제3장 의사결정나무

Decision Tree Analysis

데이터 마이닝 기법 분류

❖ **지도예측**

(Supervised Prediction)

입력변수, 목표변수가 존재

입력변수로부터 목표 값을 예측하는 모형 개발이 목적

- Binary Classirier : 이항 분류
- Neural Network : 신경망 모형
- Decision Tree : 의사결정나무
  - C5.0, CART, QUEST, CHAID
- Regression : 회귀분석
- Logistic : 로지스틱 회귀분석
- Discriminant : 판별분석
- Time series : 시계열분석

❖ **자율예측**

(Unsupervised Prediction)

목표변수가 명확히 규정되지 않음

데이터에 존재하는 여러 형태의 특징을 찾는 것이 목적

- K-Means : K-평균 군집화
- Two Step : 2단계 군집화
- Apriori : 연관성 규칙
- PCA / Factor : 주성분 / 인자분석

INDEX

1. 의사결정나무의 개념

2. 의사결정나무의 분리기준

3. 의사결정나무분석의 특징

4. 분석사례1 분류나무

5. 분석사례2 회귀나무

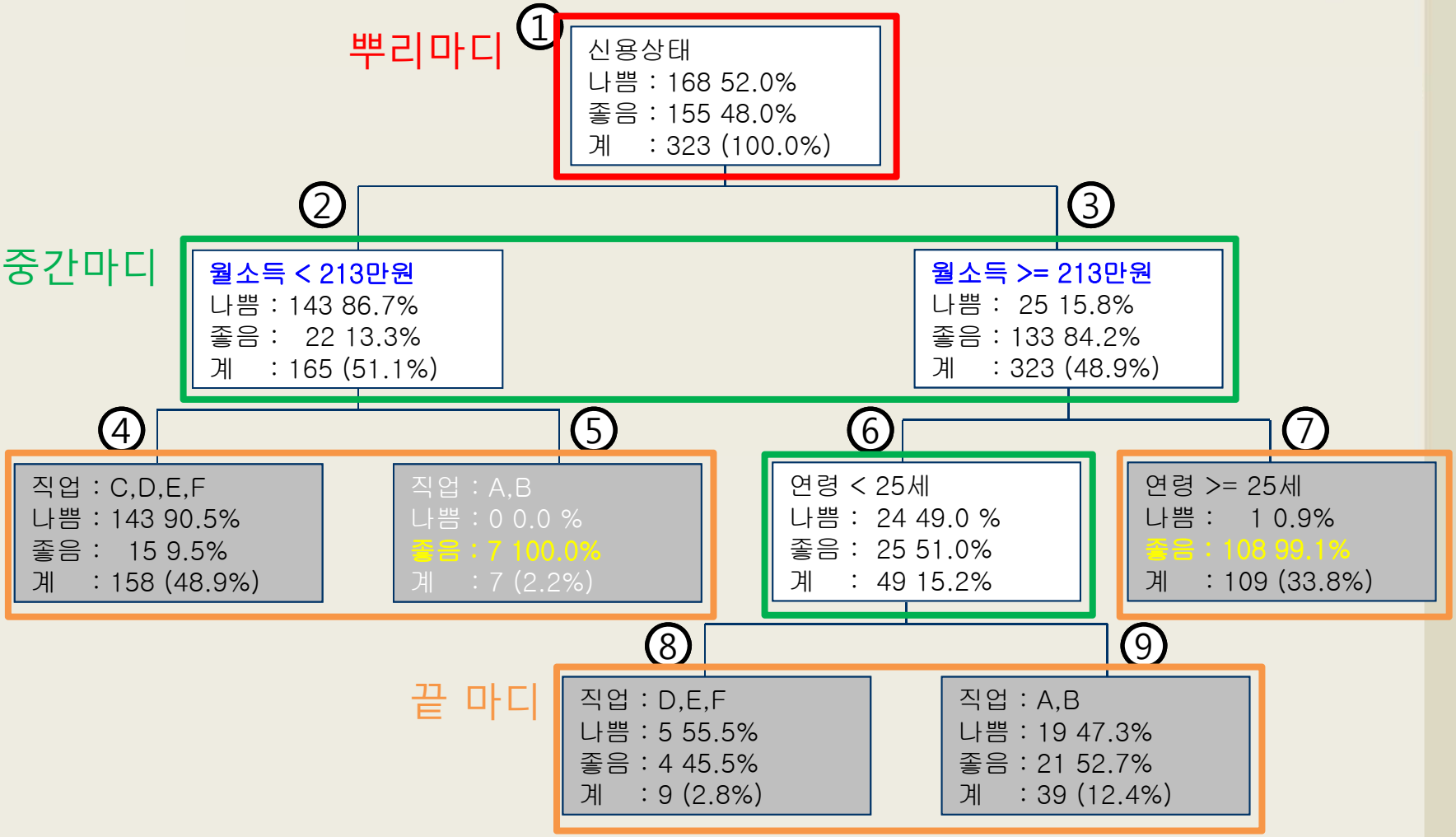
6. 분석사례3 대화식 수행

제3장 의사결정나무

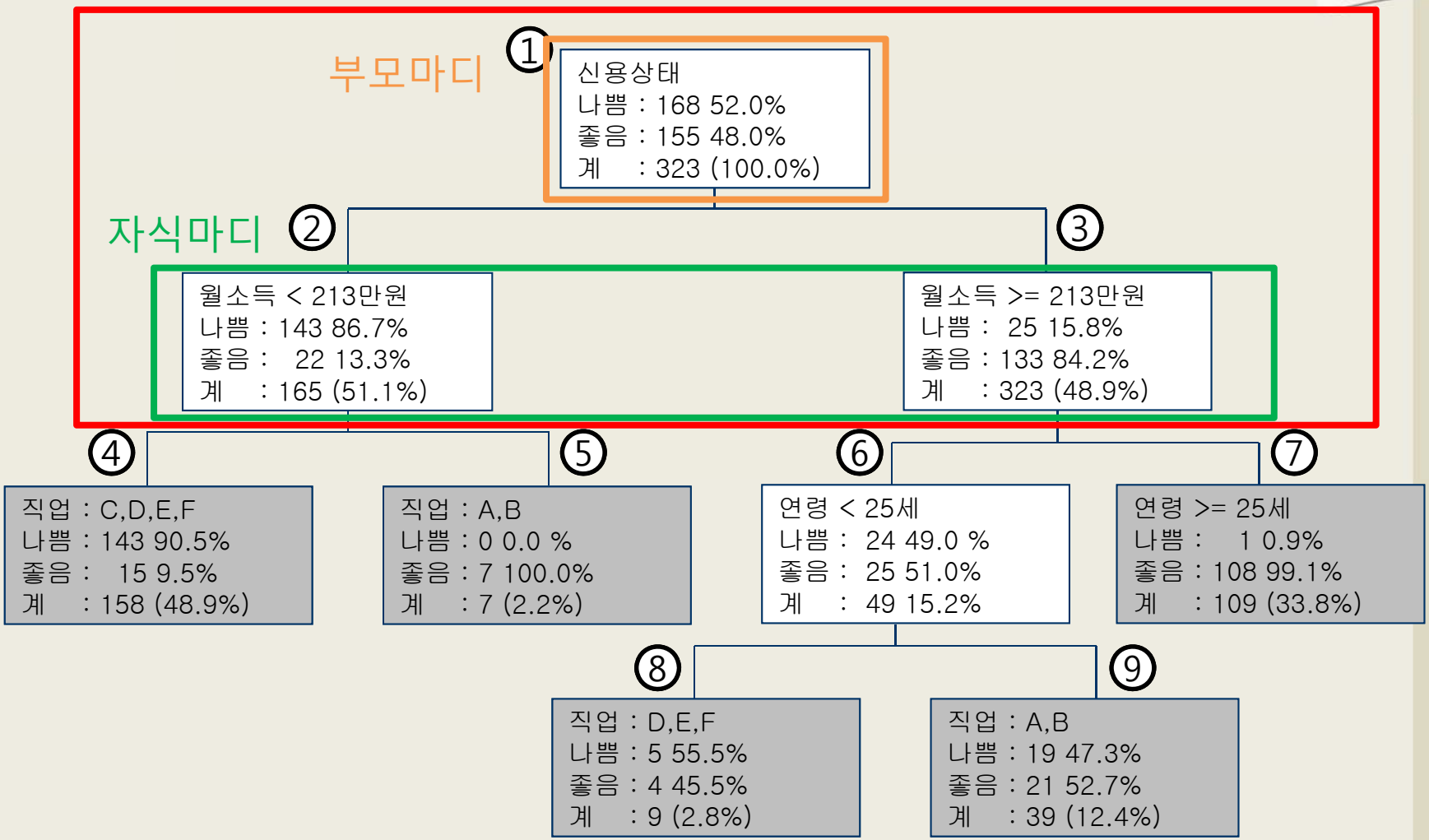
3.1 의사결정나무의 개념

- 의사결정규칙을 나무구조로 도표화 하여 분류와 예측을 수행

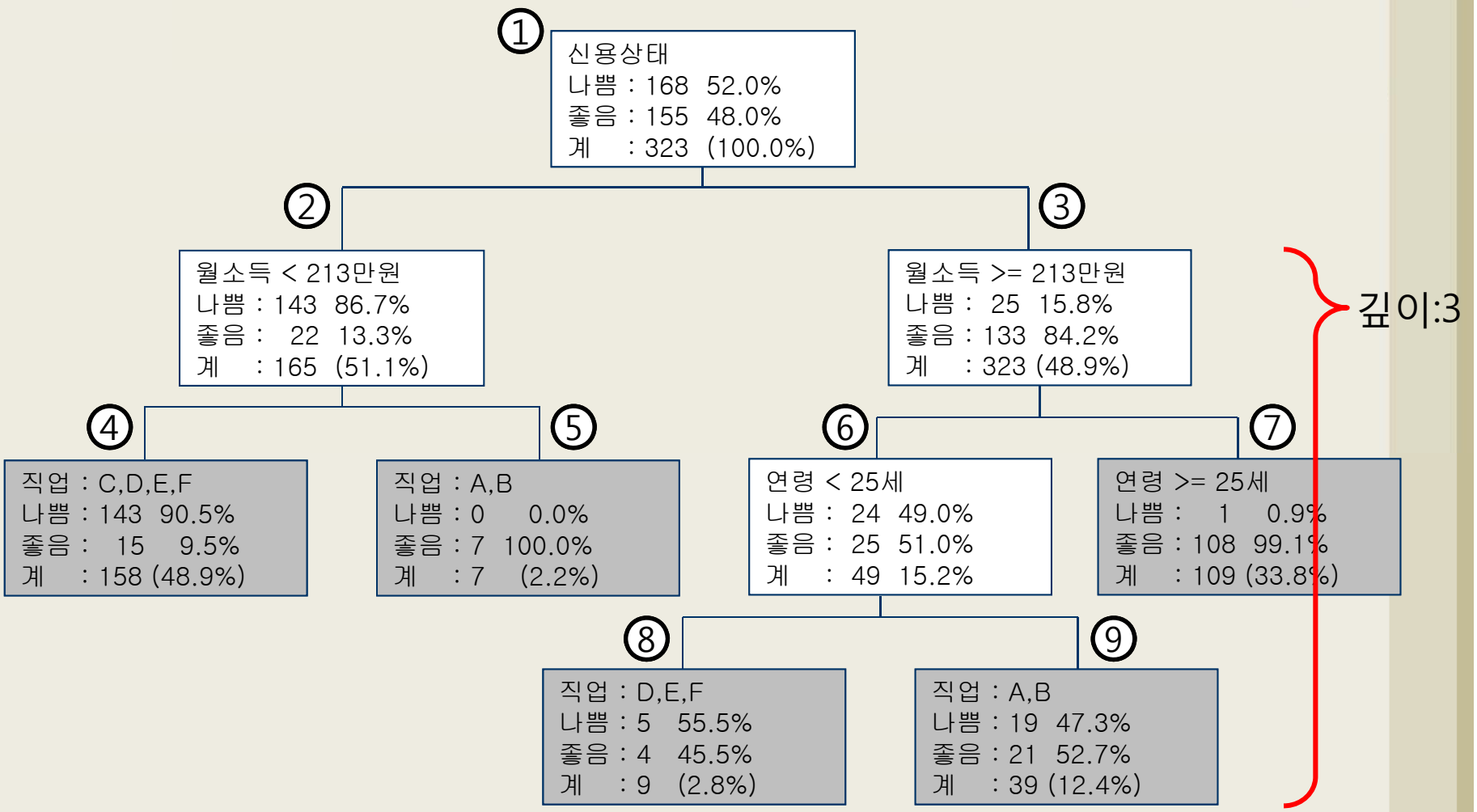
3.1.1 의사결정나무의 구성요소



- ✓ **뿌리마디** (root node) : 나무구조가 시작되는 마디
- ✓ **끝 마디** (terminal node, leaf) : 각 나무줄기의 끝에 위치하는 마디
- ✓ **중간마디** (inernal node) : 중간에 있는 끝 마디가 아닌 마디



- ✓ **자식마디** (child node) : 하나의 마디로부터 분리되어 나간 마디
- ✓ **부모마디** (parent node) : 자식마디의 상위마디



- ✓ **가지** (branch) : 하나의 마디로 부터 끝 마디까지 연결된 마디들
- ✓ **깊이** (depth) : 가지를 이루고 있는 마디의 개수

3.1.2 의사결정나무의 형성과정

- ✓ **의사결정나무의 형성**
  : 목적과 자료구조에 따라 적절한 **분리기준**과 **정지규칙**을 지정하여 의사결정나무를 얻음

- ✓ **가지치기**
  : 분류오류를 크게 할 위험이 높거나 부적절한 가지를 제거

- ✓ **타당성 평가**
  : 모형평가 도구 (이득도표나 위험도표)
  또는 평가용 데이터에 의한 교차타당성 등을 이용해 평가

- ✓ **해석 및 예측**
  : 의사결정나무를 해석하고 예측모형을 구축

제3장 의사결정나무

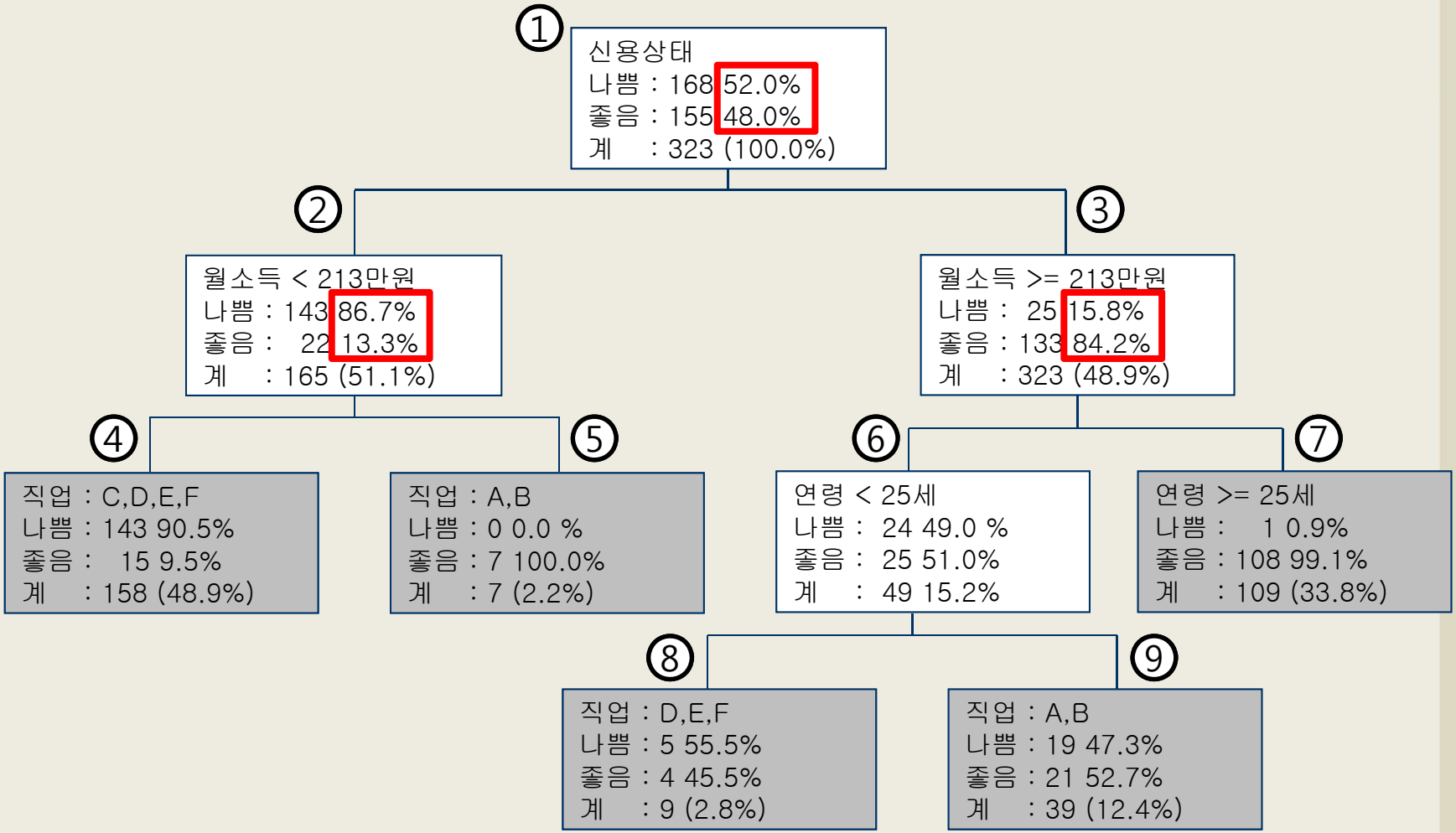
3.2 의사결정나무의 분리기준

3.2.1 의사결정나무의 분리기준

✓ **분리기준**(split criterion)
: 어떤 입력변수를 이용하여 어떻게 분리하는 것이
목표변수의 분포를 가장 잘 구별해 주는지 그 기준

- 목표변수의 분포를 구별하는 정도 : **순수도** or **불순도**
  * 순수도 : 목표변수의 특정 범주에 개체들이 포함되어 있는 정도

- 부모마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식마디를 형성함



분리기준이란, 부모마디에 비해서 자식마디들에서

순수도가 증가하는 정도를 수치화 한 것이다.

3.2.2 분류나무와 회귀나무

분류나무(classification tree) : 이산형 (범주형) 목표변수의 경우
목표변수의 각 범주에 속하는 빈도에 기초하여 분리가 일어남

✓ 카이제곱 통계량의 p-값 (p-value of Chi-square statistics) : **CHAID, QUEST**

$$\phi(g)=\chi^2=\sum_{i=1}^{g}\frac{(n_i-np_{i0})^2}{np_{i0}}$$

✓ 지니지수 (Gini index) : **CART**

$$\phi(g)=1-\sum_{i=1}^{j}\hat{p}_i(g)^2$$

✓ 엔트로피지수 (Entropy index) : **C5.0**

$$\phi(g)=-\sum_{i=1}^{j}\hat{p}_i(g)\log\hat{p}_i(g)$$

회귀나무(regression tree) : 연속형 (구간형) 목표변수의 경우
목표변수의 평균과 표준편차에 기초하여 분리가 일어남

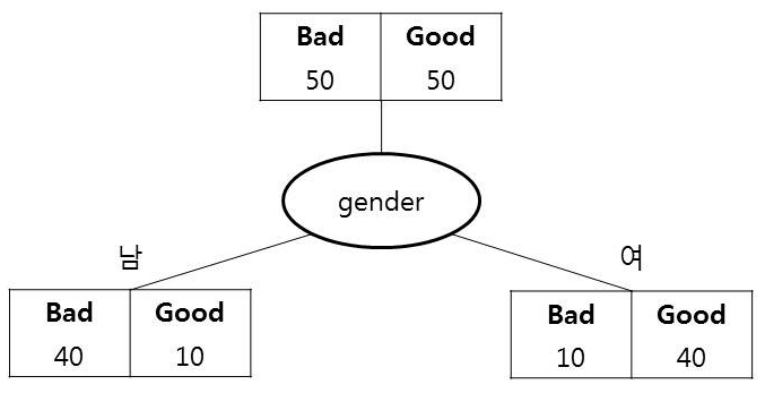
✓ 분산분석 F- 통계량의 p-값 (p-value of F-Statistics) : **CHAID**

✓ 분산의 감소량 (Variance reduction) : **CART**

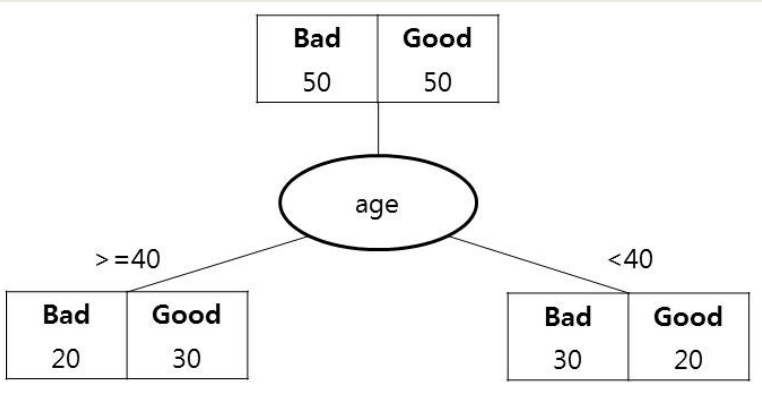
의사결정나무 구축방법 (예)

불순도 계산의 예 (Gini 지수 사용) $\phi(g)=1-\sum_{i=1}^{j}\hat{p}_i(g)^2$

성별을 기준으로 나눈 경우



나이를 기준으로 나눈 경우



	성별을 기준으로 나눈 경우	나이를 기준으로 나눈 경우
뿌리 노드의 불순도	$1-[(50/100)^2+(50/100)^2]=0.5$	
	$1-[(40/50)^2+(10/50)^2]=0.32$ $1-[(10/50)^2+(40/50)^2]=0.32$	$1-[(20/50)^2+(30/50)^2]=0.48$ $1-[(30/50)^2+(20/50)^2]=0.48$
불순도의 감소 폭	$0.5-[(50/100)*0.32]*2=0.18$	$0.5-[(50/100)*0.48]*2=0.02$

즉, 성별에 의해 자료를 나누는 것이 나이를 기준으로 분할하는 것 보다 종료 노드의 순수성의 증가에 도움이 된다.

CART (Classification And Regression Tree) Algorithm

- 1984년 Breiman과 그의 동료들이 발명
- 기계학습(machine learning) 실험의 산물
- **가장 널리 사용되는 의사결정나무 알고리즘**
- 가장 성취도가 좋은 변수 및 수준을 찾는 것에 중점

- 불순도로는 목표변수가 : 범주형인 경우 → 지니지수 (Gini Index)
- 불순도로는 목표변수가 : 연속형인 경우 → 분산의 감소량

- 분리 방법 : 이지 분리 (binary split)

C 4.5, C5.0 Algorithm

- 호주의 연구원 J. Ross Quinlan에 의하여 개발
- 초기버전은 ID 3 (Iterative Dichotomizer 3)로 1986년에 개발
- 가지치기를 사용할 때 학습자료를 사용함

- 목표변수가 반드시 범주형이어야 하며,
- 불순도로 엔트로피 지수 (Entropy index) 사용

- 분리방법 : 다지 분리 (multiple split)
- 예측변수가 범주형일 경우, 범주의 수만큼 분리가 일어남

CHAID (Chi-squared Automatic Interaction Detection) Algorithm

- 1975년 J.A. Hartigan이 발표
- AID (Automatic Interaction Detection)를 발전시킨 알고리즘
- CHAID는 가지치기를 하지 않고 나무를 적당한 크기에서 성장을 중지

- 예측변수가 반드시 범주형이어야 함
- 불순도로 카이제곱 통계량을 사용

- 분리방법 : 다지 분리 (multiple split)
- 분리변수의 각 범주가 하나의 부마디(sub-node)를 형성

QUEST Algorithm

- 1997년 Loh and Shih이 발표
- 변수의 선택에서 범주의 개수가 많은 범주형 변수로의 bias가 심각한 CART의 문제점을 개선한 알고리즘
- 변수 선택 bias가 거의 없음
- 분리규칙은 분리변수 선택과 분리점 선택의 두 단계로 나누어 시행

- 불순도로 카이제곱 통계량을 사용

- 분리방법 : 이지 분리 (binary split)

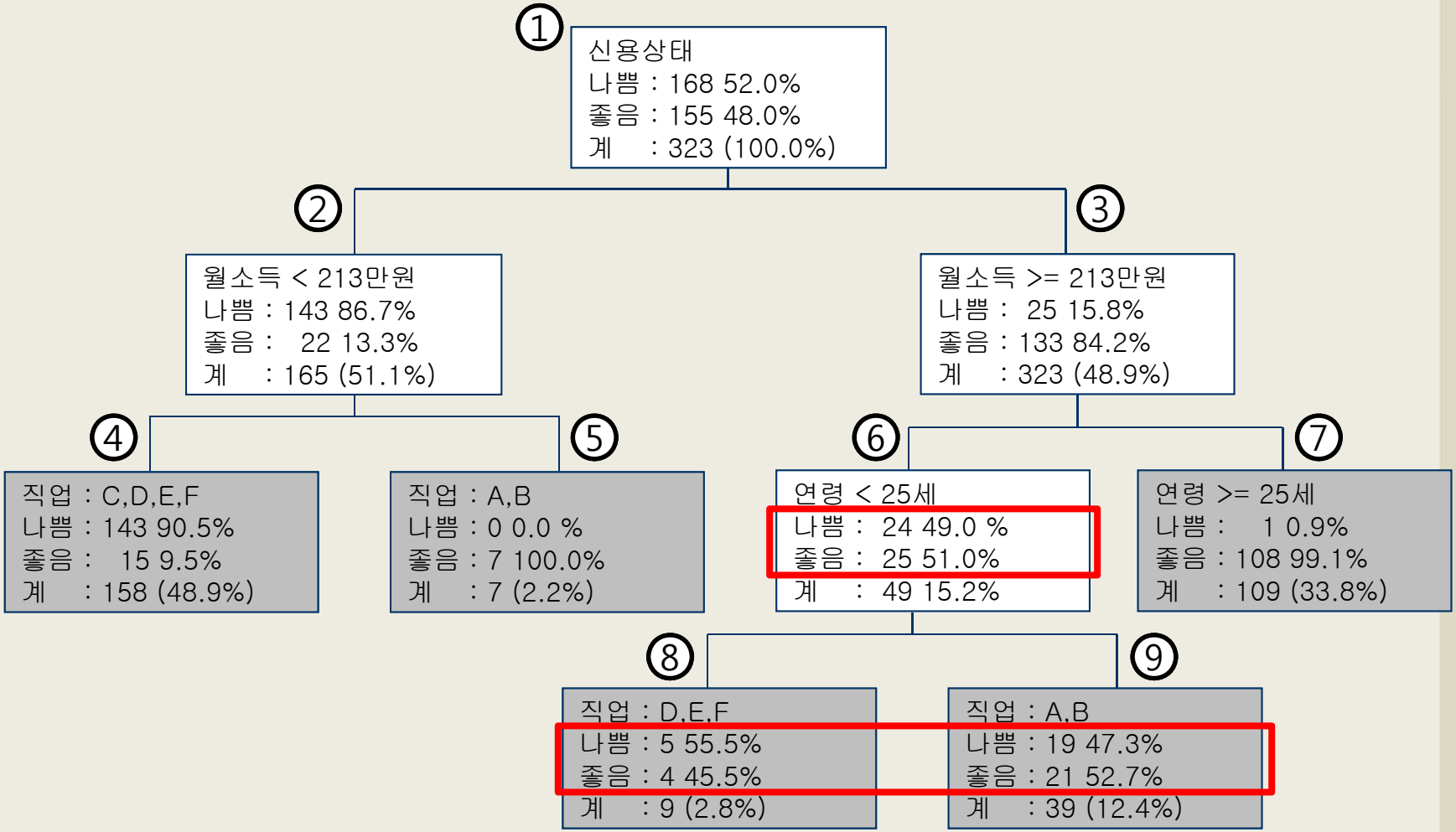
알고리즘 요약표

	CART	C 5.0	CHAID	QUEST
목표 변수	범주형 연속형	**범주형**	범주형 연속형	범주형
예측 변수	범주형 연속형	범주형 연속형	**범주형**	범주형 연속형
분리 기준	지니 지수 분산의 감소량	엔트로피 지수	카이제곱 통계량 F-검정	카이제곱 통계량 F-검정
분리 개수	이지분리	다지분리	다지분리	이지분리

3.2.3 정지규칙과 가지치기

- ✓ 정지규칙 (stoping rule)
  : 더 이상 분리가 일어나지 않고,
  현재의 마디가 끝 마디가 되도록 하는 규칙

- ✓ 가지치기 (pruning)
  : 적절하지 않은 마디를 제거하여,
  적당한 크기의 부나무(subtree) 구조를 가지도록 하는 규칙



적절한 정지규칙 or 가지치기를 수행하여 제거하는 것이 바람직하다.

제3장 의사결정나무

3.3 의사결정나무분석의 특징

3.3.1 의사결정나무분석의 장점

- ✓ **해석의 용이성**
  - 나무구조에 의해서 모형이 표현되기 때문에 해석이 쉽다.
  - 새로운 자료에 모형을 적합 시키기 쉽다.
  - 어떤 입력변수가 중요한지 파악이 쉽다.

- ✓ **교호작용 효과의 해석**
  - 두 개 이상의 변수가 결합하여 목표변수에 어떠한 영향을 주는지 알기 쉽다.

- ✓ **비모수적 모형**
  - 선형성, 정규성, 등분산성의 가정이 필요치 않다.
  - 단지 순위만 분석에 영향을 주므로 이상치에 민감하지 않다.

3.3.2 의사결정나무분석의 단점

- ✓ **비연속성**
  - 연속형 변수를 비연속적인 값으로 취급하여 예측오류가 클 가능성이 있다.

- ✓ **선형성 또는 주효과의 결여**
  - 선형 또는 주효과 모형에서와 같은 결과를 얻을 수 없다.

- ✓ **비안정성**
  - 분석용 자료에만 의존하므로 새로운 자료의 예측에 불안정하다.
    * 평가용 데이터에 의한 교차타당성 평가
      or 가지치기에 의해 안정성 있는 결과를 얻는 것이 바람직