Branding Method

의사결정나무 분석기법
-그 의미와 적용 사례-

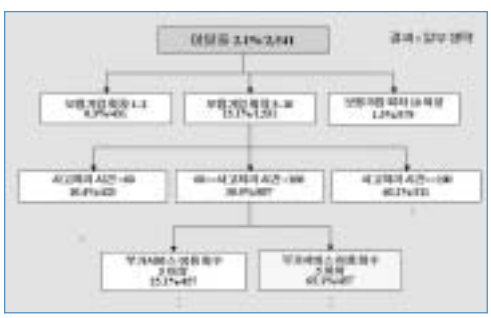
SPSS Korea
정성원

1. 의사결정나무 분석기법의 개념

의사결정나무 (Decision Tree)는 의사결정 규칙 (Decision Tree)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류 (Classification) 하거나 예측 (Prediction)을 수행하는 계량적 분석 방법이다.

분석결과는 '조건 A이고 조건 B이면 결과집단 C' 라는 형태의 규칙으로 표현되므로 이해가 쉽고, 분류 또는 예측을 목적으로 하는 다른 계량적 분석 방법에 비해 쉽게 이해하고 활용 할 수 있다는 장점이 있다.

〈그림-1 이탈고객 스코어링 모형〉



2. 의사결정나무 분석기법의 활용분야

의사결정나무 분석기법은 DB마케팅, CRM, 시장조사, 광고조사, 의학연구, 품질관리 등의 다양한 분야에서 활용되고 있으며, 구체적인 활용 예는 DM의 응답자 분석, 고객 타겟팅, 고객들의 신용점수화, 캠페인 반응분석, 고객행동예측, 고객세분화, 시장 세분화, 신상품 수용도 분석, 광고효과측정, 상표 이미지 테스트 등을 들 수 있다.

3. 의사결정나무의 구조

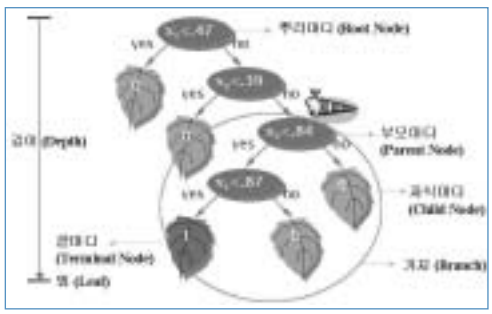
의사결정나무의 맨 위쪽에 위치하는 마디를 가리켜서 뿌리마디 (Root Node)라고 부르는데, 분류대상이 되는 모든 개체집단을 의미하게 된다. 하나의 마디가 하부마디로 분화가 될 때, 특정마디 위쪽에 존재하는 마디를 부모마디 (Parent Node)라고 부르고 특정마디 아래쪽에 존재하는 마디를 자식마디 (Child Node)라 부르며 더 이상 마디가 분화되지 않는 최종마디를 끝마디

정성원
jung@spss.co.kr

SPSS Korea / 이사
소속학회 : 통계학회, 데이터마이닝학회, 한국데이터베이스마케팅협회, 한국여론조사 협회 등
저서 : 윈도우용 SPSS 통계조사분석, 범주형데이터분석, 전문통계분석, SPSS/PC+ I II, SPSS/PC+ 데이터분석 외
주요프로젝트 : 1998 LG유통, 삼성SDS 1999 한국산업보건연구원 1차, 현대증권, LG화재 2000 한국산업보건연구원 2차, 교보생명 2001 국민카드, 롯데제과 2002 한국산업안전공단, 한국통신 2003 CJ 홈쇼핑

(Terminal Node)라고 부른다. 이와 같은 각 마디들이 분화되어있는 모습이 나무의 모양을 닮았다고 하여 이를 의사결정 나무라 부른다.

〈그림-2 의사결정나무의 구조〉

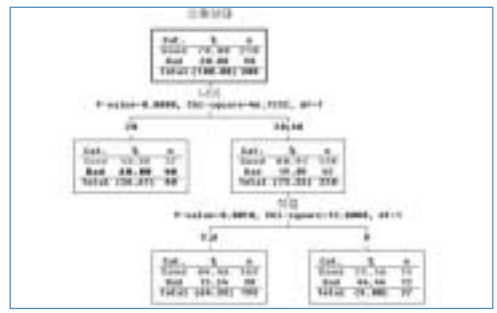


4. 의사결정나무의 종류 및 분리과정

의사결정나무는 명목형 목표변수를 기준으로 마디가 분화되는 분류나무 (Classification Tree)와 연속형 목표변수를 기준으로 마디가 분화되는 회귀나무 (Regression Tree)로 나눌 수 있다.

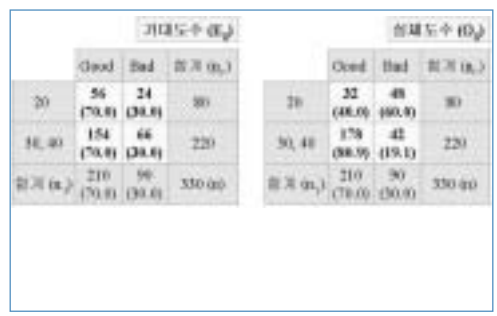
먼저 분류나무(Classification)의 경우에는 목표 변수의 각 범주에 속하는 빈도(frequency)에 기초하여 분리가 일어난다. 예를 들어, 신용상태가 'Good' 과 'Bad' 의 값을 갖는 명목형 목표변수를 기준으로 나이 (20대, 30대, 40대), 직업 (A, B, C)을 고려하여 전체고객 300명을 세분화하는 과정을 설명하면, 첫번째 단계에서 목표변수와 나이, 직업 등의 각각의 변수간의 상관도가 가장 밀접한 변수를 선택하고, 선택된 변수의 범주를 다양하게 조합하여 가장 상관도가 높은 범주 조합을 선택하여 첫번째 분리를 한다. 직업보다 나이가 신용상태 (목표변수)와 상관도가 높다면 나이를 우선 선택하고 나이의 모든 범주조합 (20, 30 40), (20 30,40), (30, 20 40) 중에 가장 상관도가 높은 조합을 선택한다. 이때 선택하는 기준은 의사결정나무 알고리즘에 따라 카이제곱 통계량 (Chi-Square Statistics), 지니 지수 (Gini Index), 엔트로피 지수 (Entropy Index)등을 사용 한다.

〈그림-3 분류나무 예시〉



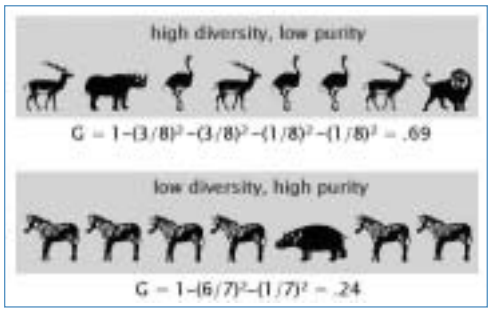
일단 첫번째 분리에서 나이를 기준으로 (20, 30 40)로 분리하는 것이 가장 신용상태와 상관도가 높게 되는 범주조합이라고 판단되면, 이러한 기준에 의해 분리를 시키고 각각의 분리가지에 대해 다음 분리를 위해 변수를 선택하고 같은 요령으로 범주조합을 선택한다.

〈그림-4 카이제곱 통계량 계산〉



CHAID (Chi-squared Automatic Interaction Detection) 알고리즘의 경우 카이제곱 통계량에 의해 목표변수를 분리하는 정도를 결정하는데 세부적인 카이제곱 통계량의 계산과정은 〈그림-4〉에서 보는 바와 같다.

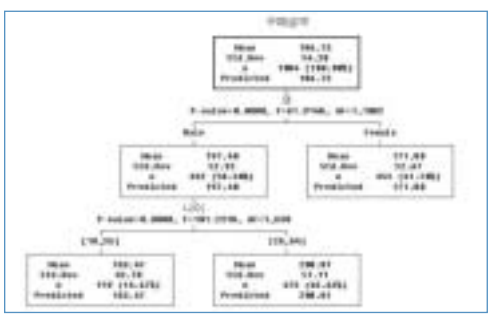
〈그림-5 지니지수 계산〉



CART (Classification & Regression Tree) 알고리즘의 경우 지니지수에 의해 목표변수를 분리하는데 세부적인 지니지수 계산과정은 〈그림-5〉와 같다.

회귀나무 (Regression Tree)의 경우에는 각 분리마디별 목표변수의 F-통계량 또는 분산의 감소량에 의해 마디를 분리한다. 목표변수의 F-통계량에 의해 분리마디를 결정하는 경우에는 목표변수와 설명변수들 간에 가장 설명력이 큰 변수를 먼저 선택하여 분리를 한다.

〈그림-6 회귀나무 예시〉

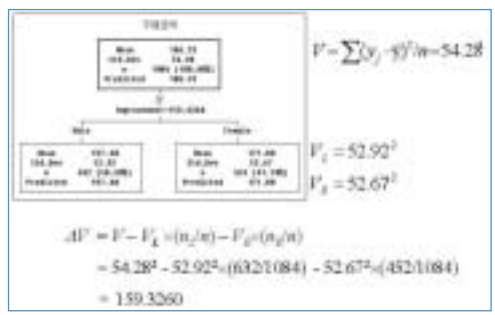


F-통계량을 계산하는 절차는 일반적으로 많이 알려진 통계분석법인 분산분석(ANOVA)과 동일한 방법이다.

〈그림-7 F-통계량 계산〉

특정한 마디를 분리함으로써 분산이 줄어든다는 것은 동질적 성격을 가지고 있는 개체들은 같은 집단으로 이질적 성격을 가지고 있는 개체들은 다른 집단으로 나뉘었다는 것을 의미하므로, 가장 분산을 크게 감소시키는 분리조합이 최선의 분리를 이끌어 내는 것이다.

〈그림-8 분산의 감소량 계산〉



5. 의사결정나무의 분석과정

의사결정나무의 분석과정을 요약하면 다음과

같다.

1. 목표변수와 관계가 있는 설명변수들의 선택
2. 분석목적과 자료의 구조에 따라 적절한 분리 기준과 정지규칙을 정하여 의사결정 나무의 구조 작성
3. 부적절한 나뭇가지는 제거 (가지치기)
4. 이익(Gain), 위험(Risk), 비용(Cost) 등을 고려하여 모형평가
5. 분류(Classification) 및 예측(Prediction)

의사결정나무의 정지규칙이란 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러가지 규칙을 의미한다. 이러한 규칙에는 최대 나무의 깊이, 자식마디의 최소 관측치 수, 또는 카이제곱 검정통계량, 지니지수, 엔트로피 지수 등이 될 수 있다.

의사결정나무 가지치기란 끝마디가 너무 많으면 모형이 과대 적합된 상태로 현실문제에 적용할 수 있는 적절한 규칙이 나오지 않게 된다. 따라서 분류된 관측치의 비율 또는 MSE (Mean Squared Error)등을 고려하여 적절한 수준의 가지치기 규칙을 제공하여야 한다.

6. 의사결정나무분석 알고리즘

현재까지 연구된 의사결정나무분석 알고리즘은 4가지 정도의 종류가 있다. 가장 널리 사용되는 알고리즘으로 1975년 J.A.Hartigan에 의해 개발된 CHAID 알고리즘은 명목형, 순서형, 연속형 등 모든 종류의 목표변수와 분류변수에 적용이 가능하다. CHAID 알고리즘은 1991년 Biggs et al. 에 의해 Exhaustive CHAID 알고리즘으로 발전 하게 되었다. 하나의 부모마디 밑에 2개의 자식마디만이 생기는 이지(Binary) 분리 알고리즘인 CART 알고리즘은 1984년 Leo Briemans & Associates 에 의해 개발되었고 CHAID와 마찬 가지로 목표변수나 분류변수의 척도에 관계없이 적용할 수 있다는 장점으로 인해 널리 사용되고 있다. 가장 최근에 만들어진 C5.0알고리즘은 호주의 수학자인 J. Ross Quinlan 박사에 의해 1986년 ID3라는 이름의 알고리즘으로 만들어 졌다가 1993년에 C4.5를 거쳐 1998년에 완성된 알고리즘으로 명목형 목표변수만을 지원하는 단점이 있는 반면에 가장 정확한 분류를 만들어 주는 알고리즘으로 평가 받아 최근 데이터 마이닝 분야에서도 폭넓게 사용되고 있다.

7. 적용사례

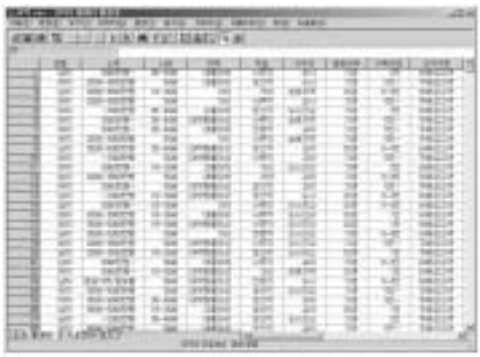
●펌프식 용기의 치약

미국 Market Studies사는 대규모 소비용품업체가 의뢰한 2단계의 치약시장 조사를 통해 소비자 1000명을 대상으로 치약에 관한 태도, 의견 및 행동을 파악하고, 기존의 로션 제품에서 사용하는 것과 유사한 "펌프식 용기"로 포장한 치약제품에 대한 관심을 측정 할 목적으로 면접조사를 실시 하였다.

만약 펌프식 용기의 치약이 펌프식 용기의 액체비누만큼 인기가 있는 것으로 판명된다면, 이것은 전체치약시장에서 하나의 중요한 세분시장이 될 것으로 판단되었다.

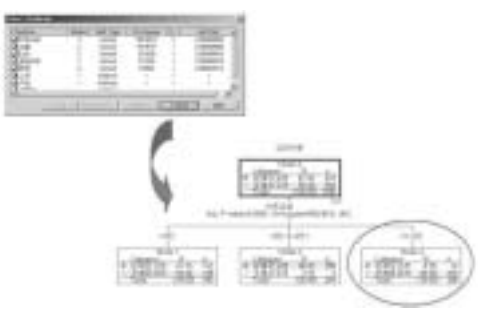
펌프식 치약제품에 대한 잠재시장을 구성하는 소비자들의 특성은 어떠한지를 파악하는 조사를 추가로 실시하였다.

〈그림-9 치약사례 데이터〉



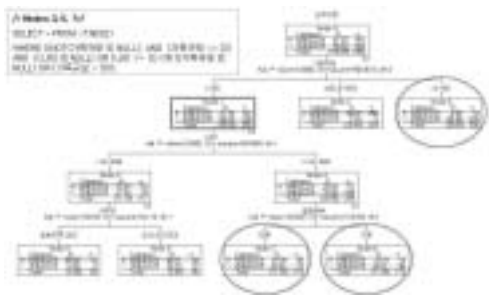
응답자 구분방법은 펌프식 용기의 액체비누를 알고 사용경험이 있으며, 펌프식 용기의 치약 아이디어를 "매우 좋아 하거나" "약간 좋아 하며", 이러한 제품을 "구입한다"거나 "구입가능성이 높은" 응답자는 주예상고객(prime prospect)으로 보았고, 펌프식 용기의 액체비누를 사용해 본 경험이 없으나 이러한 형태의 제품을 알고 있고, 펌프식 용기의 치약 아이디어를 "매우 좋아하거나" "약간 좋아하며", 이러한 제품을 "구입한다"거나 "구입가능성이 높은" 응답자는 좋은 예상고객(good prospect)으로 보았고 펌프식 용기의 액체비누를 알고 있고 사용해 본 경험이 있으며, 펌프식 용기의 치약 아이디어를 "매우 좋아하거나" "약간 좋아 하며", 이러한 제품을 "구입가능성이 있거나" "미정인" 응답자는 평균적인 예상고객(fair prospect)로 판단하였고, 나머지 모든 응답자는 비예상고객 (nonprospect)으로 보았다.

〈그림-10 치약사례 1차마디 생성〉



의사결정나무 분석에 사용한 고객구분은 주예상고객과 좋은예상고객을 주예상고객으로, 평균적인 예상고객과 비예상고객을 비예상고객으로 이분구분하였다.

〈그림-11 치약사례 나머지 마디 생성〉



〈그림-12 치약사례 모형평가〉

●가정용품 구매 정보원

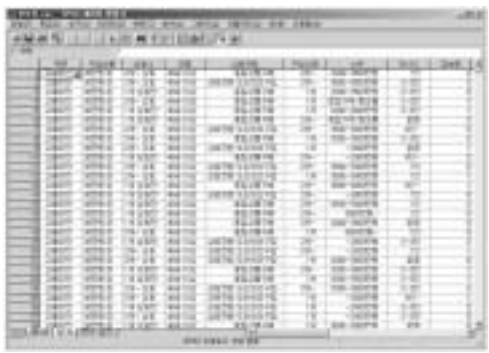
미국 가정용품회사의 마케팅관리자는 구매자들이 구매결정을 할 때 몇 개의 다른 정보원을 활용하는지 조사하였다.

활용하는 정보원의 수에 차이가 있는 구매자 집단간에 어떤 차별 특성이 있는지 분석해 보면 보다 효과적인 광고 및 판촉 캠페인을 설계하는데 도움이 될 것으로 판단되었다.

표본조사를 통해 최근에 가정용품을 구입한

1000가구를 대상으로 1)친구와 이웃 사람들, 2)책과 잡지, 3)매체광고, 4)광고 팜플렛과 소책자, 5)소매점의 점원, 6)인터넷 등의 정보원 활용 여부, 구입 고려한 가정용품의 수, 가정용품 구입비, 고려했던 상표의 수, 과거와 동일상표 구입여부, 가장의 연령과 학력, 가족소득, 자녀의 수 등을 포함하는 25개 변수에 관한 정보를 수집하였다.

〈그림-13 가정용품 사례 데이터〉



응답자의 특성을 나타내는 변수의 수가 25개나 되므로 교차집계로 중요변수를 파악하기는 힘들다.

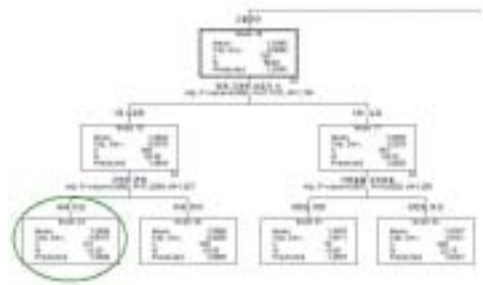
의사결정 나무분석의 CHAID 알고리즘에 의해 모든 2원 분할중에 가장의 학력이 가장 설명력이 높은 변수로 선정되었다.

정보원의 대량사용자는 가장의 학력이 고졸이상이라는 증거를 갖는 셈이 된다.

〈그림-14 가정용품 사례 1차마디 생성〉

CHAID 알고리즘은 전체표본을 하위표본의 종속변수 평균간에 통계적으로 유의미한 차이가 되도록 2개의 하위표본으로 분할하는 변수만을 선정 한다.

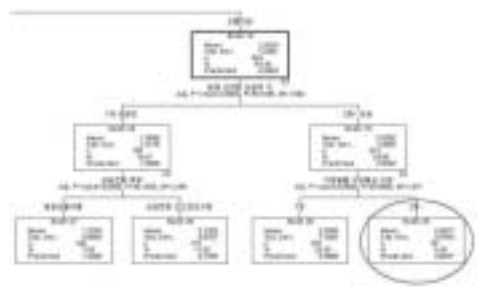
〈그림-15 가정용품 사례 추가마디 생성〉



CHAID 알고리즘은 2개 이상의 독립변수 조합이 종속변수에 미치는 효과를 잘 이해할 수 있게 해준다.

가장 정보원을 많이 활용하는 집단은 고졸이상의 가장, 2개 이상의 상표를 고려하고, 가정용품 구입예상 수량이 2개 이상인 가구집단이다

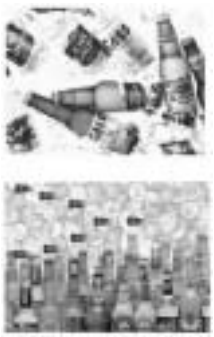
〈그림-16 가정용품 사례 추가마디 생성〉



●맥주상표 인지도 조사

미국 Brewer사는 아리조나, 뉴멕시코, 텍사스주 지역에서 맥주류 제품을 판매하고 있다.

1000명의 남성 맥주 음주자를 대상으로 광고 카피 전략에 활용할 수 있는 상표인지도 조사를 실시하였다.



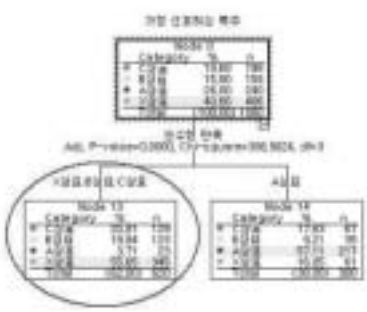
19가지 광고표어를 응답자에게 보여주고 각 광고표어별로 먼저 생각나는 상표가 무엇인지를 묻고, 응답자의 맥주 음주량 수준, 가장 선호하는 맥주상표를 물었다.

〈그림-17 맥주상표 사례 데이터〉

[19가지 광고표어]

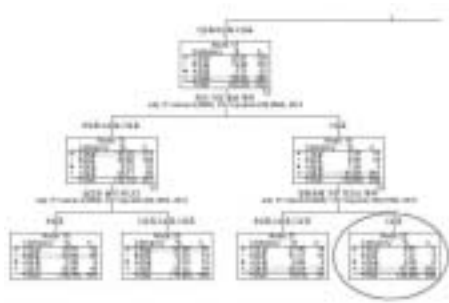
- 가장 많이 판매되는 맥주
- 가장 오래된 전통맥주
- 매력적인 광고의 맥주
- 급조된 술이 아니다.
- 대중이 좋아하는 술
- 맛이 가장 좋은 맥주
- 가장 오래된 맥주회사
- 숙녀용 맥주
- 스포츠맨용 맥주
- 알코올 도수가 낮은 맥주
- 야외활동시 가장 맛있는 맥주
- 사교모임시 가장 맛있는 맥주
- 남자에 가장 잘 어울리는 맥주
- 가장 풍미가 도는 맥주
- 가장 숙성된 맛이 나는 맥주
- 젊은이용 맥주

〈그림-18 선호맥주에 따른 1차 마디 생성〉



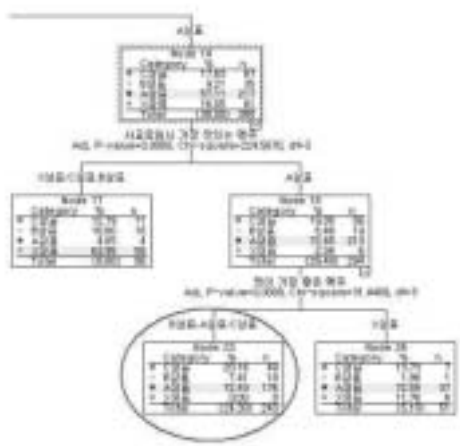
"순수한 만족"이란 광고 카피로 인지하는 맥주상표의 종류에 따라 응답자들의 맥주상표 선호도 차이가 나는 것으로 나타났다.

〈그림-19 선호맥주에 따른 2차 마디 생성〉



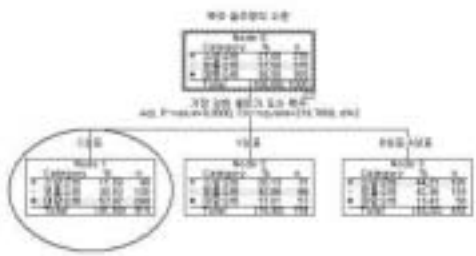
Brewer사의 X상표의 맥주를 가장 선호하는 사람들은 X상표가 "순수한 만족", "맛이 가장 좋은 맥주", "일한 후에 가장 맛있는 맥주"로 인지하고 있다.

〈그림-20 선호맥주에 따른 2차 마디 생성〉



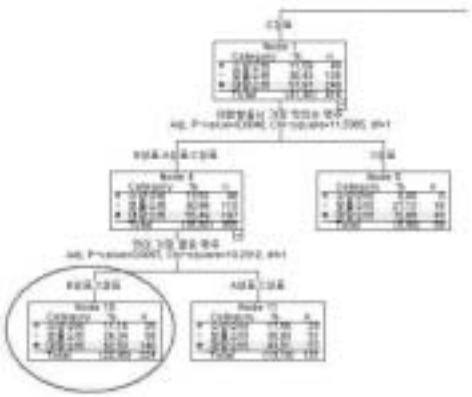
Brewer사의 X상표 선호도가 가장 떨어지는 응답자 집단은 A상표를 "순수한 만족", "사교 모임시 가장 맛있는 맥주", "맛이 가장 좋은 맥주"로 인지하고 있다.

〈그림-21 음주량에 따른 1차 마디 생성〉



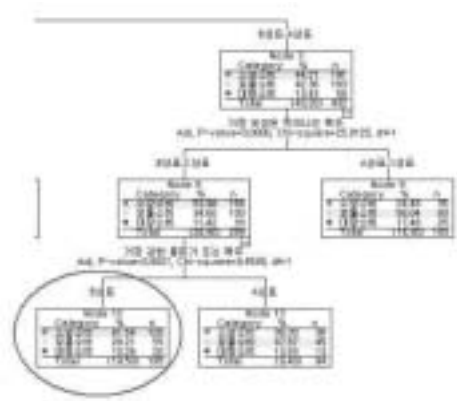
맥주 음주량의 수준별로 상표인지도를 살펴 보았을 때 대량소비 응답자집단은 C상표를 "가장 풍미가 도는 맥주"로 인식하는 집단이다.

〈그림-22 음주량에 따른 2차 마디 생성〉



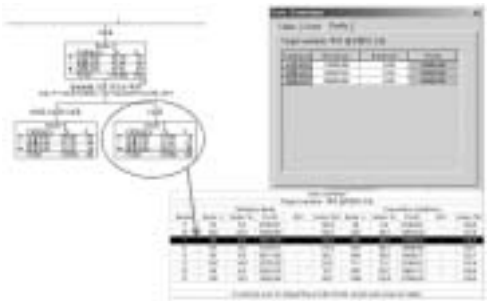
가장 대량 소비를 하는 응답자 집단은 C상표를 "가장 풍미가 도는 맥주"로 인식하고 경쟁상표(B상표,A상표,C상표)를 "야외 활동 시 가장 맛있는 맥주"로 인식하고, B상표, X상표를 "맛이 가장 좋은 맥주"로 인식하는 집단이다.

〈그림-23 음주량에 따른 2차 마디 생성〉



가장 소비를 적게 하는 응답자 집단은 경쟁사 상표(A상표, B상표, C상표)를 "가장 풍미가 도는 맥주", "가장 숙성된 맛이 나는 맥주"로 인식하는 집단이다.

〈그림-24 음주량에 따른 2차 마디 생성〉



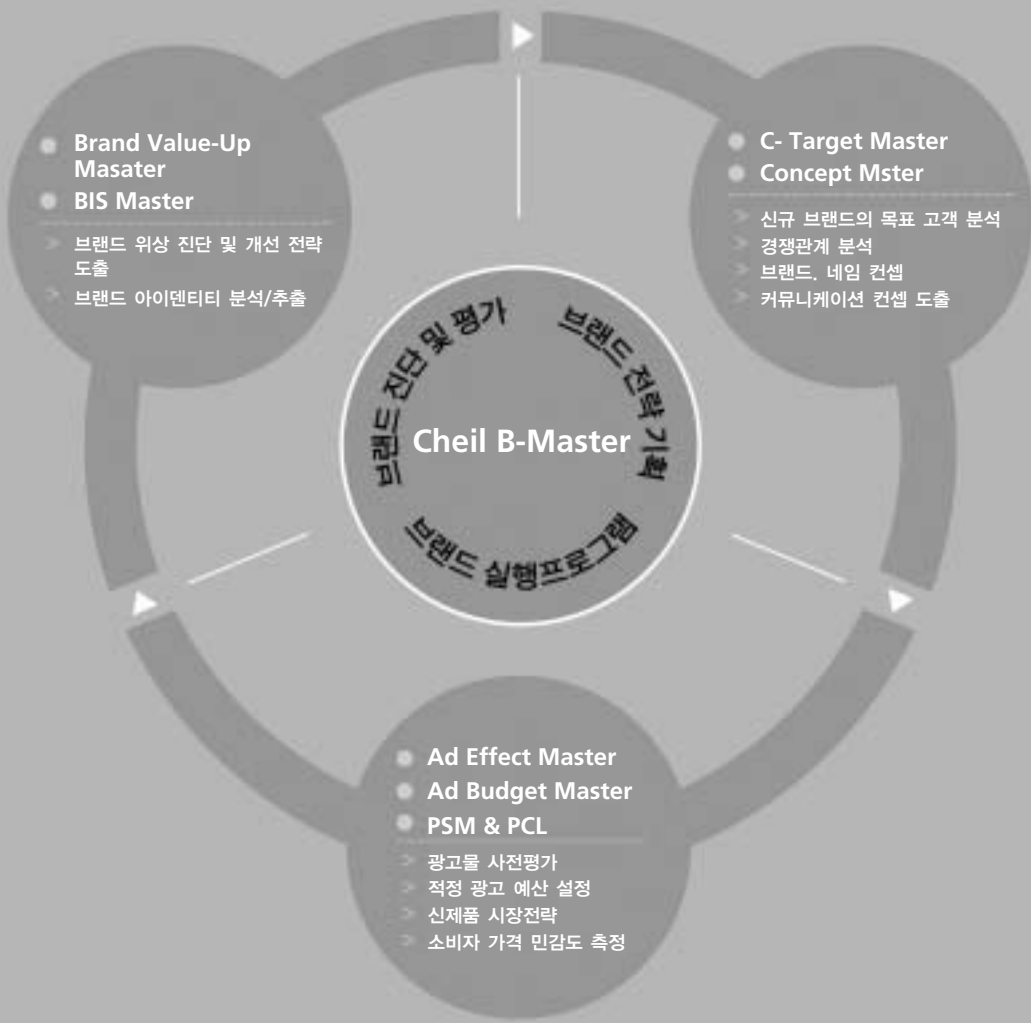
1인당 한달 평균 맥주 구입비용이 소량소비집단 1만원, 보통소비집단 2만원, 대량소비집단 5만원으로 가정 하였을 때 "가장 풍미가 도는 맥주", "일한 후 가장 맛있는 맥주"로 X상표를 인지하는 집단의 1인당 추정 맥주구입 비용은 3.5만원인 것으로 나타났고 이는 전체 집단 평균의 1.26배에 해당하는 것으로 나타났다.

1인당 추정 맥주구입 비용이 4.18만원으로 가장 큰 집단은 C상표를 "가장 풍미가 도는 맥주"로 인식하고 X상표를 "야외 활동 시 가장 맛있는 맥주"로 인식하는 집단이다.

8. 결론

의사나무기법은 계량적인 데이터에 근거한 분석기법으로서 다양한 분야에 활용되는 기법으로 의사나무기법을 지원하는 소프트웨어를 사용하면 누구나 간단히 분석하고 결과를 해석하기가 쉬운 기법이다. 특히 마케팅 조사나 광고 조사 분야에서 폭 넓게 사용될 수 있는 기법으로 좀 더 많은 사용자가 생겼으면 하는 희망과 함께 글을 마친다.

Cheil B-Master Framework



단 계	모 델	설 명
브랜드 진단 및 평가	● Brand Value-Up Master	▶ 브랜드 위상 진단 및 개선 전략 도출 모델
	● BIS Master	▶ 브랜드 원칙과 체계를 구축하기 위한 모델
브랜드 전략 기획	● C-Target Master	▶ 브랜드 목표 고객에 대한 분석 모델
	● Concept Master	▶ 소비자의 내면적 속성과심층심리를 파악하여 최적의 브랜드 컨셉 도출 모델
브랜드 실행 프로그램	● Ad Effect Master	▶ 광고물을 사전, 사후 평가하여 광고의 개선점, 개선방향을 파악하고 지속적으로 광고효과를 관리하기 위한 모델
	● Ad Budget Master	▶ 브랜드 커뮤니케이션을 위한 최적의 광고 예산 설정 모델
	● PSM[Price Sensitivity Measurement]	▶ 소비자에게 가장 저항감이 없는 가격을 찾아내는 방법으로 전략적 활용이 높은 가격 설정 모델
	● PCL[Product Clinic Laboratory]	▶ 소비자가 길제 구매 과정에서 거치는 광고, 인상, 접촉, 시용의 각 단계별로 신제품을 평가하고 출시후의 시장전망과 대응전략 도출이 가능한 신제품 마케팅 모델