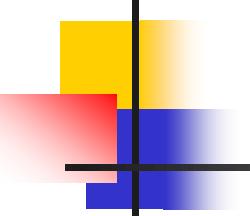


# Course Overview



# Objectives

---

## ■ 교과목 개요

- 빅데이터에서 데이터 간의 관계, 패턴, 규칙 등을 찾아내고 모형화해 유용한 경영정보로 변환시키는 일련의 과정을 이론과 함께 IBM SPSS Modeler와 R을 통한 실습을 병행하여 교육하고 마케팅, 영업, 고객관리, 금융, 생산 등 다양한 경영부문에서 이를 활용할 수 있는 방법론을 제시

## ■ 수업 목표

- 데이터마이닝의 개념, 절차, 기법에 대한 이해
- 데이터마이닝 도구(IBM SPSS Modeler와 R)의 활용능력 배양
- 다양한 비즈니스 문제를 실제 데이터를 가지고 분석함으로써 데이터마이닝 실무 능력을 배양

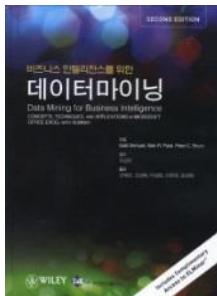
# Textbooks



## 경영을 위한 데이터마이닝 (2판)

Michael J. A. Berry , Gordon S. Linoff

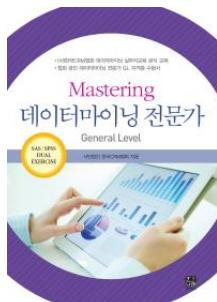
2009 / 한경사



## 비즈니스 인텔리전스를 위한 데이터마이닝 (2판)

Galit Shmueli , Nitin R. Patel, Peter C. Bruce

2012 / 이앤비플러스

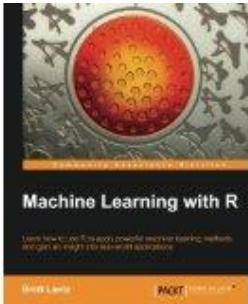


## Mastering 데이터마이닝 전문가: General Level

한국CRM협회

2014 / 한나래

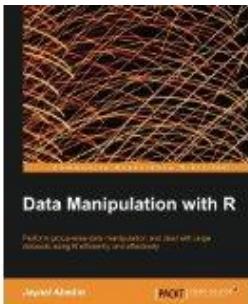
# Textbooks



Machine Learning with R

Brett Lantz

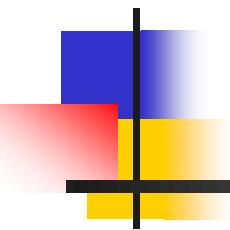
2013 / Packt Publishing



Data Manipulation with R

Jaynal Abedin

2014 / Packt Publishing

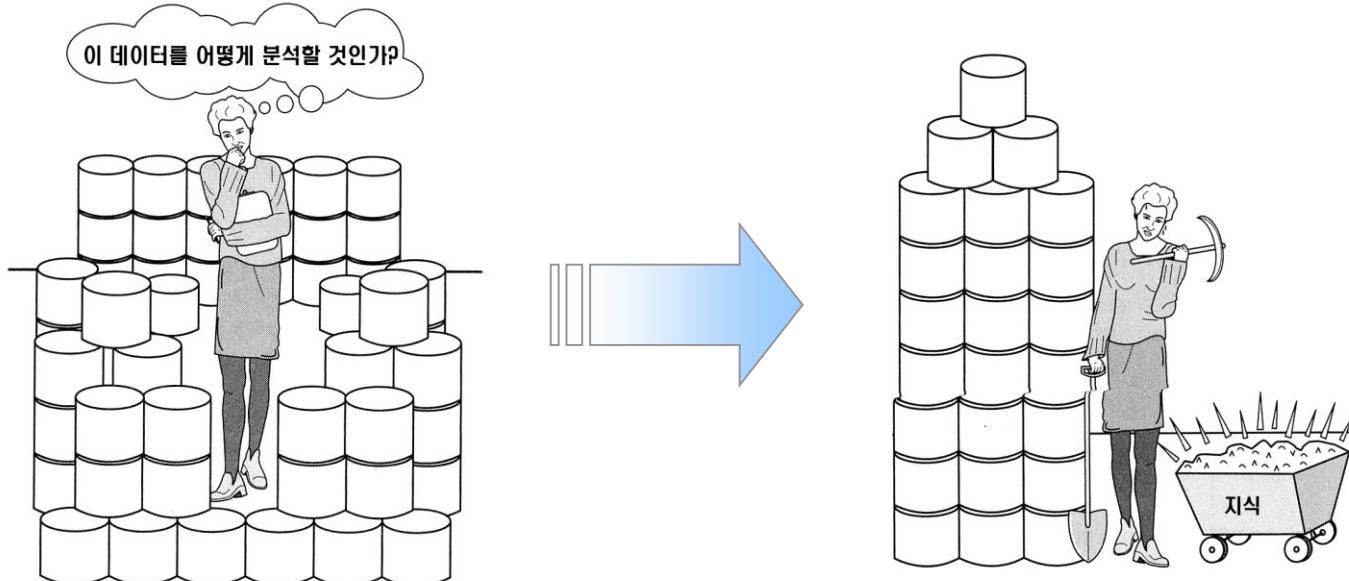


# Introduction to Data Mining

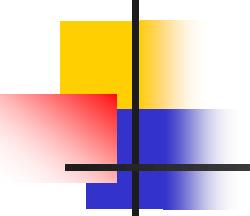
# Concept of Data Mining

## ■ 데이터마이닝(Data Mining)의 정의

- 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하여 의사결정에 활용하는 작업



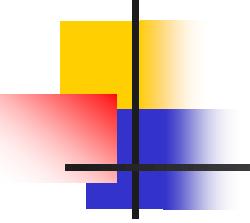
- 지식발견(KDD: Knowledge Discovery in Database)
- 정보발견(Information Discovery), 정보수확(Information Harvesting)
- 정보고고학 (Data Archeology), 자료패턴처리 (Data Pattern Processing)



# Concept of Data Mining

## ■ 데이터마이닝의 다양한 정의

- 데이터베이스에서 지지발견은 데이터에 있는 유효하고, 잠재적으로 이용가능하며 궁극적으로 이해할 수 있는 패턴을 식별하는 중요한 프로세스 (Fayyad et al., “Advance in Knowledge Discovery and Data Mining,” 1996)
- 데이터 마이닝은 비즈니스 문제를 해결하기 위해 현재 조치를 취할 수 있고, 명시적이며 새로운 정보를 추출하기 위해 세부적인 데이터를 분석하는 프로세스이다.(NCR)
- 데이터 마이닝은 큰 데이터베이스로부터 이전에 알려지지 않고, 궁극적으로 이해가능한 정보를 추출 및 중요한 비즈니스 의사결정을 하는 프로세스이다.(IBM)
- 데이터 마이닝은 비즈니스 우위를 위해 이전에 알려지지 않은 패턴을 발견하기 위해 많은 양의 데이터를 선택하고, 탐색 및 모델링하는 프로세스이다. (SAS Institute)
- **데이터 마이닝은 기업의 경영 활동에서 발생하는 대용량 데이터에서 데이터 간의 관계·패턴·규칙 등을 찾아내고 모형화해 유용한 경영 정보로 변환시키는 일련의 과정이다.**



# Statistical Analysis vs. Data Mining

- 전통적 통계분석

대상집단이 있으며, 모집단의 분포 혹은 모형 등 여러 가지 가정을 전제로 하게 되며 이 전제 조건하에서 분석을 실시

→ 표본의 관찰을 통해 모수 전체를 추론하는 과정

- 데이터마이닝

표본조사/실험에서 필연적으로 수반되는 분포라든가 모형에 대한 전제조건이 필요하지 않음

→ 모집단의 전체자료를 이용한 정보화하는 과정

# SQL/OLAP/Data Mining

- ✓ 데이터 마이닝은 분명한 미래의 분석 방향이다. 미래의 CEO들은 데이터 마이닝을 이용한 분석 결과를 첨부하지 않은 보고서는 검토의 가치가 없다고 판단할 것이다. – Elder Research



## SQL/OLAP/Data Mining

- ✓ 2008년 4월의 매출건수는 ?  
→ SQL
- ✓ 2008년 4월의 지역별 매출건수는 ?  
→ OLAP
- ✓ 2008년 4월에 제품을 구매한 홍길동이 향후 6개월 이내에 추가 구매를 할 가능성은?  
→ Data Mining

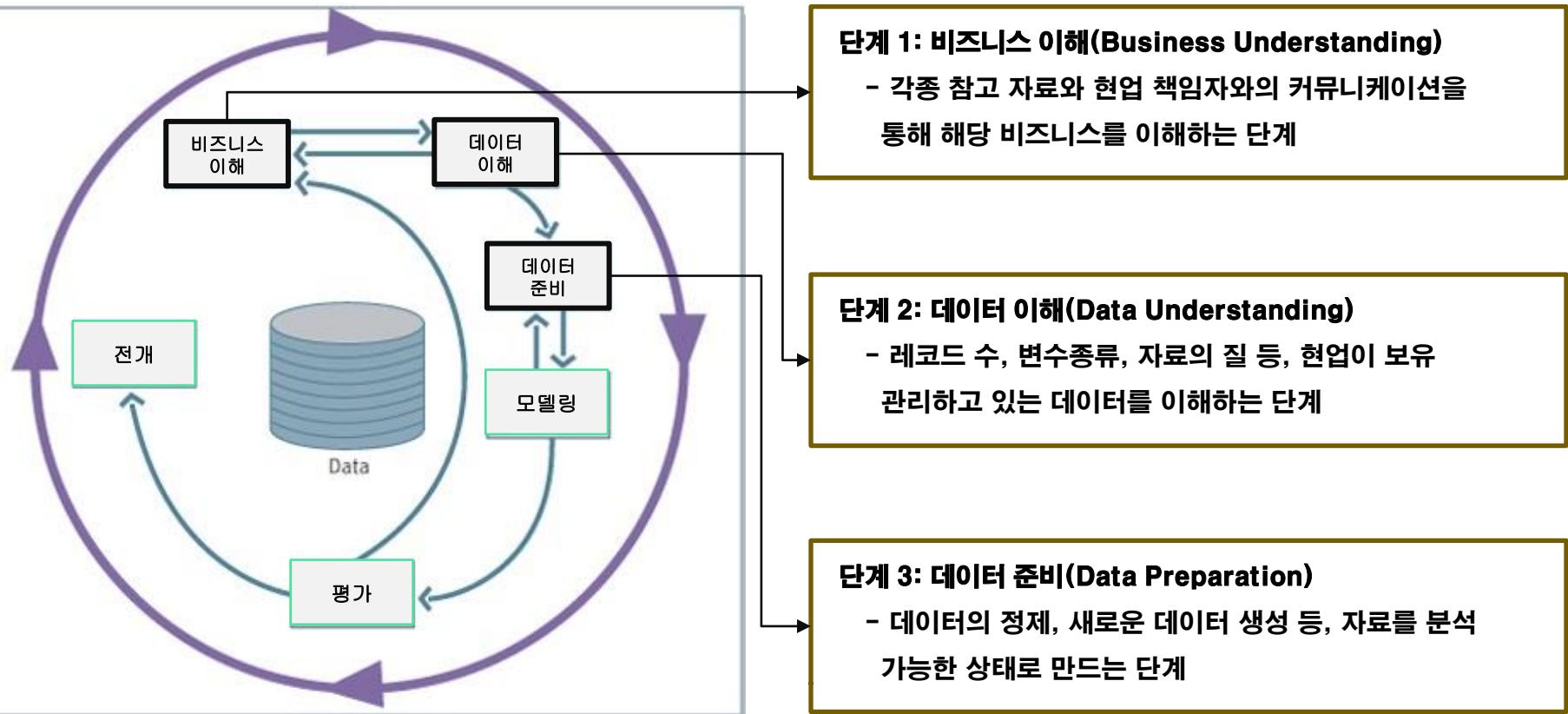
# Data Mining S/W

구분	SAS Enterprise Miner	SPSS Modeler	R	EC Miner
특징	데이터마이닝 프로세스 전반을 지원이 가능하고, 사용이 간편한 GUI를 통해 모델 구축의 가속화가 가능하다.	Data 핸들링에 강하고 사용하기 쉬운 사용자 인터페이스를 가지고 있다.	오픈 소스로서 누구나 자유롭게 실행하고 복사하고 수정하고 배포할 수 있다.	국내에서 개발한 데이터마이닝 소프트웨어로 사용하기 쉽고, 편리한 모델 관리 기능을 가지고 있다.
장점	대용량 데이터분석이 가능하고 활용영역이 다양하다.	자동 모델링 기능, 텍스트 분석 기능, 개체 분석 기능이 강화되었다.	코딩을 이용하기 때문에 다른 도구들에 비해 자유롭게 분석 할 수 있다.	마우스 조작만으로 분석에 필요한 기능 활용이 가능하고, 하나의 프로젝트 창에서 다수의 모델 병렬처리 관리가 가능하다.
단점	초반 작업 설정과 사용법을 습득하는데 다소 시간이 걸린다.	분석을 위한 설정과 연결 과정에 대한 프로세스가 다소 많은 편이다.	코딩의 어려움 때문에 전문가가 아닌 일반인은 이용하기 힘든 편이다.	SAS와 SPSS에 비해 다양한 기법을 제공하지 않다.
평가판 이용 가능 기간	없음.	14일	무료이용가능	90일
홈페이지	<a href="http://www.sas.com/korea/">http://www.sas.com/korea/</a>	<a href="http://www.spss.co.kr/">http://www.spss.co.kr/</a>	<a href="http://www.r-project.org/">http://www.r-project.org/</a>	<a href="http://www.ecminer.com/">http://www.ecminer.com/</a>

# Data Mining Process

## ▪ CRISP-DM : SPSS에서 제시하는 데이터마이닝 프로세스 (1/2)

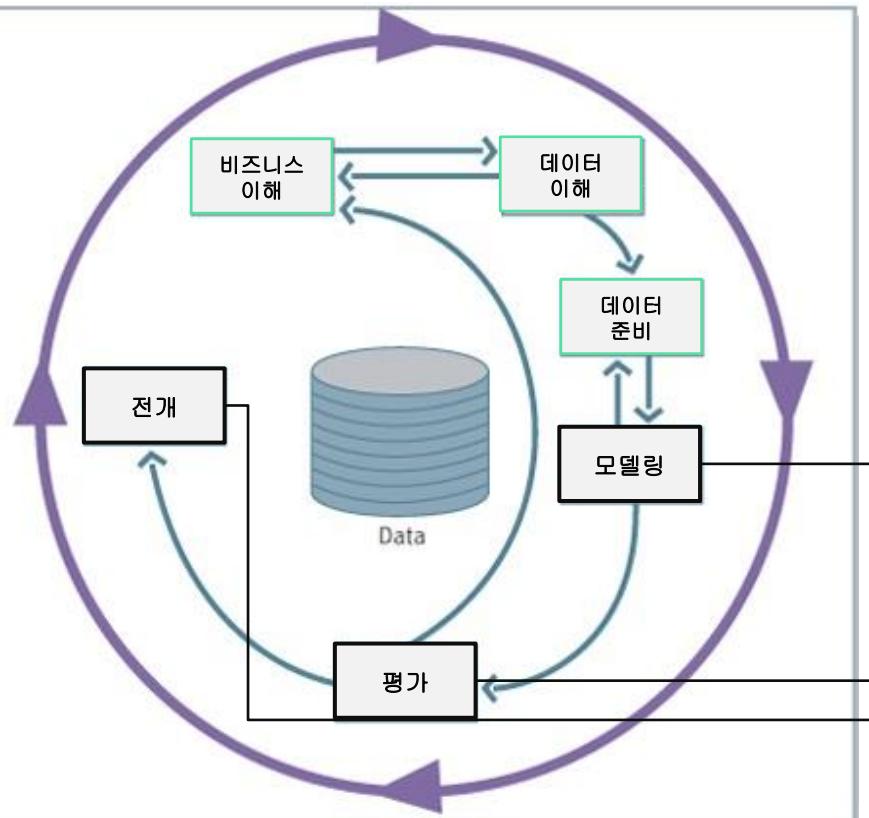
CRISP-DM(cross-industry standard process for data mining)은 데이터마이닝에 관련된 광범위한 업무의 범위를 다루고 있음.



# Data Mining Process

## ▫ CRISP-DM : SPSS에서 제시하는 데이터마이닝 프로세스 (2/2)

CRISP-DM(cross-industry standard process for data mining)은 데이터마이닝에 관련된 광범위한 업무의 범위를 다루고 있음.



### 단계 4: 모델링 (Modeling)

- 자료 기술 및 탐색을 포함하여 필요한 각종 모델링을 하는 단계

### 단계 5: 평가 (Evaluation)

- 모형의 해석 가능 여부, 독립적인 새 자료에 적용되는 경우에도 재현 가능한지를 검토하는 단계

### 단계 6: 전개 (Deployment)

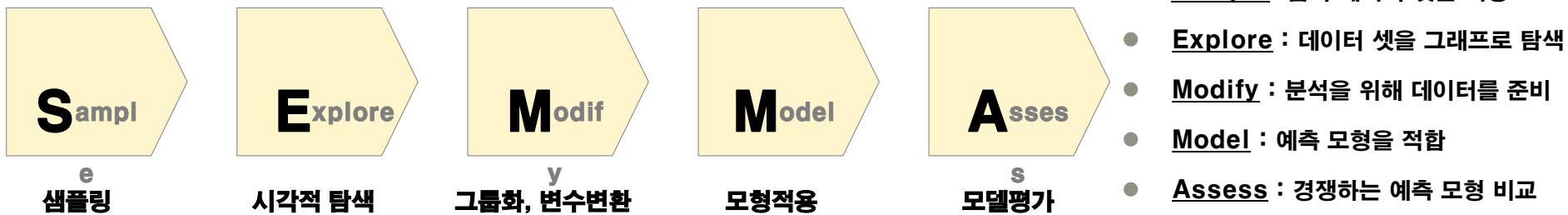
- 각 관리자에게 전달하여 필요한 조치를 취하는 등 검토가 끝난 모형을 실제 현업에 적용하는 단계

# Data Mining Process

## ▣ SEMMA : SAS에서 제시하는 데이터마이닝 프로세스

SEMMA는 데이터마이닝의 기술적 업무에 보다 집중된 과정을 포함하고 있음.

### ▷ 데이터마이닝 표준 실행 과정



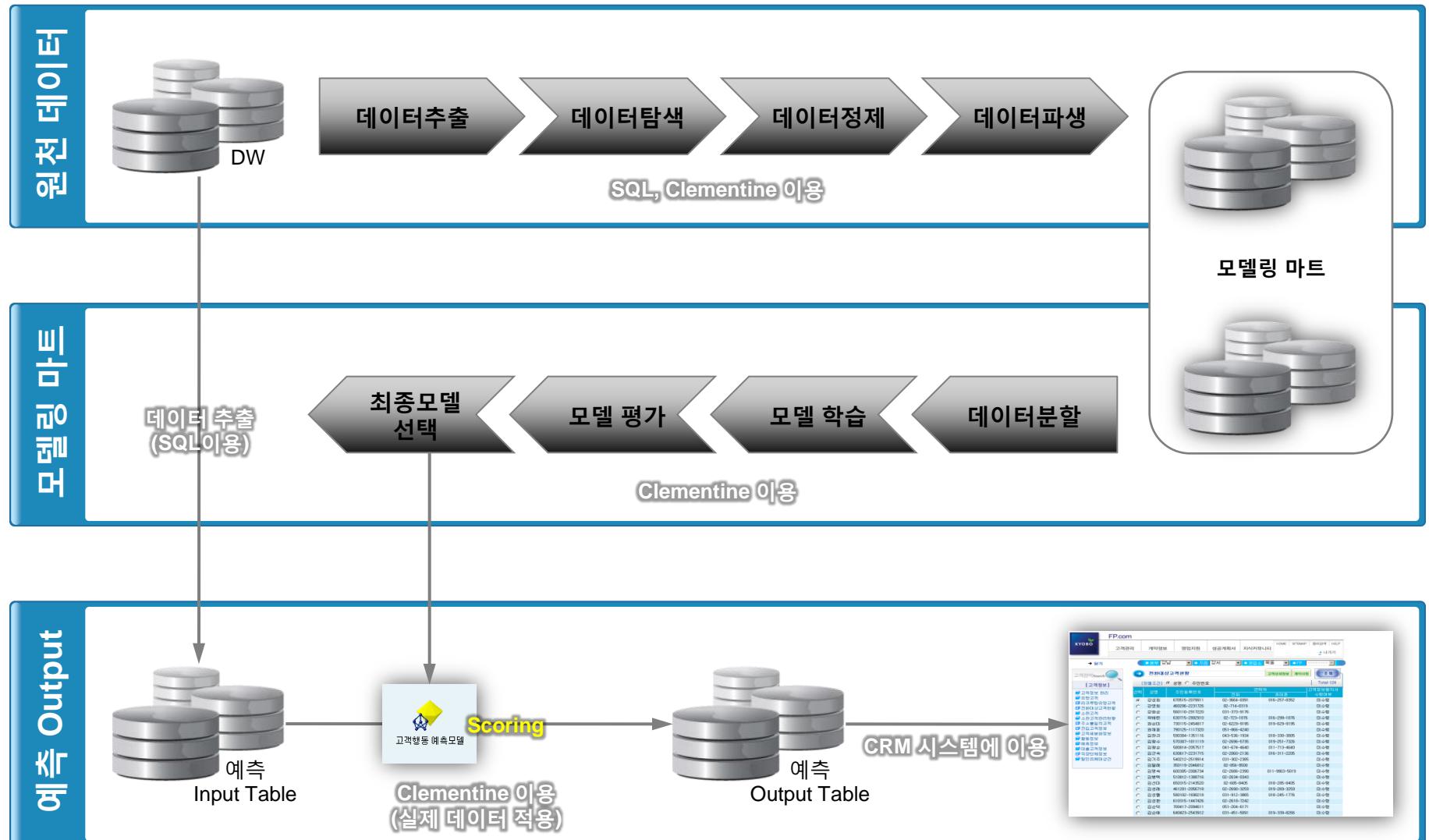
- SEMMA 방법론이란 SAS 기업에서 개발한 데이터마이닝 표준 가이드로써 Sample, Explore, Modify, Model, Assess의 단계로 되어있으며 각 5단계의 약자를 따서 만들었음
- SEMMA는 데이터마이닝을 구현하는데 있어서 하나의 가이드 역할을 할 수 있으며 5단계를 순차적으로 이루어 져 있음 (Sample, Explore, Modify, Model, Assess)
- SEMMA는 데이터 마이닝의 방법론이 아니라 SAS Enterprise Miner Tool의 작업을 수행하기 위한 기능적 논리 구성이다. 또한, SEMMA는 데이터 마이닝 모델 개발 측면에서 초점을 맞추고 있음.

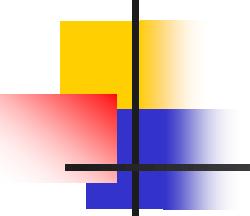
# Data Mining Process with SPSS Modeler

원천 데이터에 접근하여 각종 정제, 변환, 필드추가, 탐색, 모델링 및 평가를 통해 모델을 확정하고, 최종모델을 DB, Flat File, XML 코드로 전개한다.



# Predictive Modeling Process with SPSS Modeler



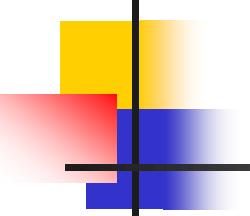


# 데이터마이닝 기법의 분류

- Supervised Modeling (지도학습, Predictive Analytics)
  - Estimation / Prediction (추정/예측: 연속형)
    - Linear Regression, Neural Network
  - Classification / Prediction (분류/예측: 이산형)
    - Decision Tree (C5.0) , Neural Network

용어의 유래: 어린아이가 말을 배우는 과정 (엄마가 *Supervisor* 역할)

- Unsupervised Modeling (비지도학습, Descriptive Analytics)
  - Clustering (군집화)
    - K-Means, SOM
  - Association rule mining (연관규칙탐사)
    - Apriori
  - Sequential rule mining(연속규칙탐사)

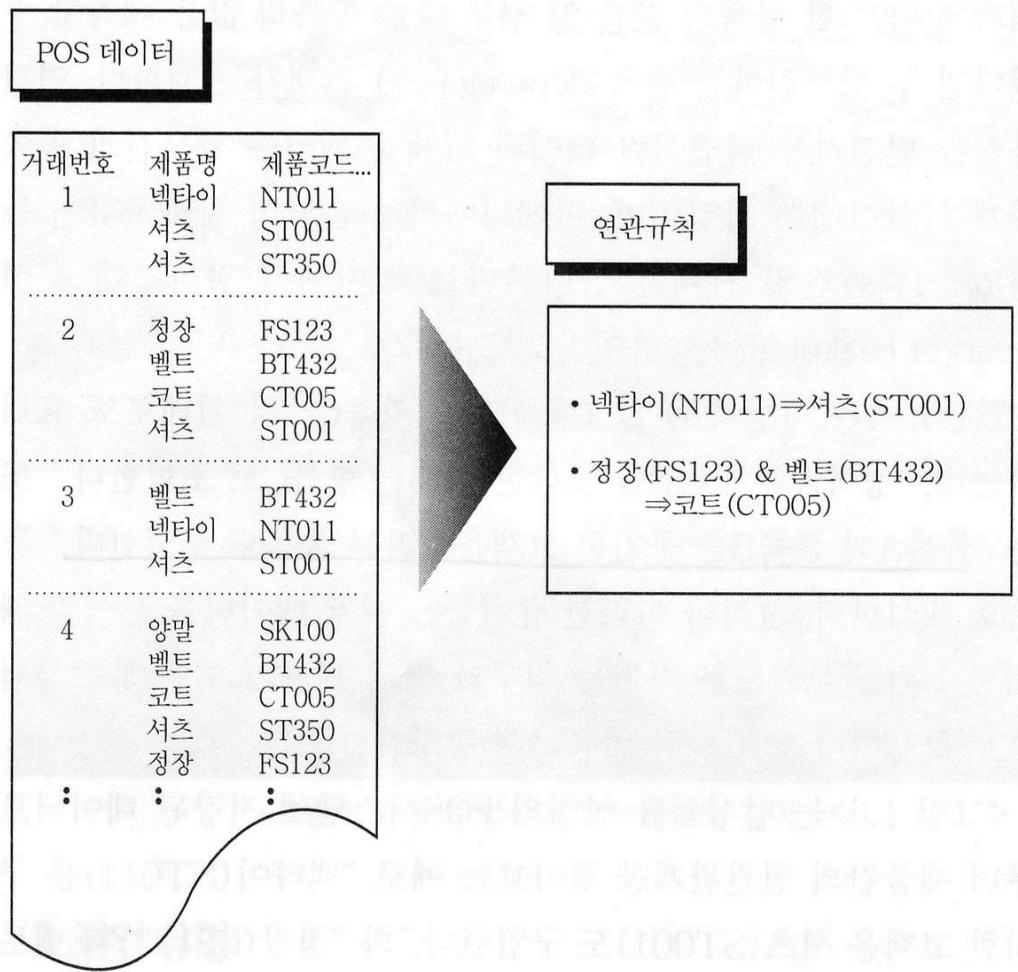


# 연관규칙탐사(Association Rule Mining)

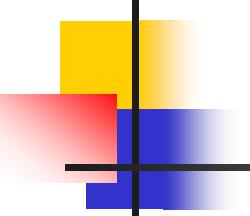
- 정의
  - 데이터 안에 존재하는 항목간의 종속 관계를 찾아내는 작업
- 장바구니 분석(market basket analysis)
  - 고객의 장바구니에 들어있는 품목 간의 관계를 발견
- 규칙의 표현
  - 항목 A와 품목 B를 구매한 고객은 품목 C를 구매한다.
  - (품목 A) & (품목 B)  $\Rightarrow$  (품목 C)
- 연관규칙의 활용
  - 제품이나 서비스의 교차판매
  - 매장진열, 첨부우편
  - 사기적발

# 연관 규칙

## ■ 연관 규칙의 예



[Source: 데이터마이닝, 장남식 외, 1999]

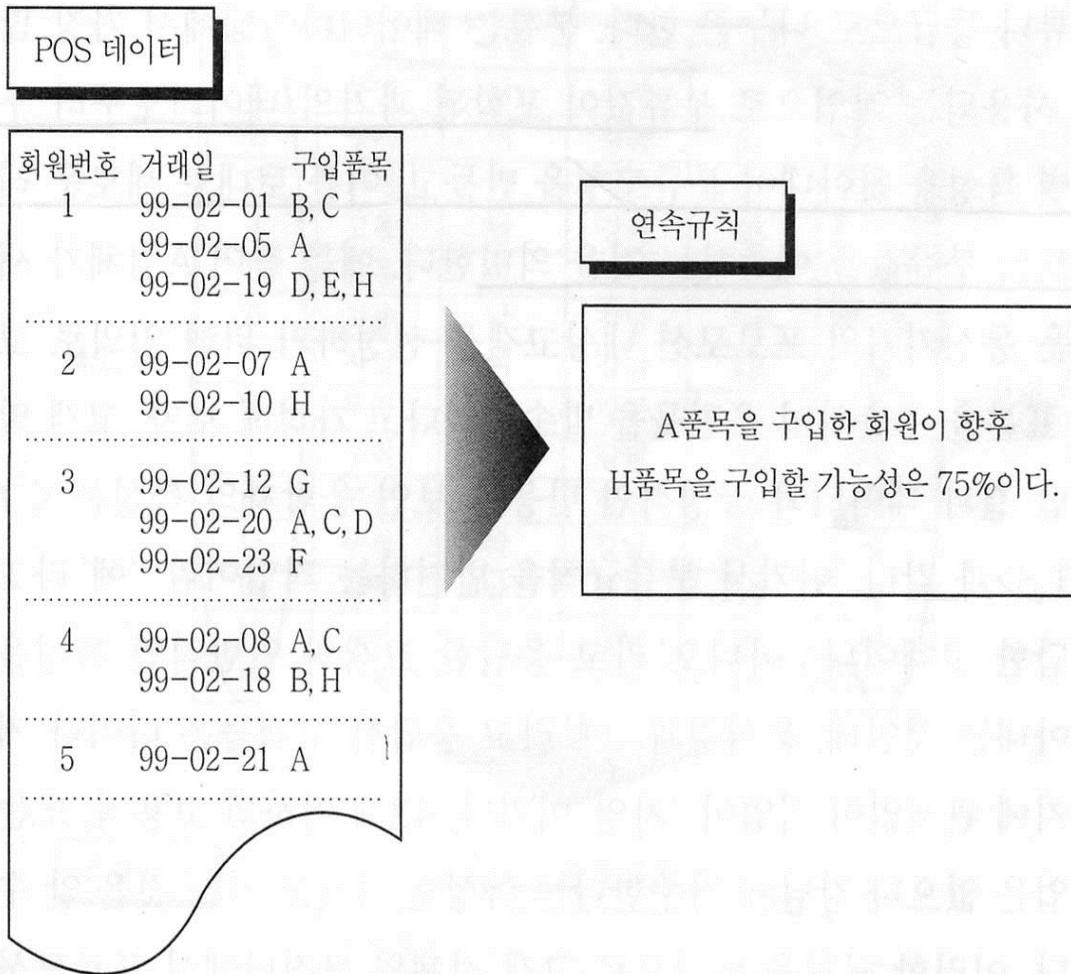


# 연속규칙탐사(Sequential Rule Mining)

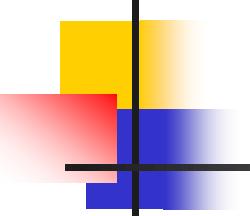
- 정의
  - 연관 규칙에 시간 관련 정보가 포함된 형태
- 규칙의 표현
  - 새 냉장고를 구입한 고객 중 한달 이내에 새 오븐을 구입하는 경향이 많다.
- 연속규칙의 활용
  - 타겟 마케팅
  - 일대일 마케팅
- 전제조건
  - 고객의 구매내역(history) 정보가 반드시 필요함

# 연속 규칙

## ■ 연속 규칙의 예



[Source: 데이터마이닝, 장남식 외, 1999]



# 분류(Classification)

## ■ 분류 프로세스

- 과거의 데이터를 부류로 구분
- 부류별 특성을 발견
- 분류 모형 생성
- 모형을 토대로 새로운 레코드의 분류 값 예측

## ■ 분류의 활용

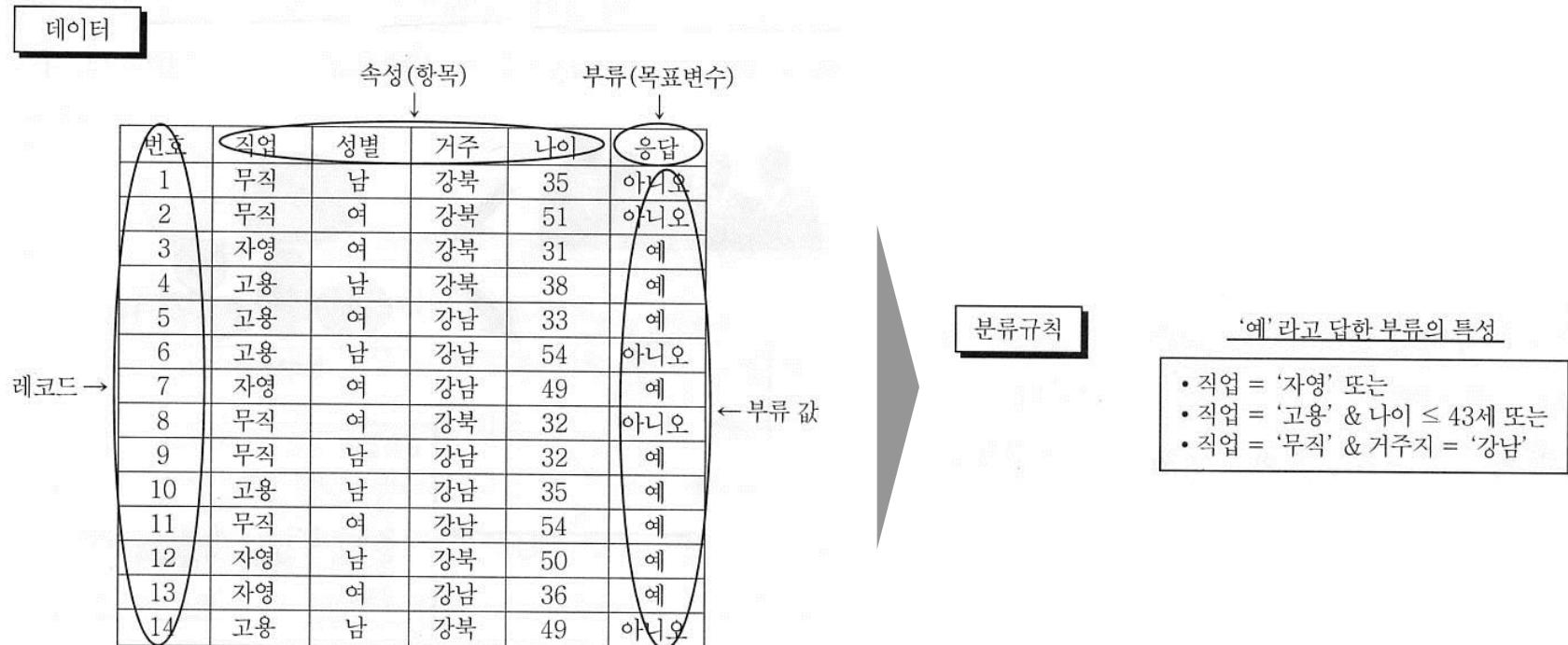
- 고객의 신용등급 분류
- 기업의 도산 예측
- 프로모션 대상고객 선정

## ■ 분류 기법

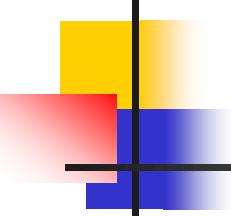
- 의사결정나무(Decision Tree)
- 인공신경망(Neural Network)

# 분류(Classification)

## ■ 분류의 예



[Source: 데이터마이닝, 장남식 외, 1999]

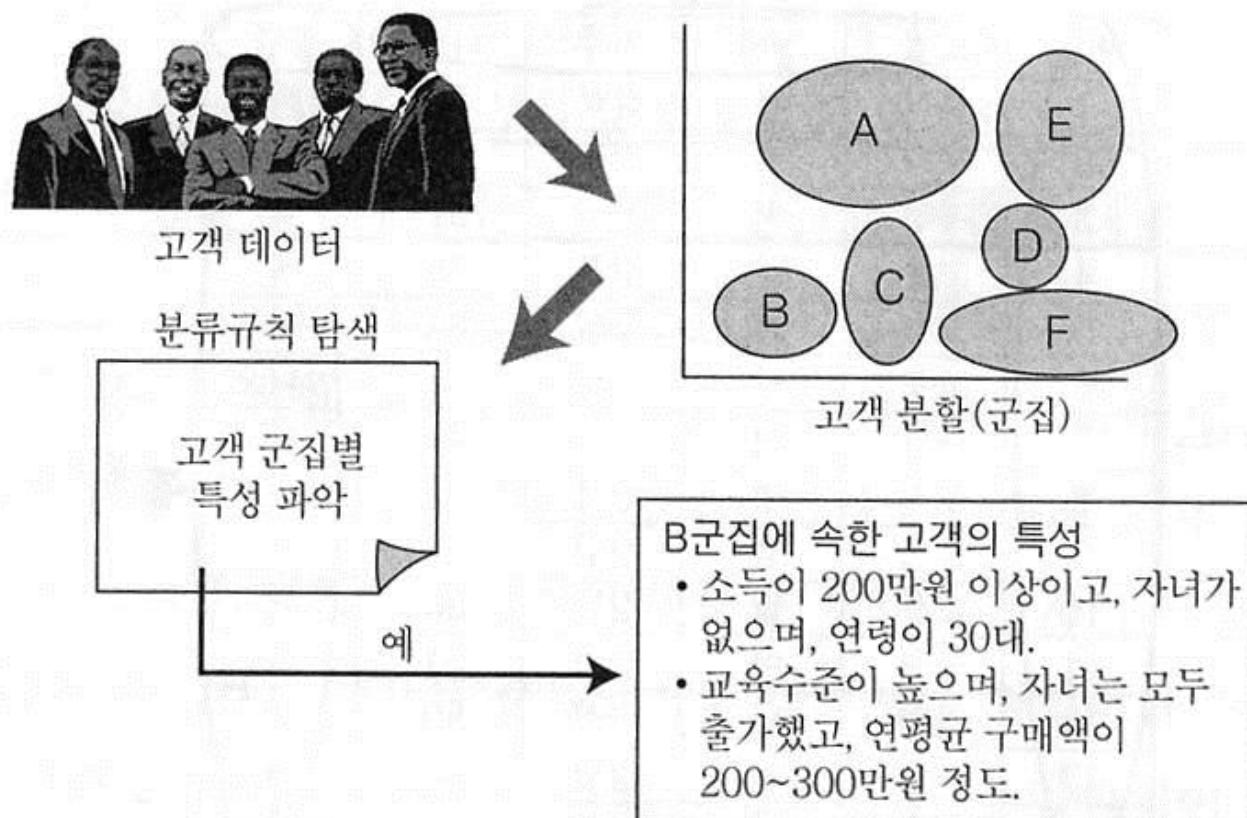


# 군집화(Clustering)

- 정의
  - 레코드들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업
- 군집화의 활용
  - 다른 데이터마이닝 기법의 선행 작업으로써 많이 이용
- 분류 vs 군집화
  - 분류 값의 유무

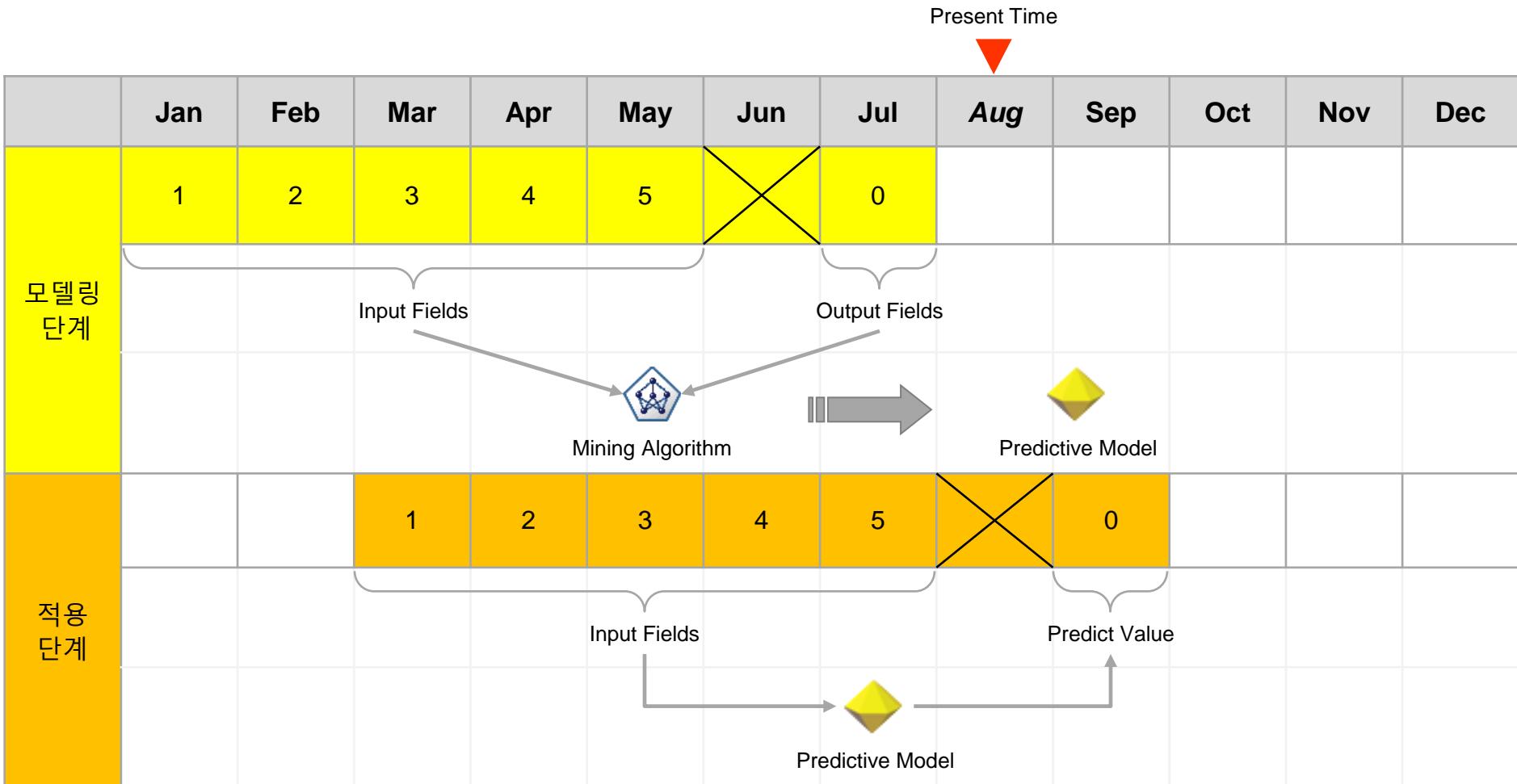
# 군집화(Clustering)

## ■ 군집화의 예



[Source: 데이터마이닝, 장남식 외, 1999]

# 예측 모델링 시점 및 모델 적용 시점



X marks the month of latency  
Numbers to left of X are months in the past

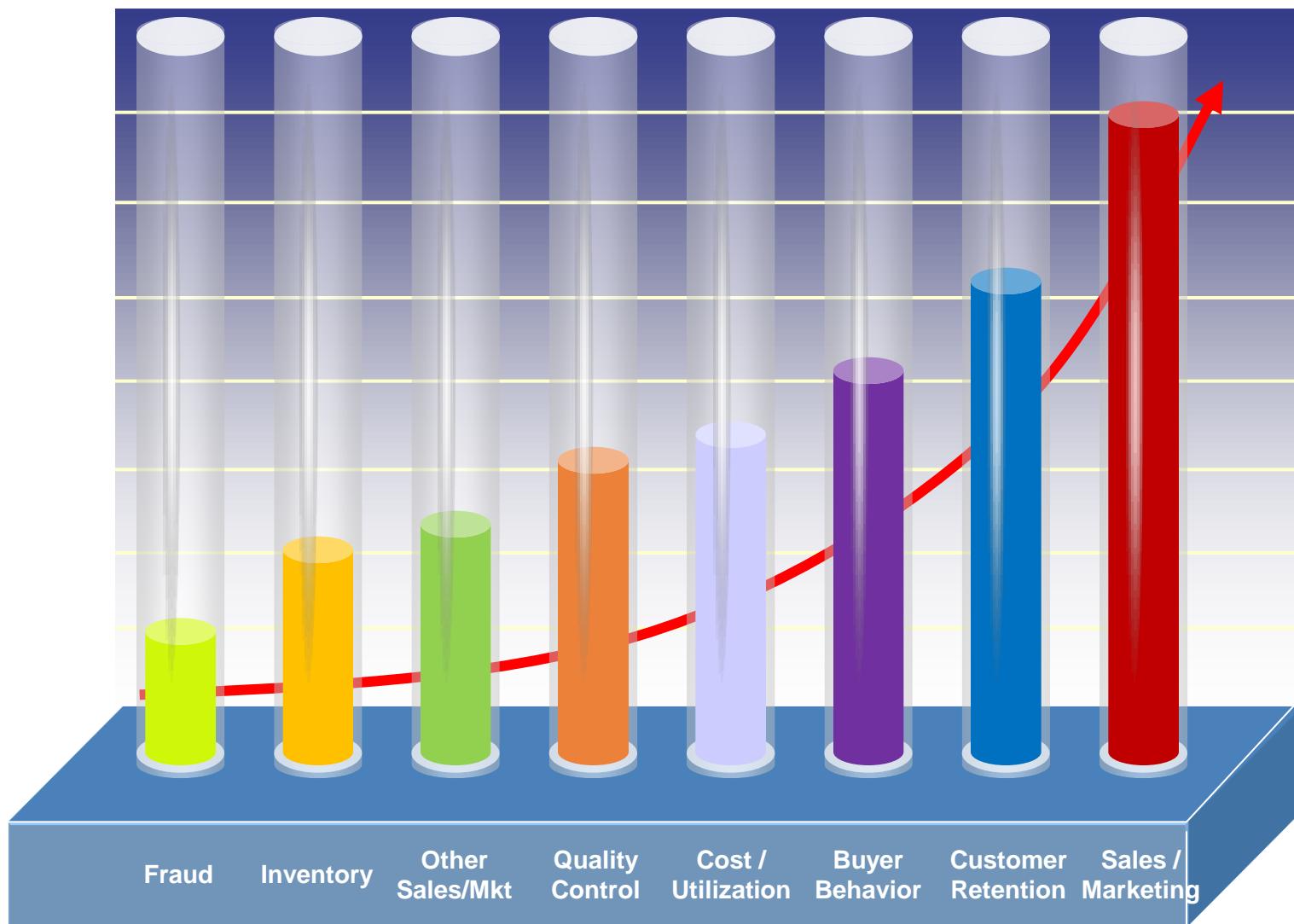
# 데이터 분할 (Data Partitioning)



- ❖ 데이터를 용도에 따라 분할
- ❖ 학습데이터 (training data) → 모델 적합
- ❖ 검증데이터 (test data) → 모델 평가
- ❖ 50% – 50% 분할
- ❖ 대안 : 60% – 40% 분할, 75% – 25% 분할

학습데이터 = Training data  
검증데이터 = Test data

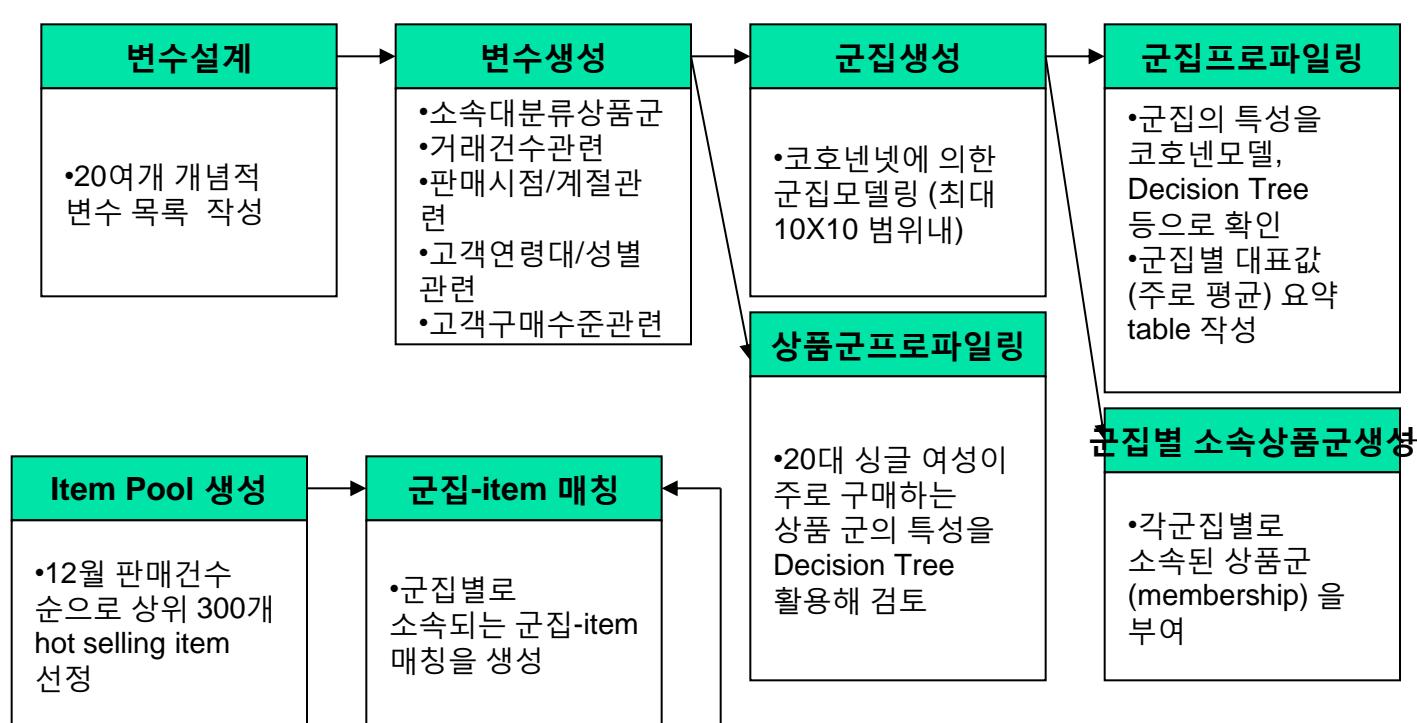
# Data Mining Applications



# Data Mining Cases: A社의 추천시스템

- 상품과 상품을 상품군 단위에서 군집한 후 동일 군집에 속하는 어떤 개별 item을 click/view 하거나 구매한 고객에게 해당 item이 속한 상품군 군집 내에 속한 다른 개별 item을 추천
- 상품군 선택은 군집화를 활용. 개별 item 선정(item sorting)은 최근판매순위(다른 기준 적용 가능) 사용

Recommendation Process Flow



# Data Mining Cases: A社의 추천시스템

## Association Analysis with Clementine

	consequent	Antecedent	Instances	Support %	Confidence %	Rule Support%	Lif	Deployability
188	A079115472	A078003730	1	1.163	100.000	1.163	86	0.000
189	A079552998	A077061413	1	1.163	100.000	1.163	86	0.000
190	A077061413	A079552998	1	1.163	100.000	1.163	86	0.000
191	A076293037	A076293241	1	1.163	100.000	1.163	86	0.000
192	A076293241	A076293037	1	1.163	100.000	1.163	86	0.000
193	A0714689954	A071468868	1	1.163	100.000	1.163	86	0.000
194	A077231302	A077145789	2	2.326	50.000	1.163	43	1.163
195	A077145789	A077231302	2	2.326	50.000	1.163	43	1.163
196	A076248144	A073946134	2	2.326	50.000	1.163	43	1.163
197	A076352317	A073946134	2	2.326	50.000	1.163	43	1.163
198	A077122352	A073946134	2	2.326	50.000	1.163	43	1.163
199	A074025508	A073946134	2	2.326	50.000	1.163	43	1.163
200	A065403381	A073946134	2	2.326	50.000	1.163	43	1.163
201	A079527593	A077145789	2	2.326	50.000	1.163	43	1.163
202	A079527593	A077231302	2	2.326	50.000	1.163	43	1.163
203	A075585631	A077145789	2	2.326	50.000	1.163	43	1.163
204	A075585631	A077231302	2	2.326	50.000	1.163	43	1.163
205	A076780606	A077145789	2	2.326	50.000	1.163	43	1.163
206	A077102665	A077231302	2	2.326	50.000	1.163	43	1.163
207	A077034904	A071365252	2	2.326	50.000	1.163	43	1.163
208	A072172559	A076831460	2	2.326	50.000	1.163	43	1.163
209	A077231302	A077102665	2	2.326	50.000	1.163	43	1.163
210	A073676775	A073946134	2	2.326	50.000	1.163	43	1.163
211	A075409254	A071365252	2	2.326	50.000	1.163	43	1.163
212	A078091839	A077145789	2	2.326	50.000	1.163	43	1.163
213	A076780606	A077231302	2	2.326	50.000	1.163	43	1.163
214	A075585631	A077145789	2	2.326	50.000	1.163	43	1.163
215	A073316796	A070864473	3	3.488	33.333	1.163	29	2.325
216	A073410585	A070864473	3	3.488	33.333	1.163	29	2.325
217	A072513619	A070864473	3	3.488	33.333	1.163	29	2.325
218	A073946134	A073676775	3	3.488	33.333	1.163	29	2.325
219	A072704937	A070864473	3	3.488	33.333	1.163	29	2.325
220	A067938510	A070864473	3	3.488	33.333	1.163	29	2.325

Table Annotations

▶ 경매 이용고객의 특성과 상품의 특성을 고려하고 데이터마이닝 기법인 Association Analysis를 활용하여 개별 경매 물건간의 동일 고객 구매 관계를 발견

## Cases of Association Analysis in Auction

 <p>【마름보이 큰청바지】  빅사이즈 와일드청바지 힙합청바지 통바지 남녀공용 2장이상 충일무료배송 <a href="#">미리보기 열기</a></p> <p><b>A079552998</b></p>	<b>CON:</b> <b>100.000</b> <b>SUP:</b> 1.163 <b>LIF:</b> 86.000
 <p>【마름보이 컨티후드티】  빅사이즈 후드짚업티셔츠 큰웃기슴둘레142까지 2007년굿바이세월 5000원할인 세월7길 <a href="#">미리보기 열기</a></p> <p><b>A077061413</b></p>	<b>CON:</b> <b>35.000</b> <b>SUP:</b> 10,800 <b>LIF:</b> 132
 <p>【컬러스카니전】  힐리웃&amp;남쁘해보이는 스키니진 면스카니전 26~36 <a href="#">미리보기 열기</a></p> <p><b>A078091838</b></p>	<b>CON:</b> <b>18,500</b> <b>SUP:</b> 18,500 <b>LIF:</b> 124
 <p>【모파상(빅사이즈)】  견장나트가디건/기모노소매 어깨견장 가디건 볼신상품 L(66)~XXL(120) <a href="#">미리보기 열기</a></p> <p><b>A077145789</b></p>	<b>CON:</b> <b>15,000</b> <b>SUP:</b> 2.326 <b>LIF:</b> 43.000

▶ 연관규칙 발견 방식과 고객/상품 군집화 등 복수의 방식에 의해 수백만 Item중 적절한 상품을 추천하는 로직을 개발

# Data Mining Cases: B社의 URC-Selling 모델



과거 거래 경험 Data 기반

**Up-Selling Mining모델(지도학습 모델)**

**Up-Selling**

High Plus 캠페인

**Re-Selling Mining모델(주기 모델)**

**Re-Selling**

Cycle 캠페인

**Cross-Selling Mining모델(연관성 모델)**

**Cross-Selling**

X 캠페인

**Up-Selling Mining모델(지도학습 모델)**

- ✓ 고객의 주 가격대를 계산
- ✓ 고가격대의 고객 특성 파악 (의사결정나무 분석)
- ✓ 저가격대의 고객 중 고가격대의 고객과 동일한 특성을 가지는 고객 파악(오류 모델)

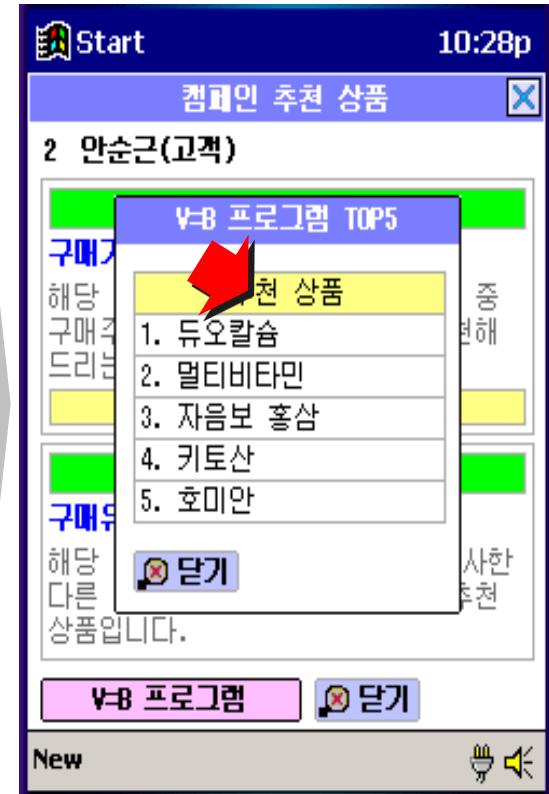
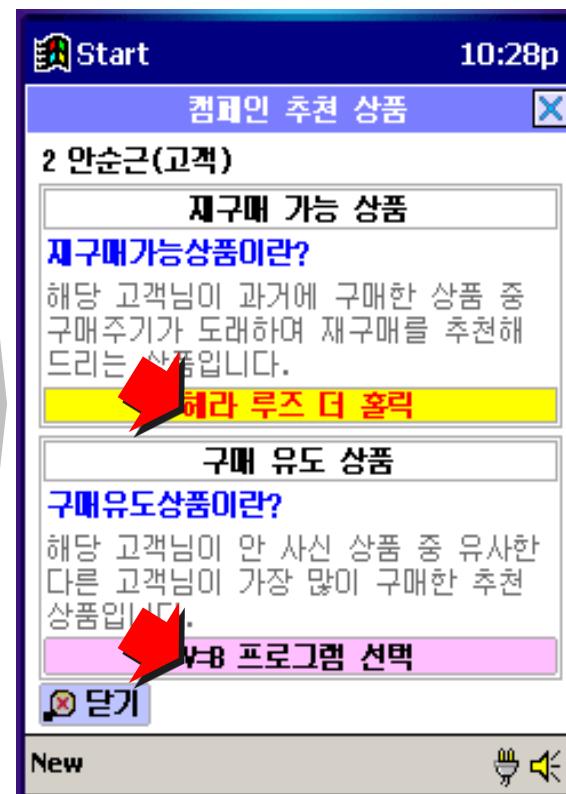
**Re-Selling Mining모델(주기 모델)**

- ✓ 고객별 또는 고객/상품별 평균 주기 계산
- ✓ 최종 구매일 + 평균 주기일수 - 3일 전 날짜를 계산
- ✓ 해당 날짜에 캠페인 수행

**Cross-Selling Mining모델(연관성 모델)**

- ✓ 고객이 1회 구매한 상품 데이터 정리
- ✓ 1회 구매 시 교차한 상품들을 연관성 분석을 수행함
- ✓ 특정 상품과 가장 연관성이 높은 상품 선정(with 확률 값)

# Data Mining Cases: B社의 URC-Selling 모델



# Data Mining Cases: B社의 URC-Selling 모델

**NEXT system**

나의메뉴 | 판매/수금 | 주문/매입 | 재고/회계 | 카운셀러인사 | 카운셀러평가보상 | 기본정보 | 영업계획 | 영업분석  
 관리지표 | 역매관리 | 캠페인관리 | 카운셀러포인트관리 | 고객마일리지관리

영업계획분석 :: 현재위치 | 홈 > 영업계획 > 캠페인관리 > 캠페인조회(수정)

● 캠페인조회(수정) AmorePacific NEXT system

**영업계획**

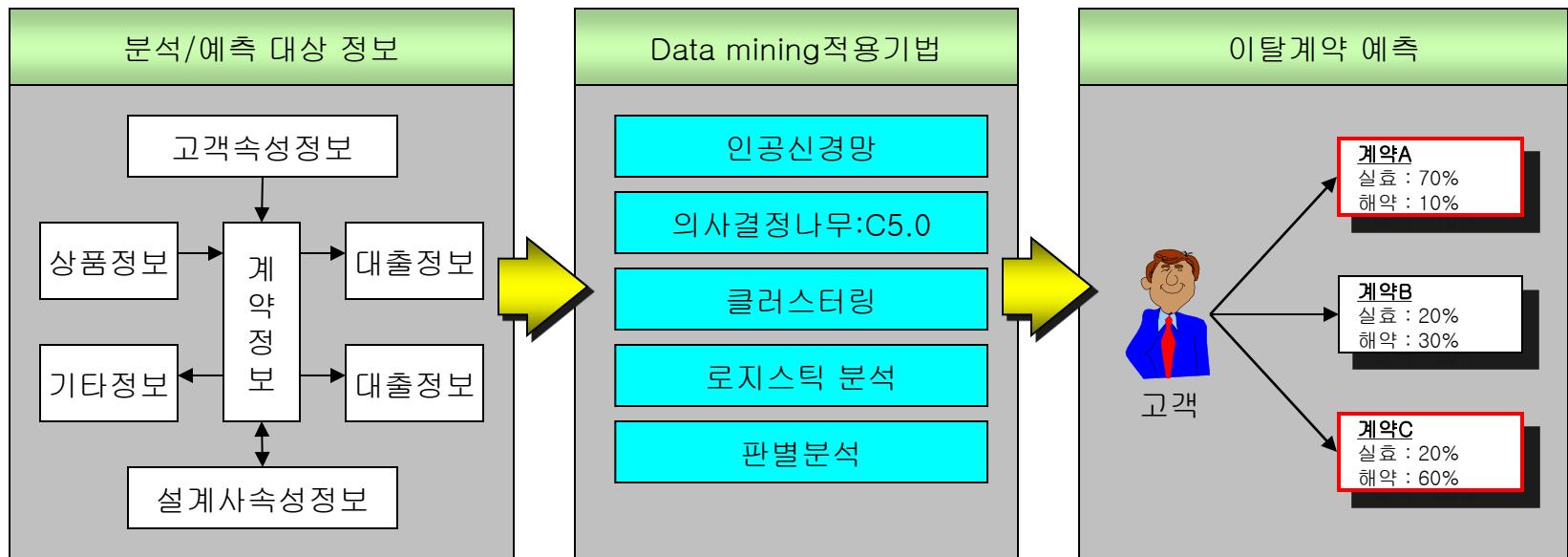
- 관리지표
- 역매관리
- 캠페인관리**
  - 캠페인등록
  - 캠페인조회(수정) ■
  - 캠페인결과조회
  - 캠페인정보보류고객 조회
  - 캠페인설적조회
  - 캠페인인증번호확인 등록
  - 캠페인인증번호확인 현황
  - 캠페인기간별설적조회
  - 캠페인고객유지조회
- 카운셀러포인트관리

캠페인 코드: 200706-10001 | 캠페인명: [고객]거래우수고객\_6월합계 | 조회 | 엑셀 | 도움말

카운셀러코드:  ~

제외	카운셀러코드	고객번호	주 소	기간구매수량	기간구매금액
	카운셀러명	고객명	전화번호	재구매가능상품	구매유도상품
<input type="checkbox"/>	V-10001	3	경기	0	0
<input type="checkbox"/>	권	최		0	0
<input type="checkbox"/>	V-10002	3	경기	0	0
<input type="checkbox"/>	권	이		0	0
<input type="checkbox"/>	V-10003	3	경기	0	0
<input type="checkbox"/>	권	김		0	0
<input type="checkbox"/>	V-10004	3	경기	0	0
<input type="checkbox"/>	권	김		0	0
<input type="checkbox"/>	V-10005	2	경기	0	0
<input type="checkbox"/>	보	구		0	0
<input type="checkbox"/>	V-10006	2	경기	0	0
<input type="checkbox"/>	보	정		0	0
<input type="checkbox"/>	V-10007	2	경기	0	0
<input type="checkbox"/>	보	구		0	0
V31002 40 경기 동드체 시 생여동 자동바 생여린 102호					

# Data Mining Cases: K社의 이탈고객관리



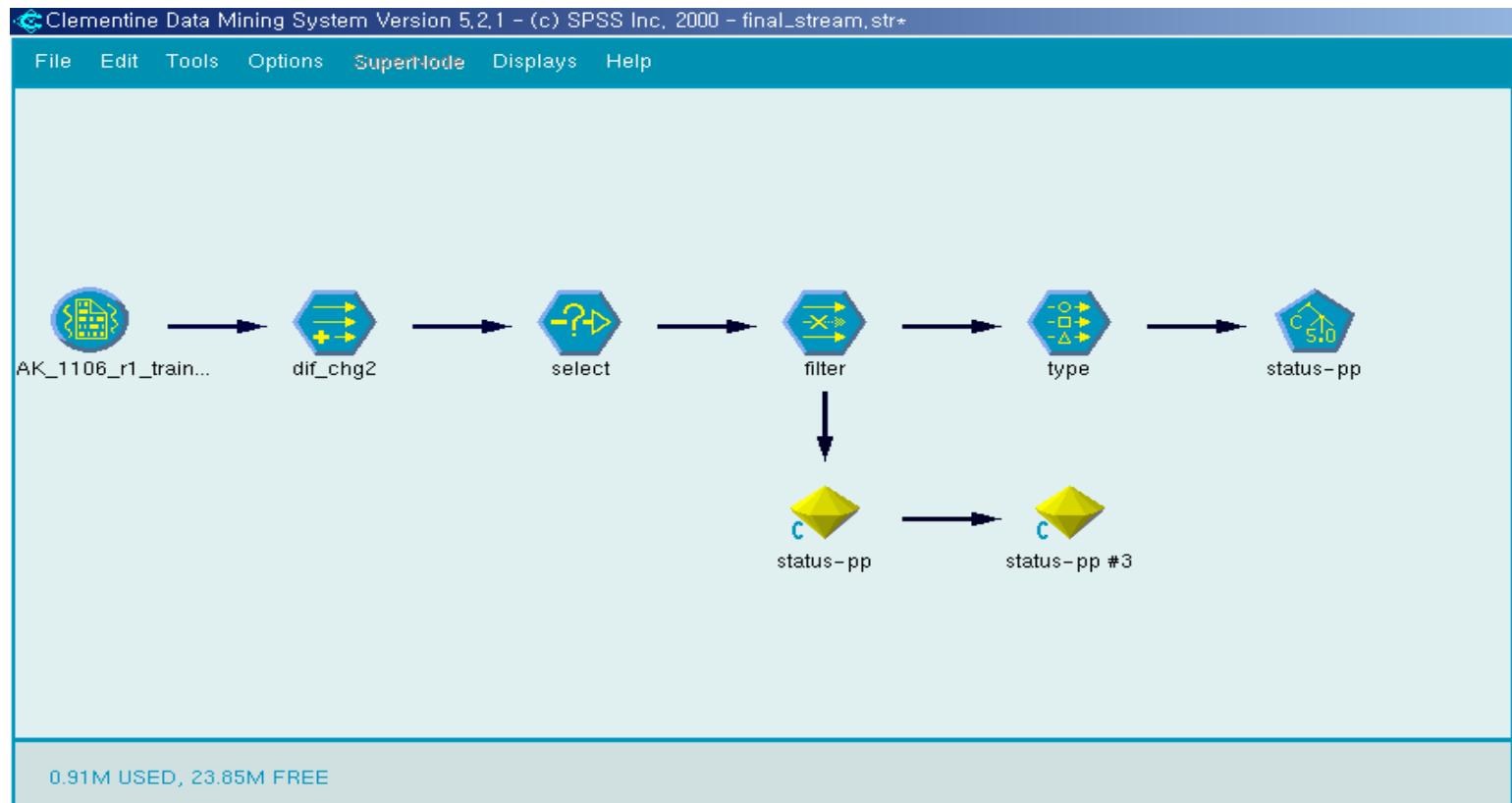
# Data Mining Cases: K社의 이탈고객관리

## ■ 모델설계: 입력 변수

구분	변수명	개수
고객속성	이탈시 연령-계약자	1
계약속성	납입방법, 납입기간, 보험기간, 주보험금, 수금방법, 무배당여부코드, 이탈시 연령-주피보험자, 유지건수율, 직전6개월 실효후해약건(재해), 주요 수금방법(전체)	10
거래속성	납입횟수_할인횟수, 기준일과 계약시기와의 거리, 총납입기간(유지), 총납입기간(실효), 최근1년동안의 총 연체횟수, 기준일과 최종 유지계약일과의 거리	6

# Data Mining Cases: K社의 이탈고객관리

- 적용 데이터마이닝 도구 : SPSS Clementine
- 적용 데이터마이닝 기법 : Decision Tree



# Data Mining Cases: K社의 이탈고객관리

## ■ 모델 평가

모델 검증 결과			89,287	
예측결과				
실제반응	유지	해약		
	유지	50,931	59,391	
	해약	10,909	29,896	
	61,840	27,447		

• 해약 비율 :  $29,896 / 89,287 = 33.5\%$   
• 해약 적중률 :  $18,987 / 29,896 = 63.5\%$   
• 유지 적중률 :  $50,931 / 59,391 = 85.8\%$   
• TOTAL :  $69,918 / 89,287 = 78.3\%$

# Data Mining Cases: K社의 이탈고객관리

## ■ 모델 분석 및 전개: 도출된 해약 규칙

Rule #1 for 해약

if 유지건수율 <= 8  
then -> 해약 (7307.6, 0.942)

조건을 만족하는 7,307 건의 계약 중 94.2%에 해당하는 계약이 ‘해약’으로 분류

Rule #3 for 해약

if 유지건수율 > 8  
and 납입방법 == 1  
and 직전6개월 실효후해약건(재해) < 0  
and 납입기간 > 4  
and 총납입기간(실효) < 12  
and 보험기간 > 17  
and 무배당여부코드 == 0  
and 기준일과 계약시기와의 거리 > 976  
and 기준일과 계약시기와의 거리 < 4081  
and 주보험금 > 6700  
and 기준일과 최종유지계약일과의 거리 > 83  
and 납입횟수\_할인횟수 <= 39  
and 이탈시연령-주피보험자 <= 45  
and 총납입기간(유지) < 126  
and 주요수금방법(전체) == 4  
then -> 해약 (316.0, 0.766)

Rule #2 for 해약

Rule #2 for 해약:  
if 유지건수율 > 8  
and 납입방법 == 1  
and 직전6개월 실효후해약건(재해) < 0  
and 납입기간 <= 4  
and 수금방법 == 1  
then -> 해약 (8444.8, 0.732)

조건을 만족하는 8,444 건의 계약 중 73.2%에 해당하는 계약이 ‘해약’으로 분류

조건을 만족하는 316건의 계약 중 76.6%에 해당하는 계약이 ‘해약’으로 분류

# 이탈고객관리: 모델 분석 및 전개(2)

## ■ 모델 분석 및 전개: 도출된 해약 규칙

Rule #4 for 유지

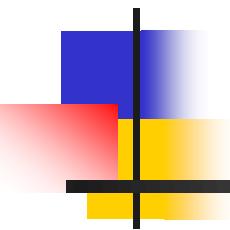
```
if 유지건수율 > 8  
and 납입방법 == 1  
and 직전6개월 실효후해약건(재해) =< 0  
and 납입기간 > 4  
and 총납입기간(실효) =< 12  
and 보험기간 > 7  
and 무배당여부코드 == 1  
and 기준일과 계약시기와의 거리 =< 1232  
and 수금방법 == 2  
then -> 유지 (38053.4, 0.776)
```

조건을 만족하는 38,053명의 고객 중 77.6%에 해당하는 고객이 '유지'로 분류

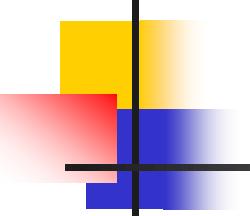
Rule #5 for 유지

```
if 유지건수율 > 8  
and 납입방법 == 1  
and 직전6개월 실효후해약건(재해) =< 0  
and 납입기간 > 4  
and 총납입기간(실효) =< 12  
and 보험기간 > 7  
and 무배당여부코드 == 1  
and 기준일과 계약시기와의 거리 =< 1232  
and 수금방법 == 4  
then -> 유지 (283.9, 0.754)
```

조건을 만족하는 286명의 고객 중 75.4%에 해당하는 고객이 '유지'로 분류



# **IBM SPSS Modeler를 활용한 금융업체의 캠페인분석 실습**

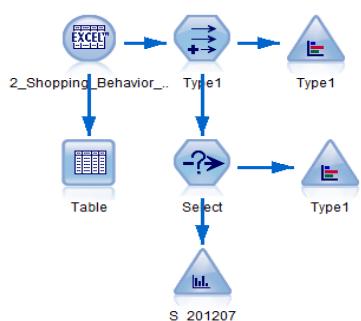


# IBM SPSS Modeler의 특징

- 1986년 세계 최초로 개발된 SPSS社의 Data Mining Solution
- 데이터마이닝 Workbench로 상호작용적인 데이터마이닝 프로세스를 지원하기 위한 다양한 도구와 기술을 병합시킨 작업환경을 제공한다.
- IBM SPSS Modeler은 데이터마이닝 기술자 중심이 아니라 비즈니스 지식을 접목시키는 사용자 중심으로 설계되어 있다.
- Release History
  - Clementine (Version 12까지)
  - PASW Modeler (Version 13)
  - IBM SPSS Modeler (Version 13이후): 현재버전 15.0

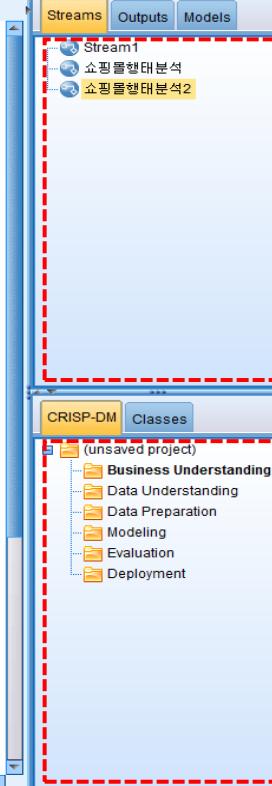
# IBM SPSS Modeler의 GUI

메뉴 및 도구 막대



작업 창  
(Stream Canvas)

: 데이터 처리 및 모델링



관리자 창  
(Manager Window)

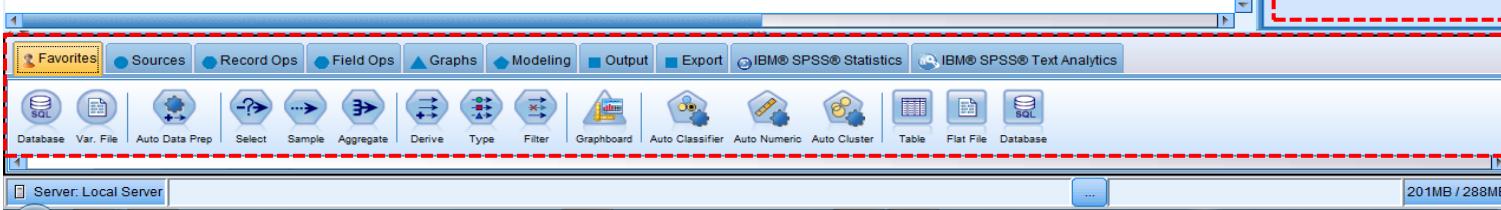
: 프로그램 실행 후 사용된 스트림 및 출력 결과, 생성된 모델 결과 표시

프로젝트 창  
(Project Window)

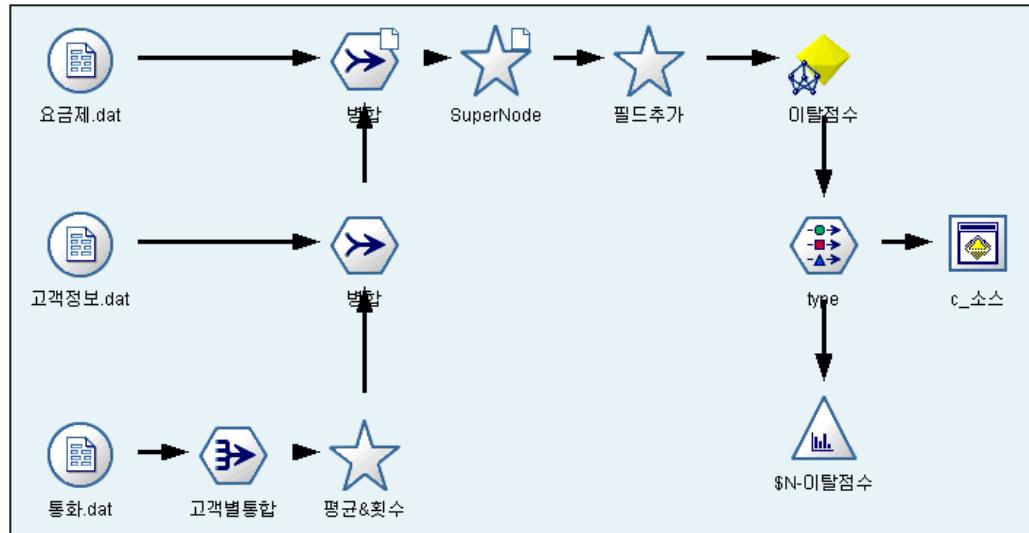
: 방법론(CRISP-DM)에 따른 프로젝트 단위의 개체 표시

노드 팔레트  
(Node Palette)

: 유사한 속성을 갖는 노드들의 그룹으로 구성



# 실행과정 – Visual Programming



- ✓ 팔레트의 각 노드 아이콘을 선택한 후, 작업 창(Stream Canvas)에 배치하고 노드들을 연결하여 노드들의 속성을 정의, 스트림(Stream)을 작성한 뒤, 이를 실행한다.

팔레트의 각 노드 선택 → 작업 창에 배치 → 노드 연결 → 노드 속성 정의 → 실행

# 마우스 및 키보드 기본 동작

## 가운데 버튼

**drag & drop:** 노드 연결

**double click:** 노드 연결 해제

## 오른쪽 버튼

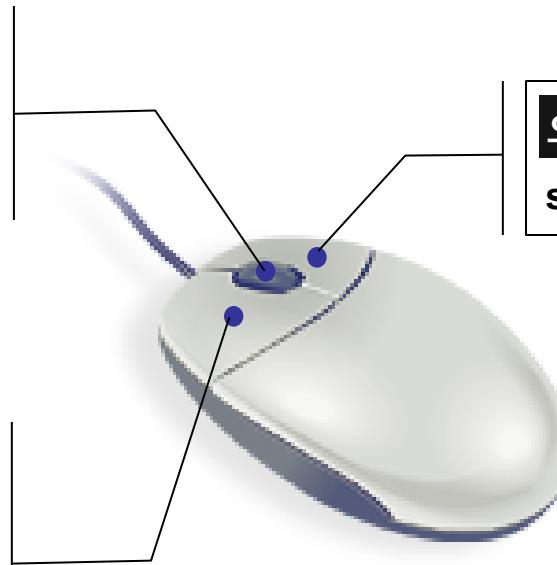
**single click:** 팝업 메뉴 호출

## 왼쪽 버튼

**single click:** 노드 선택

**drag & drop:** 이동

**double click:** 속성편집(대화상자 호출)



## 키보드 관련 Information

✓ **Delete 키** : 노드 및 연결 삭제

✓ **CTRL + V** : 노드 붙여넣기

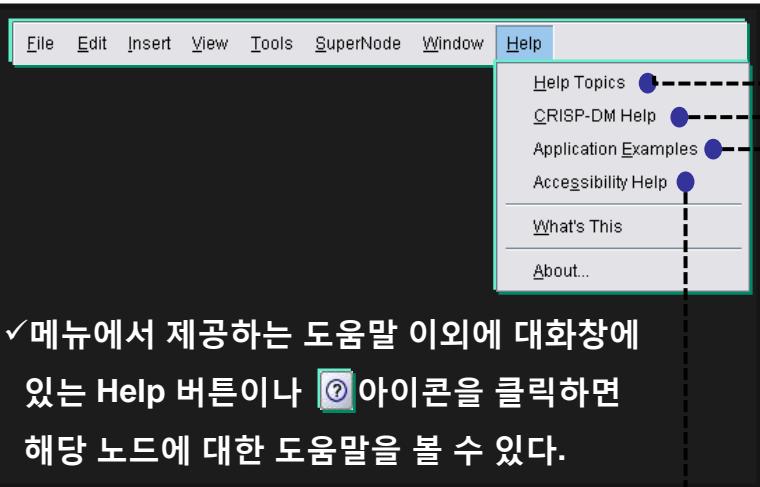
✓ **CTRL + E** : 스트림 실행

✓ **F2** : 노드 연결

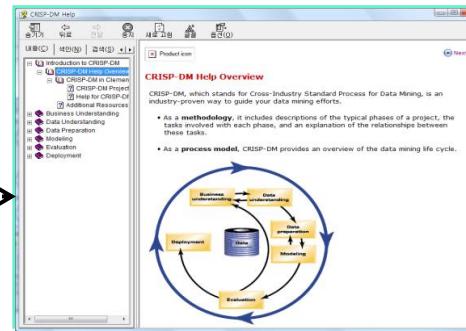
✓ **CTRL + C** : 노드복사

✓ **F3** : 노드 연결 해제

# 도움말

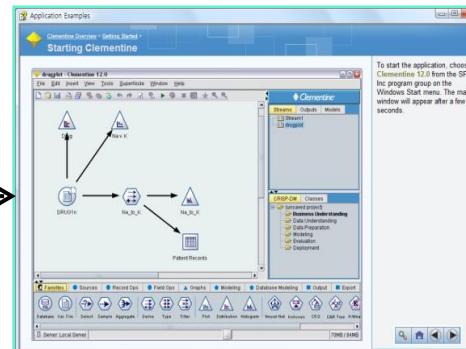
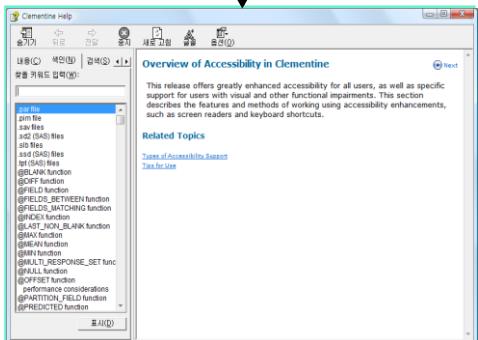


- ✓ ‘목차’와 ‘색인’으로 구성되어 있으며, 보통 ‘색인’을 통해 사용자가 원하는 항목을 직접 입력하여 도움말을 찾아 볼 수 있다.



- ✓ 데이터 마이닝 프로세스의 표준인 CRISP-DM에 관한 도움말을 제공한다.

✓ 키보드 단축키와 같이 작업하는데 편리한 사용자 팀 등에 대한 정보를 얻을 수 있다.



- ✓ SPSS Modeler를 처음 접한 사용자가 가장 빨리 기본적인 기능을 익힐 수 있는 온라인 자습서이다.

# Node의 종류

구분	
	<b>Sources</b>
	<b>Operation</b>
	<b>Graphs</b>
	<b>Modeling</b>
	<b>Output</b>



Super



## ▪ Sources Node

- 1) 데이터 연결 노드
- 2) 데이터베이스 연결 또는 가변형식, 고정형식 파일의 데이터, SPSS, SAS 파일 등의 다양한 파일들을 데이터로 읽어온다.

## ▪ Operations Node

- 1) 데이터 변환 작업 노드
- 2) 샘플링, 레코드 또는 필드 단위의 데이터 병합 및 필터, 변수파생, 모형평가를 위한 파티션 작업 등이 포함된다.

## ▪ Graphs Node

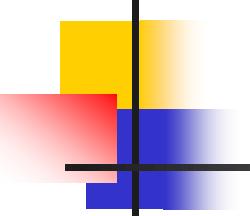
- 1) 데이터 도식화 노드
- 2) 크게 데이터 탐색으로 이용되는 히스토그램, 2차원 및 3차원 도표와 ROI Chart 등과 같은 평가도표로 이용된다.

## ▪ Modeling Node

- 1) 데이터 모형화 노드
- 2) Decision Tree, Regression, Neural Network, Clustering, Association 등 다양한 종류가 이용된다.

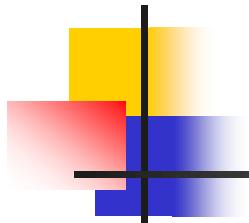
## ▪ Output Node

- 1) 마이닝 결과 출력 노드
- 2) 최종 결과를 테이블, 외부파일로 출력하는 기능, 생성된 모델들 간의 예측력을 평가하는 기능 등이 포함된다.



# 신규상품의 DM 프로모션

- 한 은행이 새로운 개인연금상품(PEP)을 신설하여 기존 고객들을 대상으로 가능한 많은 계좌를 유치하고자 한다
- 고객의 금융상품(PEP: Personal Equity Plan, 연금보험) 구매 여부 예측에 의한 신규고객 창출
  - 고객 프로파일 개발
  - 다이렉트 메일 광고 효율성 제고
  - 타겟 메일링에 의한 응답률 제고
- 분석 절차
  1. 기존고객 DB로부터 시험메일 발송을 위한 표본고객목록을 추출
  2. 새로운 금융상품(PEP)의 제안 메일을 발송
  3. 고객의 반응을 기록
  4. SPSS Modeler를 이용하여 캠페인 결과를 분석



# 캠페인 데이터의 구성

- 학습용 데이터 300건 (`snapshottrainN.db`)
- 검증용 데이터 300건 (`snapshottestN.db`)
- 신규고객 데이터 200건 (`newcustomersN.db`)

# 데이터 연결 & 확인

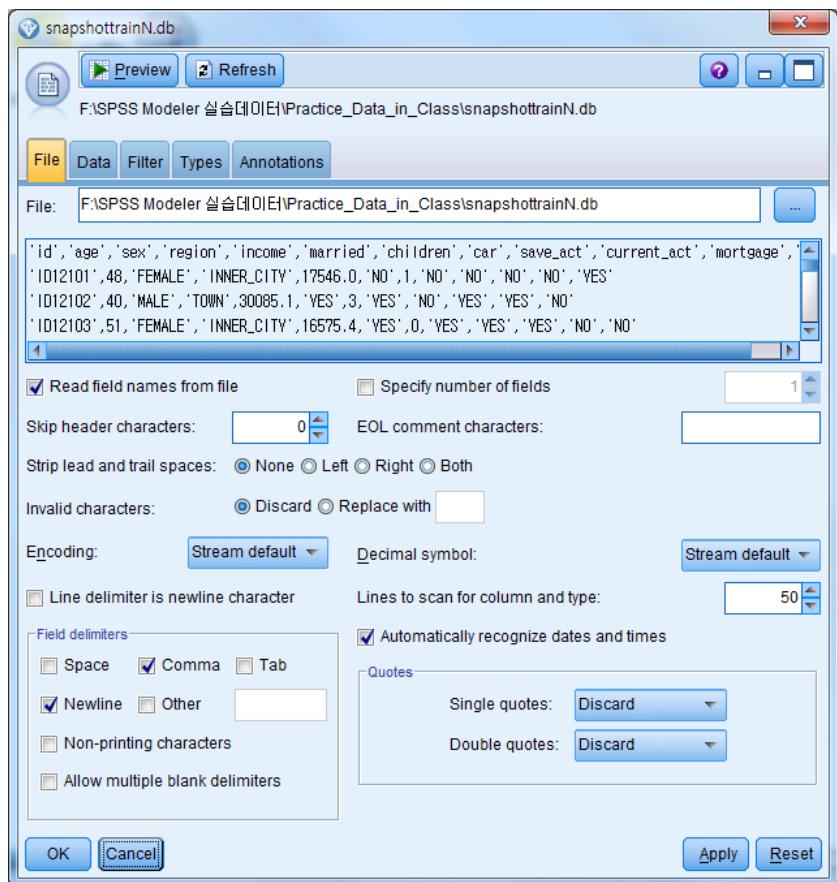


Diagram illustrating the connection between a database and a table:

```

graph LR
    Database[snapshottrainN.db] --> Table[Table]
  
```

The diagram shows a blue circular icon with a document symbol connected by a blue arrow to a blue square icon with a grid symbol. Below the icons, the text "snapshottrainN.db" is followed by an arrow pointing to the word "Table".

Table (12 fields, 300 records) #1

	id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
1	ID12101	48	FEMALE	INNER_CITY	17546...	NO	1	NO	NO	NO	NO	YES
2	ID12102	40	MALE	TOWN	30085...	YES	3	Y...	NO	YES	YES	NO
3	ID12103	51	FEMALE	INNER_CITY	16575...	YES	0	Y...	YES	YES	NO	NO
4	ID12104	23	FEMALE	TOWN	20375...	YES	3	NO	NO	YES	NO	NO
5	ID12105	57	FEMALE	RURAL	50576...	YES	0	NO	YES	NO	NO	NO
6	ID12106	57	FEMALE	TOWN	37869...	YES	2	NO	YES	YES	NO	YES
7	ID12107	22	MALE	RURAL	8877.0...	NO	0	NO	NO	YES	NO	YES
8	ID12108	58	MALE	TOWN	24946...	YES	0	Y...	YES	YES	NO	NO
9	ID12109	37	FEMALE	SUBURBAN	25304...	YES	2	Y...	NO	NO	NO	NO
10	ID12110	54	MALE	TOWN	24212...	YES	2	Y...	YES	YES	NO	NO
11	ID12111	66	FEMALE	TOWN	59803...	YES	0	NO	YES	YES	NO	NO
12	ID12112	52	FEMALE	INNER_CITY	26658...	NO	0	Y...	YES	YES	YES	NO
13	ID12113	44	FEMALE	TOWN	15735...	YES	1	NO	YES	YES	YES	YES
14	ID12114	66	FEMALE	TOWN	55204...	YES	1	Y...	YES	YES	YES	YES
15	ID12115	36	MALE	RURAL	19474...	YES	0	NO	YES	YES	YES	NO

# 새로운 필드의 생성

Derive

Derive as: Conditional

Settings Annotations

Mode:  Single  Multiple

Derive field:  
realincome

Derive as: Conditional

Field type:  <Default>

If:  
children=0

Then:  
income

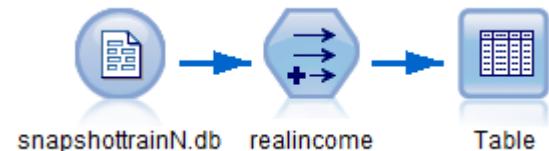
Else:  
income / (1+0.5 \* children)

OK Cancel

Table (13 fields, 300 records)

	car	save_act	current_act	mortgage	pep	income	realincome
1	NO	NO	NO	NO	YES	17546.000	11697.333
2	Y...	NO	YES	YES	NO	30085.100	12034.040
3	Y...	YES	YES	NO	NO	16575.400	16575.400
4	NO	NO	YES	NO	NO	20375.400	8150.160
5	NO	YES	NO	NO	NO	50576.300	50576.300
6	NO	YES	YES	NO	YES	37869.600	18934.800
7	NO	NO	YES	NO	YES	8877.070	8877.070
8	Y...	YES	YES	NO	NO	24946.600	24946.600
9	Y...	NO	NO	NO	NO	25304.300	12652.150
10	Y...	YES	YES	NO	NO	24212.100	12106.050
11	NO	YES	YES	NO	NO	59803.900	59803.900
12	Y...	YES	YES	YES	NO	26658.800	26658.800
13	NO	YES	YES	YES	YES	15735.800	10490.533
14	Y...	YES	YES	YES	YES	55204.700	36803.133
15	NO	YES	YES	YES	NO	19474.600	19474.600
16	Y...	YES	YES	YES	NO	22342.100	22342.100
17	NO	NO	NO	YES	NO	17729.800	8864.900
18	NO	YES	NO	YES	NO	41016.000	41016.000
19	NO	YES	NO	NO	YES	26909.200	26909.200
20	Y...	YES	YES	NO	NO	22522.800	22522.800

OK



자녀수가 0일 때는 자녀 양육비에  
지출하는 비용이 없으므로 수입을  
그대로 사용하고,  
자녀수가 1명 이상일 때는  
수입을  $(1+0.5 \times \text{자녀수})$ 로 나눈다.

# 필드 유형 및 역할 지정

**Type**

Preview

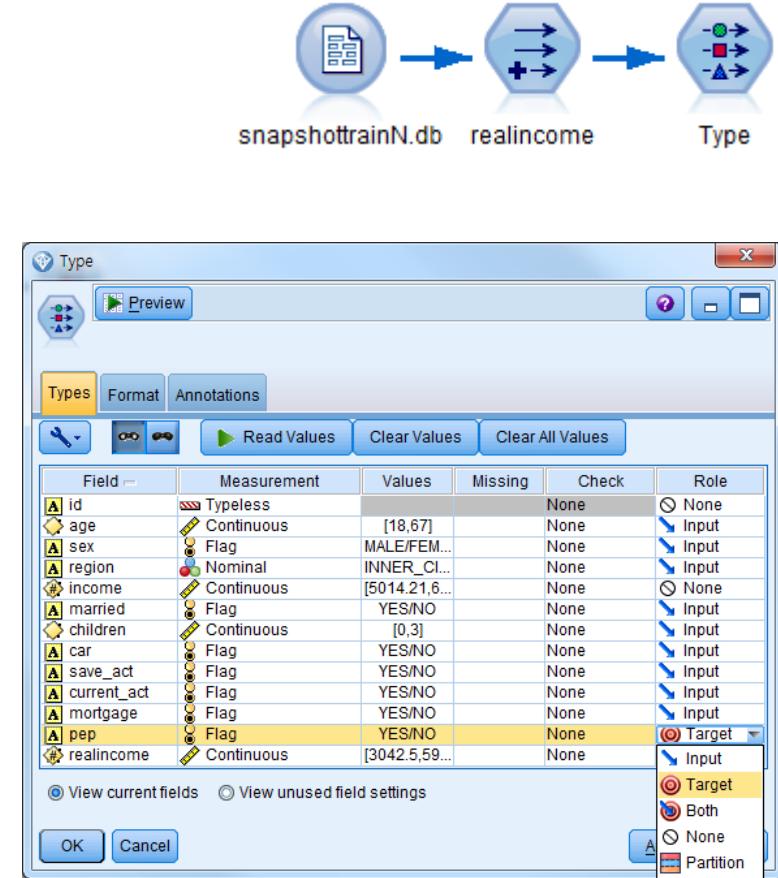
Types Format Annotations

Read Values Clear Values Clear All Values

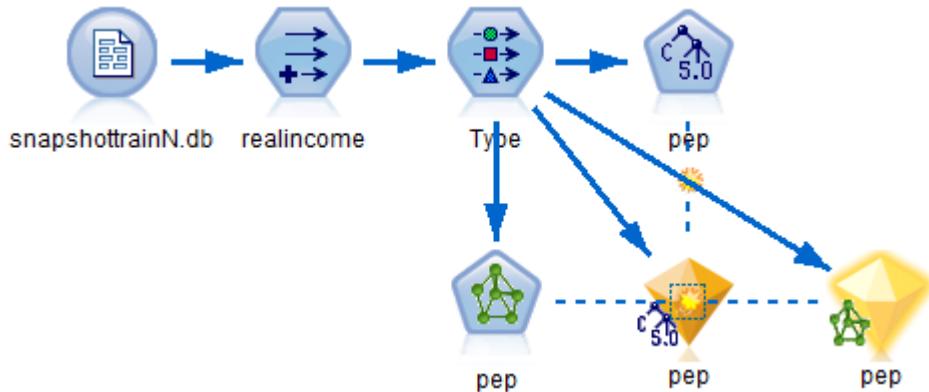
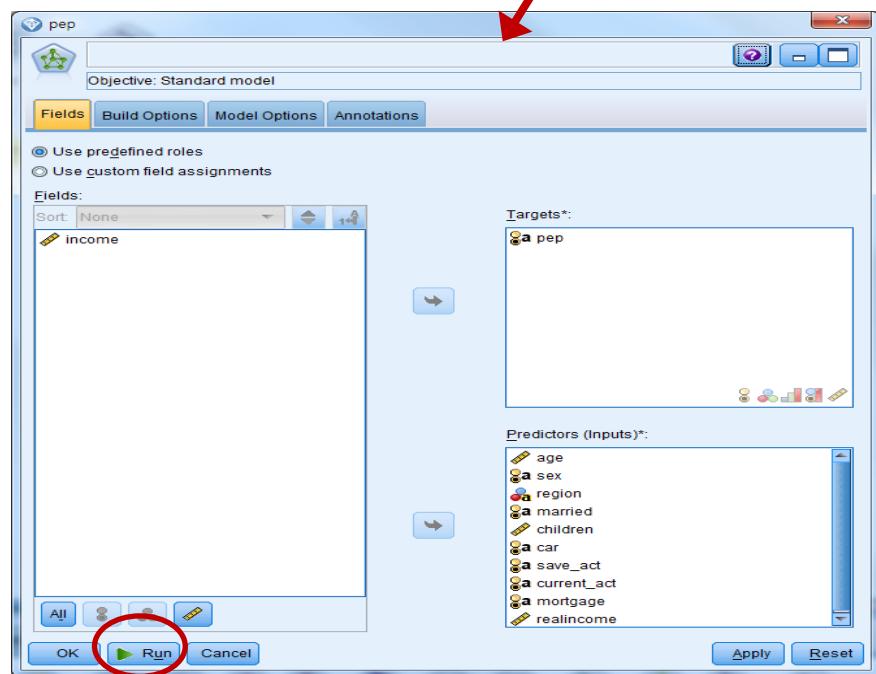
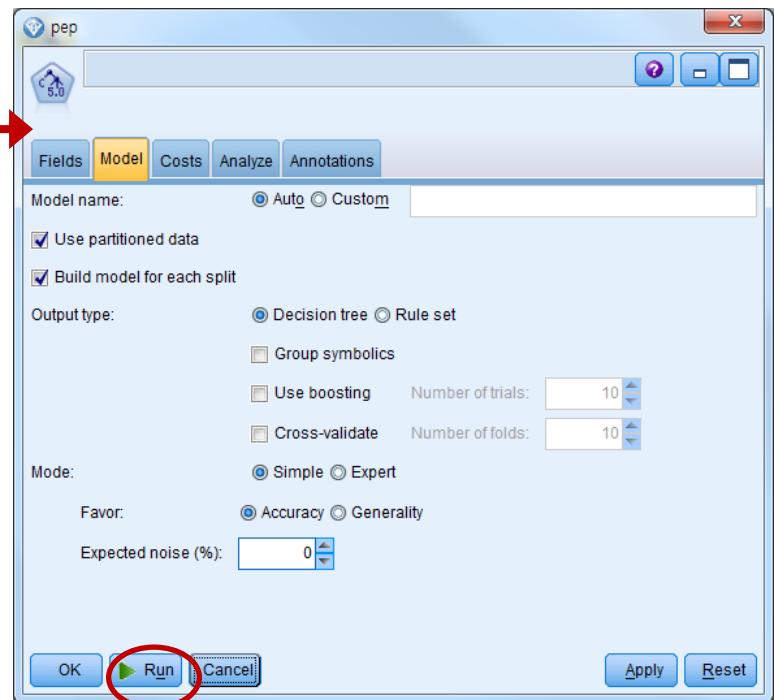
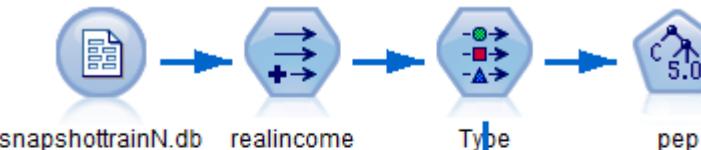
Field	Measurement	Values	Missing	Check	Role
[A] id	Typeless		None	Input	None
[A] age	Continuous	[18,67]	None	Input	Input
[A] sex	Flag	MALE/FEM...	None	Input	Input
[A] region	Nominal	INNER_C...	None	Input	Input
[A] income	Continuous	[5014.21,...	None	Input	Input
[A] married	Flag	YES/NO	None	Input	Input
[A] children	Continuous	[0,3]	None	Input	Input
[A] car	<Default>	YES/NO	None	Input	Input
[A] save_act	Continuous	YES/NO	None	Input	Input
[A] current_act	Categorical	YES/NO	None	Input	Input
[A] mortgage	Flag	YES/NO	None	Input	Input
[A] pep	Flag	YES/NO	None	Input	Target
[A] realincome	Continuous	[3042.5,59...	None	Input	Target

View current file

OK Cancel



# 모형 생성



# 모형 검토

The figure illustrates the process of model review in a data mining environment, specifically using the RapidMiner software.

**Left Panel: Model View**

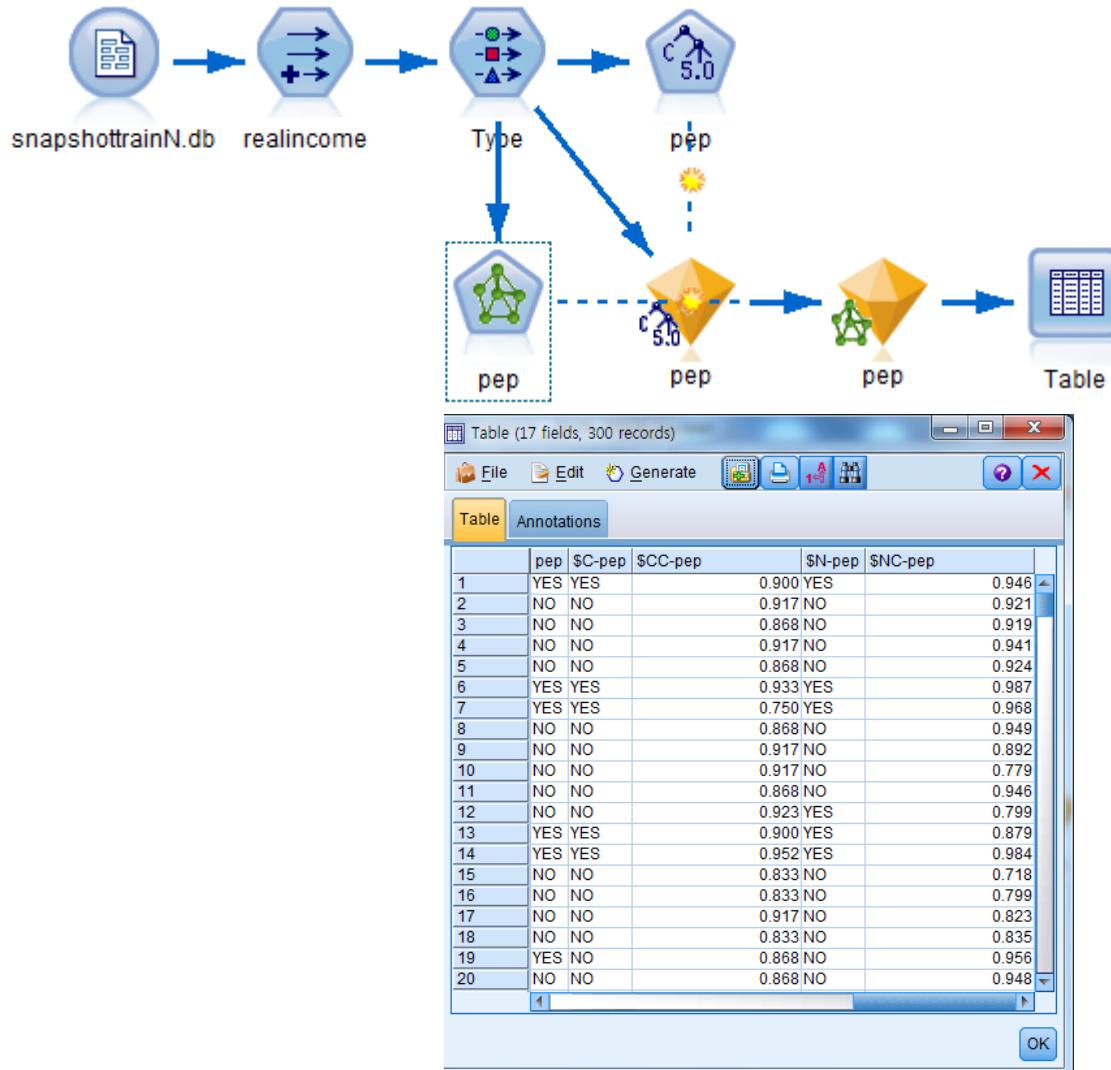
- The top navigation bar shows "Streams", "Outputs", and "Models".
- The main menu on the left includes options like "Add To Stream", "Browse", "Rename and Annotate", "Generate Modeling Node", "Save Model", "Export PMML", and "Delete".
- A red arrow points from the "Browse" option to the "Model" tab in the top navigation bar of the right panel.
- The bottom navigation bar includes "Style", "Effects", and "OK".

**Middle Panel: Model Structure**

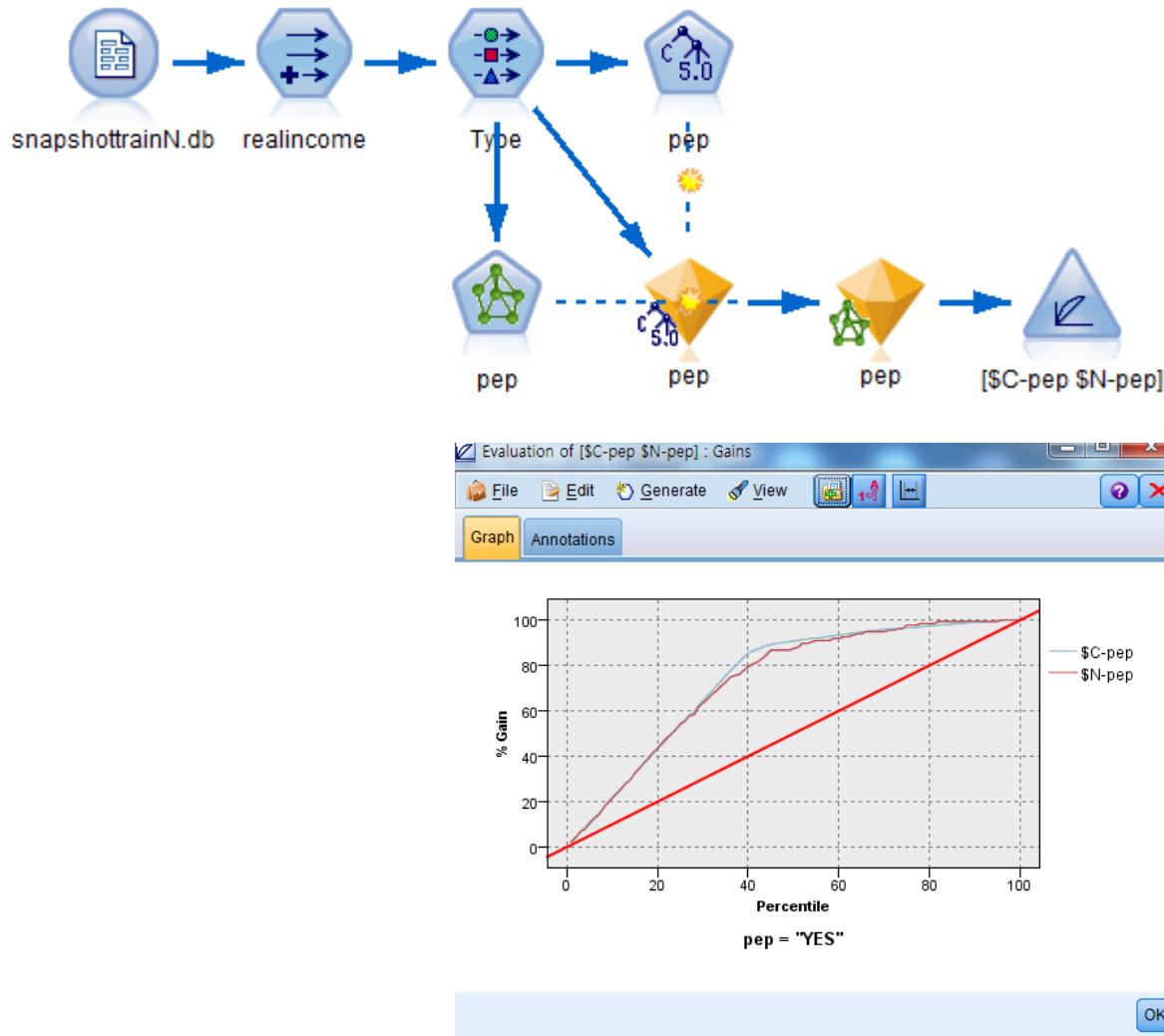
**Right Panel: Model Details and Predictor Importance**

- The top navigation bar shows "Model", "Viewer", "Summary", and "Annotations".
- The "Model" tab displays a decision tree structure for the target variable "pep". The root node is "realincome <= 13236.400 [Mode: NO] (107)". The tree branches based on "children", "age", "region", and other variables.
- The "Viewer" tab shows a "Predictor Importance" chart for the target "pep". The x-axis ranges from 0.0 (Least Important) to 1.0 (Most Important). The chart indicates that "children" is the most important predictor, followed by "region", "save\_act", "mortgage", "married", "age", and "realincome" (which is labeled as "Least Important").
- The "Summary" and "Annotations" tabs are also visible in the top navigation bar.

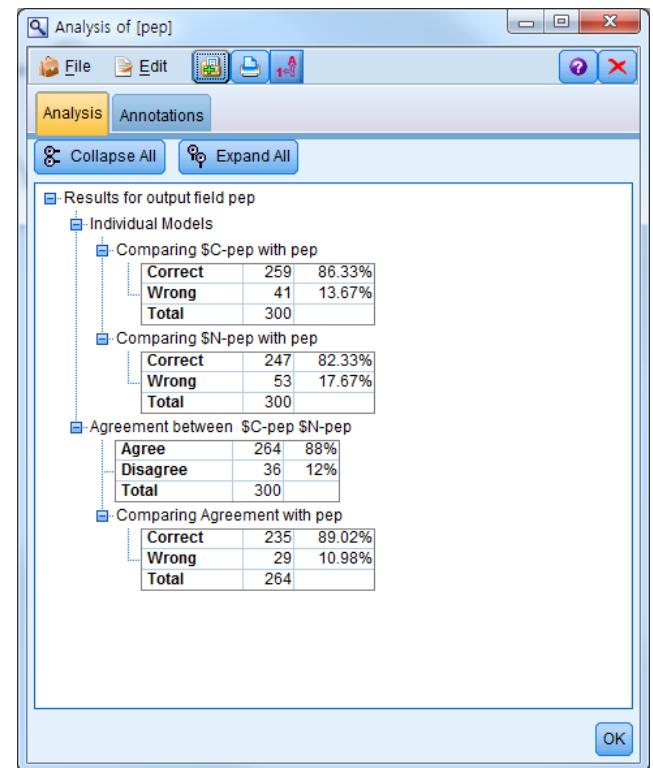
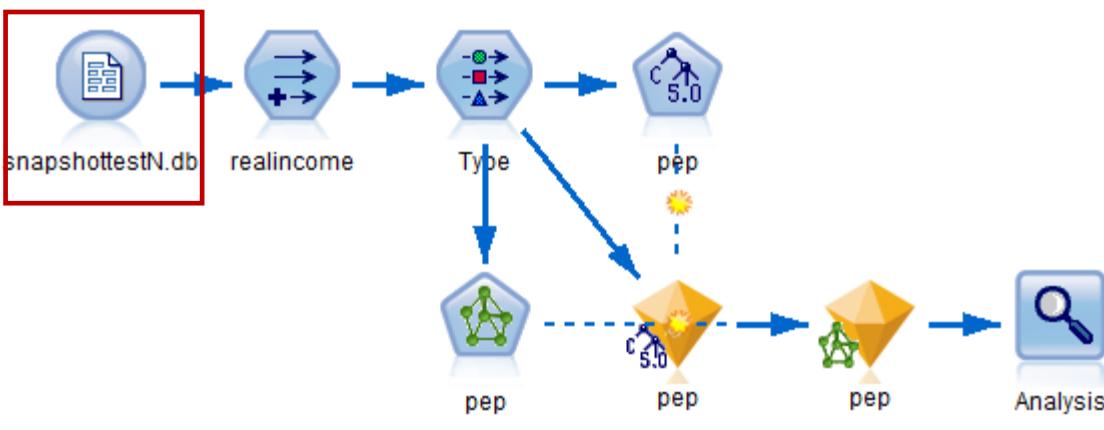
# 모형 평가



# 모형 평가 (Cont.)



# 검증용 데이터에 의한 모형 평가



# 모형 전개(Model Deployment)

