\_37
  - Chapter4\_38
  - Chapter4\_39
  - Chapter4\_40
  - Chapter4\_41
  - Chapter4\_42
  - Chapter4\_43
  - Chapter4\_44
  - Chapter4\_45
  - Chapter4\_46
  - Chapter4\_47
  - Chapter4\_48
  - Chapter4\_49
  - Chapter4\_50
  - Chapter4\_51
  - Chapter4\_52
  - Chapter4\_53
  - Chapter4\_54
  - Chapter4\_55
  - Chapter4\_56
  - Chapter4\_57
  - Chapter4\_58
  - Chapter4\_59
  - Chapter4\_60
  - Chapter4\_61
  - Chapter4\_62
  - Chapter4\_63
  - Chapter4\_64
  - Chapter4\_65
  - Chapter4\_66
  - Chapter4\_67
  - Chapter4\_68
  - Chapter4\_69
  - Chapter4\_70
  - Chapter4\_71
  - Chapter4\_72
  - Chapter4\_73
  - Chapter4\_74
  - Chapter4\_75
  - Chapter4\_76
  - Chapter4\_77
  - Chapter4\_78
  - Chapter4\_79
  - Chapter4\_80
  - Chapter4\_81
  - Chapter4\_82
  - Chapter4\_83
  - Chapter4\_84
  - Chapter4\_85
  - Chapter4\_86
  - Chapter4\_87
  - Chapter4\_88
  - Chapter4\_89
  - Chapter4\_90
  - Chapter4\_91
  - Chapter4\_92
  - Chapter4\_93
  - Chapter4\_94
  - Chapter4\_95
  - Chapter4\_96
  - Chapter4\_97
  - Chapter4\_98
  - Chapter4\_99
  - Chapter4\_100
- 모델

그래프 탐색  
노드의  
속성 패널

속성

일반

노드 ID	GrExpl
가져온 데이터	
내보낸 데이터	
노트	

분석

변수

표본 속성(Sample Prop)

방법	First N
크기	기본
난수초기값	12345

리포트

타겟	예
타겟 별 그룹	예

상태

생성 시간	14, 2, 9 오후 1:06
실행 ID	
최근 오류	
최근 상태	
최근 실행 시간	
실행 기간	
그리드 호스트	

일반

일반 속성

실행 완료

다이어그램

로그

변수 편집...

- 업데이트
- 실행
- 모델 패키지 생성...
- 결과...
- 경로를 SAS 프로그램으로 내보내기
- 자르기
- 복사(C)
- 삭제
- 이름 바꾸기
- 모두 선택
- 노드 선택
- 노드 연결
- 노드 연결 해제

WAGES

데이터 분할(Data...)

의사결정트리(Decision Tree)

회귀(Regression)

모델비교(Model Comparison)

그래프 탐색(Graph...)

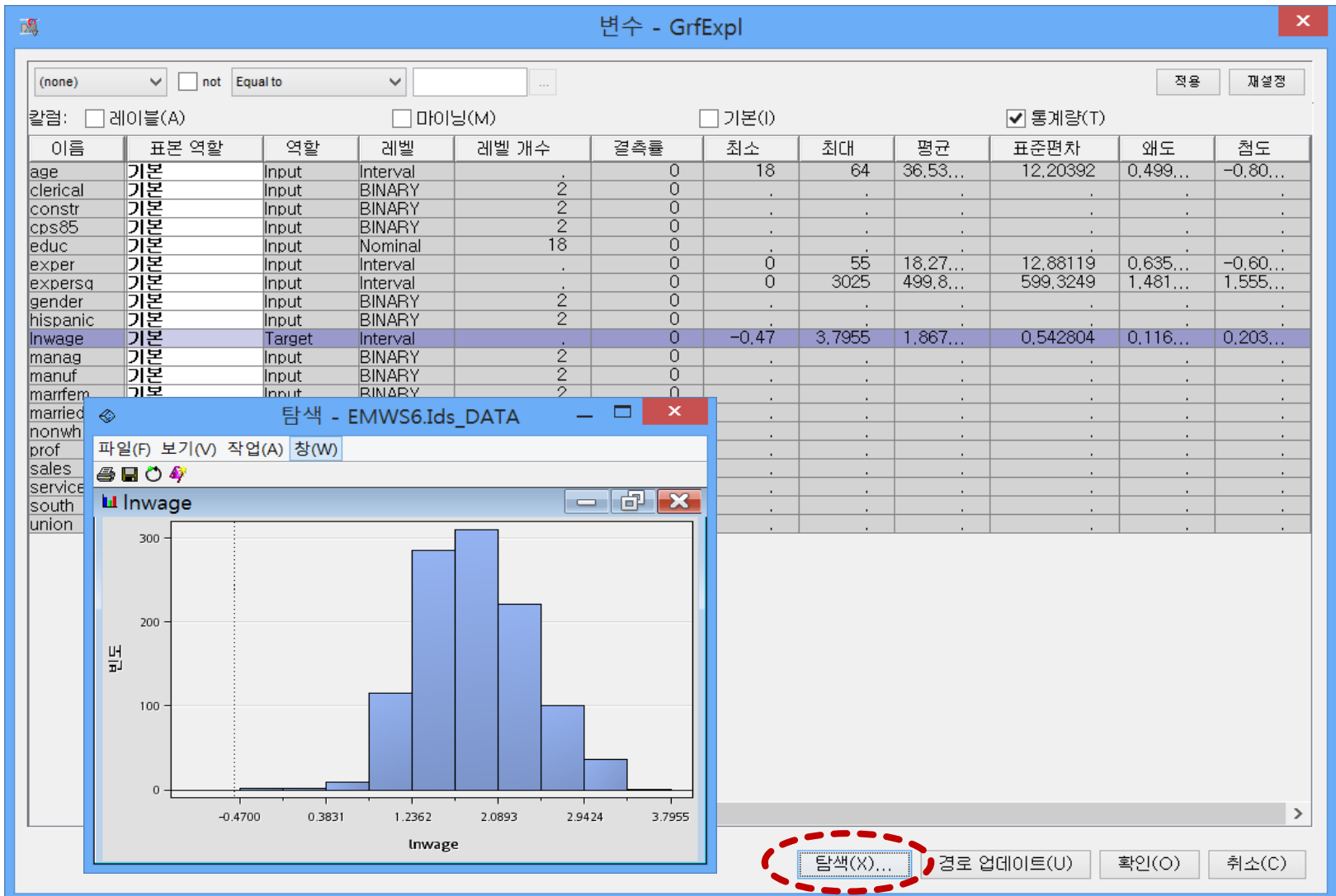
등계량 탐색(...)

멀티플롯(Multi Plot)

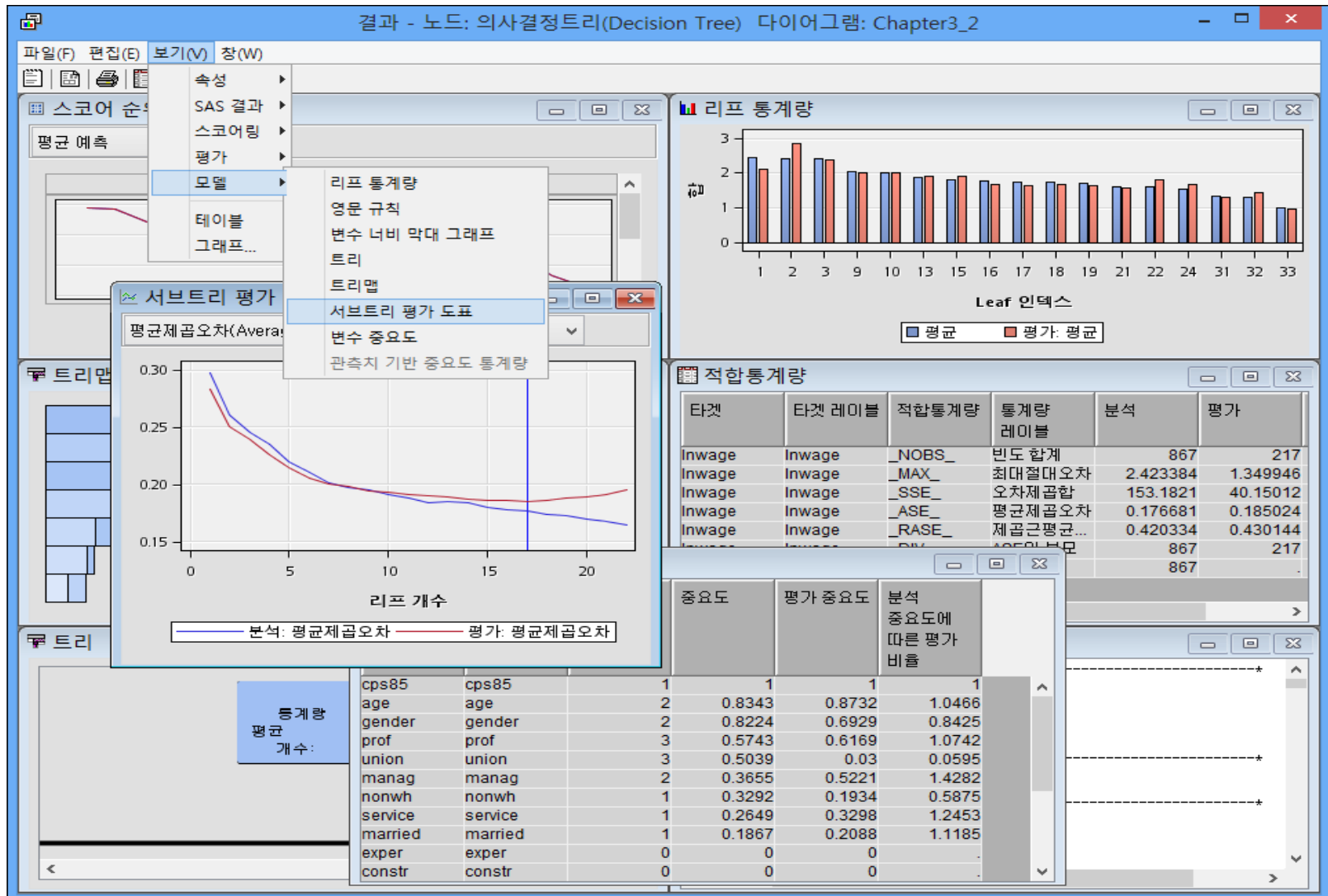
100%

hckang(으)로서의 hckang hckang-pc에 연결

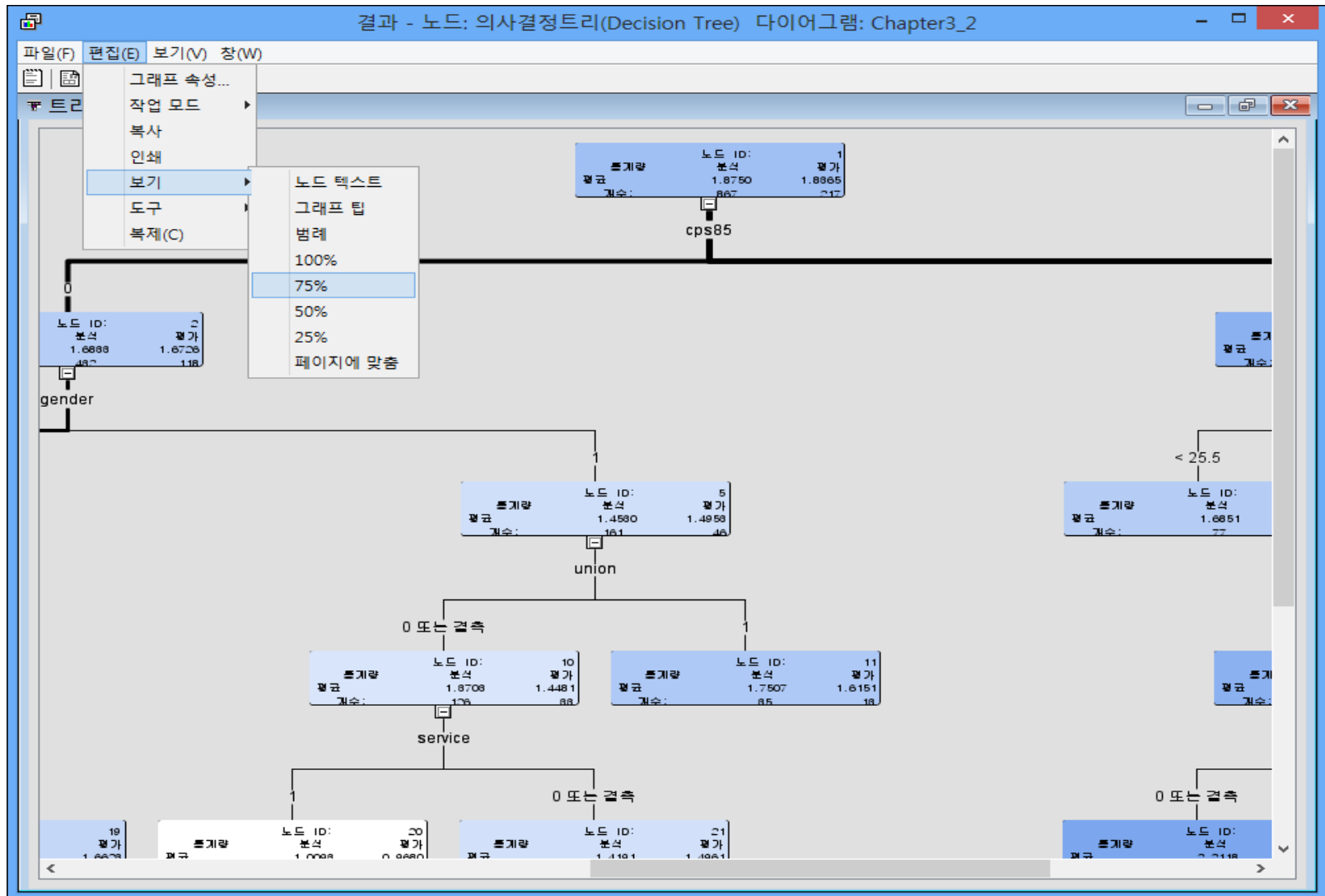
### 3.5.1 변수들의 분포에 대한 탐색



## 3.5.2 의사결정트리 노드의 실행과 결과 보기



# 의사결정트리(Decision Tree) 노드 - 결과: 트리





## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

# 의사결정트리(Decision Tree) 노드의 속성 패널

Enterprise Miner - DM Project

파일(F) 편집(E) 보기(V) 작업(A) 옵션(O) 창(W) 도움말(H)

표본추출 탐색 수정 모델 평가 유틸리티 응용 프로그램 시계열

DM Project

- 데이터 소스
  - All: German Credit Data
  - BUYROLL
  - BUYTEST
  - DMAGESCR
  - HMEQ
  - WAGES
- 다이어그램
  - Chapter3\_1

**의사결정트리 노드의 속성 패널**

속성

일반

노드 ID	Tree
가져온 데이터	...
내보낸 데이터	...
노트	...

분석

변수	...
대화식(Interactive)	...
기본 트리 사용(Use From)	...
다중 타겟 사용(Use Multiple)	...
정확도(Precision)	4

분리 규칙(Splitting Rule)

Interval 기준(Interval CrProbF)	...
Nominal 기준(Nominal ProbChisq)	...
Ordinal 기준(Ordinal CrEntropy)	...
유의수준(Significance L0.2)	...
결측값(Missing Values) 검색에 사용	...
입력변수 한 번만 사용(LOH)이	...
최대 가지(Maximum Br4)	...
최대 깊이(Maximum De10)	...

**대화식(Interactive)**

대화식 분석 창이 실행됩니다.

**Chapter3\_1**

다이어그램

로그

실행 완료

hckang(으)로서의 hckang hckang-pc에 연결

**대화식 분석 버튼**

**의사결정트리(Decision Tree) 노드의 속성 패널**

다이어그램

로그

실행 완료

hckang(으)로서의 hckang hckang-pc에 연결

[illegible]

# 노드 분리 및 분리 규칙 편집 대화상자

대화식 의사결정트리 - EMWS2.TREE\_BROWSETREE[EMWS2.EM\_TREE]

파일 편집 뷰 작업 창

트리 뷰

동계량	분석	평가
1:	19.93%	20.00%
0:	80.07%	80.00%
개수:	4170	1790

그래프 속성...

인쇄

보기

도구

노드 통계량

노드 분리...

노드 분석

노드 가지치기

타겟 전환...

하위 항목 복사

하위 항목 붙여넣기

저장된 트리 붙여넣기...

가지치기 실행 취소

분리 개수

경로의 노드 ID

분리 변수

0

1(선택됨)

노드 1 분리

타겟 변수: BAD

변수	변수 설명	-Log(p)	가지
DEBTINC	DEBTINC	331.026	4
DELINQ	DELINQ	105.3485	4
VALUE	VALUE	76.7498	4
DEROG	DEROG	54.1836	4
CLAGE	CLAGE	31.07	4
NINQ	NINQ	28.3353	3
LOAN	LOAN	26.1572	3
YOJ	YOJ	19.9992	4
JOB	JOB	12.1423	4
CLNO	CLNO	8.7765	4
MORTDUE	MORTDUE	8.228	4
REASON	REASON	0.5932	3

규칙 편집...

DEBTINC - Interval 분리 규칙

타겟 변수: BAD

결측값 할당

☒ 특정 가지

1

☐ 별도의 결측값 가지

☐ 모든 가지

가지

가지	분할점
1	< 11,7112
2	< 41,0928
3	< 45,1848
4	>= 45,1848

새로운 분할점:

가지 추가

확인

취소

적용

재설정

규칙 편집...


취소

적용

새로 고침

평가	
4170	1790
0	0
19.93%	20.00%
80.07%	80.00%
80.07%	80.00%

# 규칙 편집 대화상자


DEBTINC - Interval 분리 규칙

타겟 변수: BAD


결측값 할당  
☐ 특정 가지    1 ▼  
☒ 별도의 결측값 가지  
☐ 모든 가지

가지

가지		분할점
1	<	0.0000
2	<	11.7112
3	<	41.0928
4	<	45.1848
5	>=	45.1848
6		결측값

새로운 분할점: 4    가지 추가    가지 제거

확인    취소


DEBTINC - Interval 분리 규칙

타겟 변수: BAD

결측값 할당  
☐ 특정 가지    1 ▼  
☒ 별도의 결측값 가지  
☐ 모든 가지

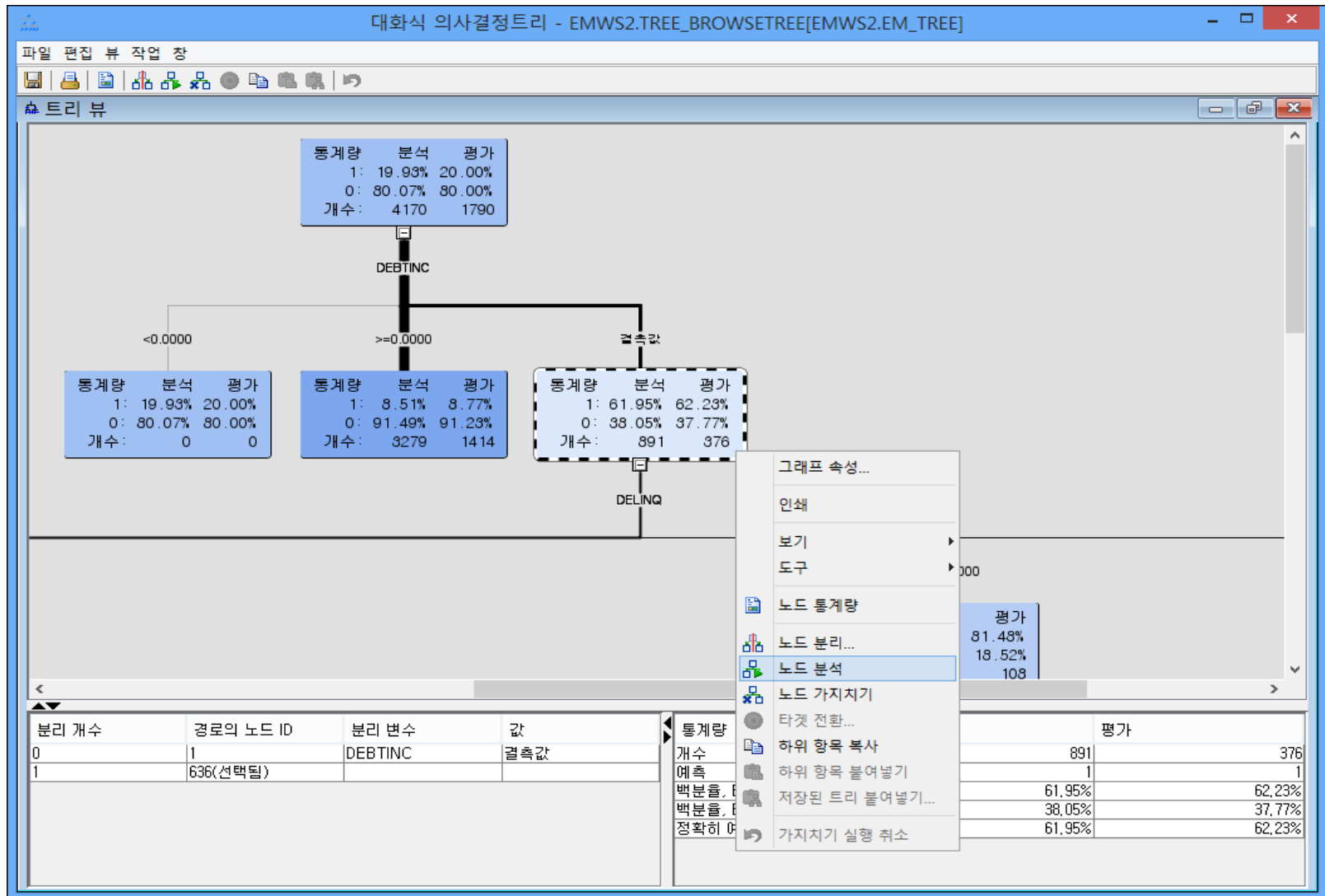
가지

가지		분할점
1	<	0.0000
2	>=	0.0000
3		결측값

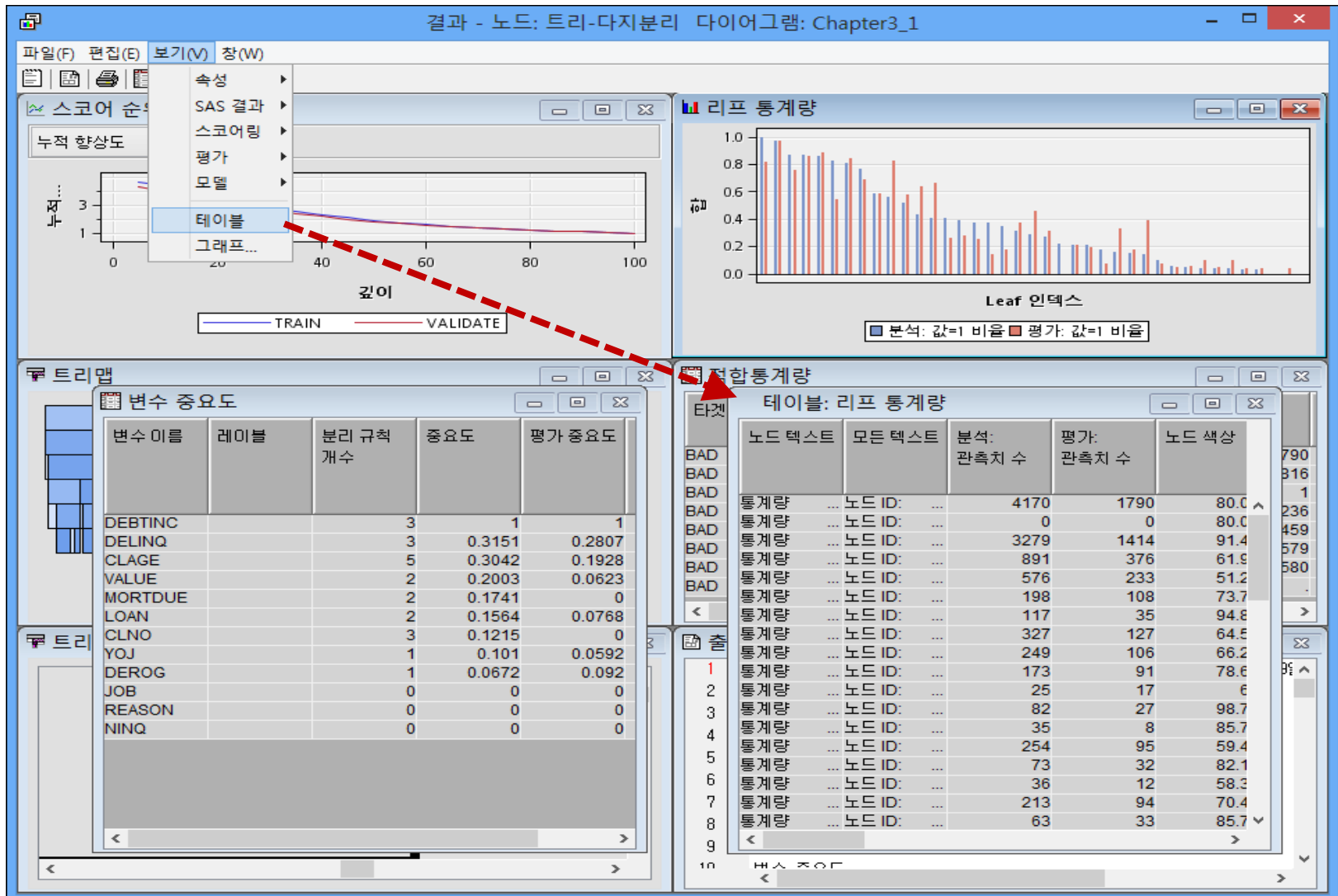
새로운 분할점:    가지 추가    가지 제거

확인    취소    적용    재설정

# 노드 분석이 수행된 결과



# 의사결정트리(Decision Tree) 노드 - 결과



## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제



## 의사결정나무모형에 대한 요약 테이블의 예

Segment	n	%Bad	DEBTINC	DELINQ	CLAGE	VALUE
전체	4,172	20.0%	-	-	-	-
1	347 (8.3%)	83.0%	43.5 이상 Missing	0.5 이상	-	-
2	406 (9.7%)	63.8%	43.5 이상 Missing	0.5 미만 Missing	178.2 미만 Missing	-
3	237 (5.7%)	29.1%	43.5 이상 Missing	0.5 미만 Missing	178.2 이상	-
4	498 (11.9%)	17.3%	43.5 미만	0.5 이상	-	-
5	231 (5.5%)	14.7%	43.5 미만	0.5 미만 Missing	-	49328.5 미만 Missing
6	476 (11.4%)	12.4%	31.2 이상 43.5 이하	0.5 미만 Missing	123.5 미만 Missing	49328.5 이상
7	100 (2.4%)	7.0%	43.5 미만	0.5 미만 Missing	123.5 이상	209545 이상
8	1,588 (38.1%)	1.7%	43.5 미만	0.5 미만 Missing	123.5 이상	49328.5 이상 209545 미만
9	289 (6.9%)	1.4%	31.2 미만	0.5 미만 Missing	123.5 미만 Missing	49328.5 이상
중요도 (Importance)			1.000	0.355	0.293	0.108

## 차례

---

- 3.1 의사결정나무의 개념
- 3.2 의사결정나무의 분리기준
- 3.3 의사결정나무분석의 특징
- 3.4 분석사례 - 1(분류나무): 신용평가 문제
- 3.5 분석사례 - 2(회귀나무): 평균임금의 예측
- 3.6 분석사례 - 3: 의사결정나무분석의 대화식 수행
- 3.7 의사결정나무모형에 대한 요약 테이블 작성
- 3.8 연습문제

# 연습문제 3-8을 위한 다이어그램

The screenshot displays the SAS Enterprise Miner interface. On the left, a project tree shows a diagram named 'Ex3\_8'. A callout bubble points to the '속성' (Properties) panel, which contains various settings for the selected node. A red dashed arrow points from the '속성' panel to the '분석용 코드 - 코드 노트' (Analysis Code - Code Note) window.

**속성 (Properties) Panel:**

- 일반 (General):**
  - 노드 ID: EMCODE
  - 가져온 데이터: ...
  - 내보낸 데이터: ...
  - 노트: ...
- 분석 (Analysis):**
  - 변수: ...
  - 코드 편집기(Code Editor): ...
  - 도구 유형 (Tool Type): 유틸리티
  - 선행 노드 필요(Data Node): 아니요
  - 재실행(Rerun): 아니요
  - 사전확률 사용(Use Prior): 예
- 스코어 (Score):**
  - 유형 (Advisor Type): 기본
  - 게시 코드(Publish Code): ...
  - 코드 형식(Code Format): DATA 스타일
- 상태 (Status):**
  - 생성 시간: 14. 2. 16 오후 1:49
  - 실행 ID: 7fe8a251-e9bd-4622-b8...
  - 최근 오류: ...
  - 최근 상태: 완료
  - 최근 실행 시간: 14. 2. 16 오후 1:56

**분석용 코드 - 코드 노트 (Analysis Code - Code Note) Window:**

```

.. 매크로
EM_IMPORT_DATA
EM_IMPORT_DATA_EMINFO
EM_IMPORT_DATA_CMETA
EM_IMPORT_VALIDATE
EM_IMPORT_VALIDATE_CMETA
EM_IMPORT_TEST
EM_IMPORT_TEST_CMETA

DATA &em_export_train;
  SET &em_import_data;
  IF respond=1;
RUN;
    
```

The status bar at the bottom indicates: hckang(으)로서의 hckang - DM Project - Ex3\_8 - EMCODE - COMPLETE