The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 217-227

Data Mining Model Approach for The Risk Factor of BMI - By Medical Examination of Health Data -

Jea-Young Lee[1] and Yong-Won Lee[2]

Abstract

The data mining is a new approach to extract useful information through effective analysis of huge data in numerous fields. We utilized this data mining technique to analyze medical record of 35,671 people. Whole data were assorted by BMI score and divided into two groups. We tried to find out BMI risk factor from overweight group by analyzing the raw data with data mining approach. The result extracted by C5.0 decision tree method showed that important risk factors for BMI score are triglyceride, gender, age and HDL cholesterol. Odds ratio of major risk factors were calculated to show individual effect of each factors.

Keyword : Data Mining, BMI, HDL cholesterol

1. 서론

컴퓨터와 네트워크의 엄청난 발전으로 데이터베이스를 만들어 많은 양의 자료를 저장, 보관하게 되었다. 여러 분야에서 저장, 보관되는 대량의 데이터를 효과적으로 분석하여 유용한 정보를 획득하기 위해 새로이 등장한 것이 데이터 마이닝이다. 데이터 마이닝이란 대량의 데이터나 복잡한 구조의 데이터들을 정교한 통계분석과 모델링(Modeling) 테크닉을 이용하여 정확히 식별되지 않는 패턴이나 자료간의 상관관계를 밝혀내어 여러 가지 결과를 예측해 내는 새로운 통계적 기법이다. 이러한 데이터 마이닝 기법을 이용하여 제조업의 부도예측에 영향을 주는 주요변수를 선택하거나(최병권, 2004), 웹 페이지 분석을 통한 성별과 관심분야에 대한 웹 페이지 성향 파악이나(Baglion 등, 2003), 두 개의 서로 다른 분포에서 나온 데이터가 섞여있을 때 데이터 마이닝의 신경망 기법을 이용하여 올바른 판별을 할 수 있게 되었다(이성원, 2001). 기술, 경제적인 분야 등 다양한 분야에 활용되고 있는 데이터 마이닝을 의학 분야에 이용하여 고혈압 위험요인들의 정확한 평가를 위하여 성별에 따라 로지스틱 분석을 적용하고, 로지스틱 분석을 통하여 위험요인들의 상대 위험도(Relative Risk)를 구하였다(오희숙 등, 2000). 또한, 한국인에게 비만이 고혈압 발생에 독립적인 위험요인을 규명하였다 (이성희, 2001). 의학에서 로지스틱 모형을 사용하여 자료를 분석하였지만,

1) Professor, Department of Mathematics and Statistics, Yeungnam University, 214-1 Daedong Kyungsan Kyungbuk, 712-749, South Korea
E-mail : jlee@yu.ac.kr
2) Research fellow, Institute of Medical Science, Yeungnam University College of Medicine, 317-1 Daemyeung-dong Namgu Daegu, 705-717, South Korea

본 연구는 대규모 데이터를 바탕으로 데이터 마이닝 분석을 실시하기로 하였다.

2. 연구 내용 및 대상

최근 소득수준의 향상, 과다한 영양섭취 및 영양불균형이 심화됨에 따라 비만유병인구가 증가하는 양상을 보이고 있다. 비만은 체내에 지방조직이 과다하게 축적된 상태를 나타내는 것으로 유전적, 사회 환경적, 경제적인 요소 등 많은 요소들이 복합적으로 연관되어 있는 것으로 알려져 있다. 1990년대 이후 비만이 중요한 건강문제로 대두되면서 국내에서도 비만에 관한 많은 연구가 진행되고 있으나, 대부분 소규모 산발적으로 진행되고 있어 국가차원의 대책 마련을 위한 자료로 사용하기에 불충분한 경우가 많았다. 동시에 건강에 대한 인식이 점점 높아짐에 따라 체중조절을 하는 사람들이 많아지며, 본인의 체형에 대한 부정확한 인식으로 인한 검증되지 않은 무분별한 다이어트가 사회문제가 되고 있다. 또한, 비만은 만성퇴행성질환의 발병과 밀접한 관련이 있으며, 특히 당뇨병, 고혈압, 관상 동맥성 심장질환, 암, 중풍과 관련이 높은 것으로 보고되고 있고(WHO, 1997), 개인적인 행동양식 및 정신적인 면까지 영향을 미치는 주요한 건강저해 요소이다.

하지만 비만에 관한 연구내용은 비만의 유병률이나 건강에 미치는 영향에 관한 연구가 많았고 비만의 위험인자나 예방에 관한 연구는 적음을 알 수 있다.(탁양주 등, 2004)

이러한 비만유병에 대한 주요인을 파악하기 위하여 2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 종합건강검진센터에 건강검진을 받은 20세 이상의 35,671명을 대상으로 분석을 실시하였다. 대상자 전원에서 흡연력, 음주력, 운동습관, 결혼상태, 성별, 직업을 면접조사 하고, 신체질량지수(BMI, Body Mass Index)수치, 총 콜레스테롤(Total cholesterol), HDL 콜레스테롤(High-density Lipoprotein Cholesterol), (Low-density Lipoprotein Cholesterol) LDL 콜레스테롤을 측정하였다.

본 논문에서 종속변수가 되는 BMI수치를 근거로 하여 비만 여부를 결정하였다. BMI는 다음과 같이 산출되어지며, BMI에 대한 분류(WHO, 2000)는 <표 1>과 같다.

$$BMI(Body\ Mass\ Index) = Weight(kg)/[Height(m)]^2$$

<표 1> BMI 분류기준

분류	BMI(kg/㎡)
저체중	<18.5
정상	18.5~22.9
위험체중	23~24.9
1단계 비만	25~29.9
2단계 비만	>30

WHO에서 서양인은 BMI 30kg/m^{2}이상을 비만으로 정의하며, 동양인은 BMI 25kg/m^{2}이상을 비만으로 정의한다(WHO, 2000). 본 논문에서는 WHO에서 정의된 동양인 기준 BMI지수 25를 경계로 하여 BMI지수 25이상을 비만집단, BMI지수 25미만을 비만이 아닌 집단으로 종속변수를 이

분류 하였다.

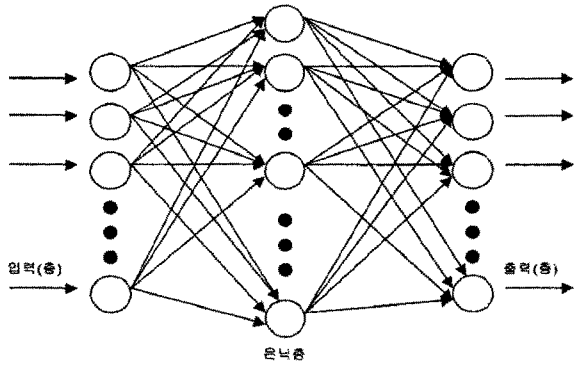
그리고, 독립변수가 되는 흡연력, 음주력, 운동습관, 결혼상태, 성별, 직업, 나이, 총 콜레스테롤, HDL 콜레스테롤, LDL 콜레스테롤에 관한 내용은 다음과 같다.

흡연력에 대하여 현재 흡연자, 과거 흡연자, 미 흡연자로 분류하였으며, 음주력에 대하여 즐겨 마시는 그룹, 거의 마시지 않지만 어쩌다 1~2잔정도 마시는 그룹, 전혀 음주하지 않는 그룹으로 분류하였다. 운동습관에 대하여 운동을 전혀 하지 않는 그룹과 운동을 하는 그룹으로 분류하였고, 나이는 10대 단위로 범주화 하여 20~29세를 20대, 30~39세를 30대, 40~49세를 40대, 50~59세를 50대, 60~69세를 60대 그리고 70대 이상으로 분류하였다. 총 콜레스테롤은 200미만을 정상그룹, 200~239를 경계그룹, 240이상을 위험그룹으로 분류하고, HDL 콜레스테롤은 46이상을 정상그룹, 35~45를 경계그룹, 35미만을 위험그룹으로 분류하였다. LDL 콜레스테롤은 130미만을 정상그룹, 130~159를 경계그룹, 160이상을 위험그룹으로 분류하였으며, 중성지방은 150미만을 정상그룹, 150~200을 경계그룹, 201이상을 위험그룹으로 분류하였다.

3. 데이터 마이닝 기법의 배경

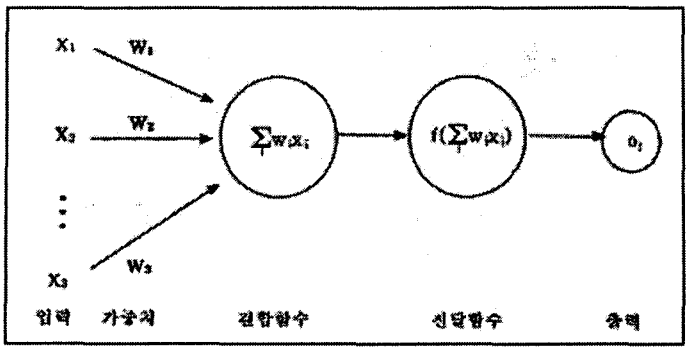
BMI의 주요인을 분석하기 위하여 데이터 마이닝 기법을 사용한다. 데이터 마이닝은 보통 탐색, 변형, 모형화, 평가의 단계를 거쳐 수행하게 되는데, 이 중 모형화 단계에서 사용되는 기법들에 대한 연구가 활발히 진행되고 있다. 이러한 데이터 마이닝 기법은 기계학습 방법의 개념으로 대표되며, 이 중 가장 유명한 것은 신경망과 의사결정 나무분석방법이다.

신경망 기법은 인간의 신경-두뇌 시스템을 흉내낸 것으로 <그림 1>과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer) 등 3개의 층으로 구성되며, 입력층에서 보내지는 값을 가중치에 따라 은닉층이 합산하고 이를 활성함수(activation function)에서 변환하여 출력층으로 보내는 구조를 가지고 있다.



<그림 1> 신경망의 원리

일반적으로 로지스틱 함수와 쌍곡탄젠트 함수(hyperbolic tangent function)가 활성함수로 가장 보편적으로 쓰이며, 활성함수를 사용하여 입력층에서 이루어지는 자료처리 과정의 모습은 <그림 2>와 같다.



<그림 2> 신경망에서의 자료처리 과정

입력변수 n개, 1개의 은닉층과 은닉노드 m개, 1개의 출력층과 출력노드 c개인 신경망 모형에서 j번째 은닉노드의 값을 H_j라고 한다면, 이것은 다음과 같이 표현된다.

$$H_j = f(b_0 + \sum_{i=1}^{n} w_{ji} \times x_i), \quad j = 1, 2, \cdots, m$$

여기서 x_i는 독립변수, w_{ji}는 j번째 은닉노드와 i번째 독립변수간의 가중치, b_0은 bias, f는 활성함수가 되며, 여기서 활성함수를 로지스틱 함수로 사용한다면 다음과 같다.

$$f(s_j) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s_j}}$$

따라서, 은닉노드의 값 H_j는 0과 1사이의 값을 취하게 되며, 계산된 H_j는 다시 출력층의 함수로 들어가서 Y_k값을 구하게 된다.

$$Y_k = f_k(b_1 + \sum_{j=1}^{m} w_{kj} H_j), \quad k = 1, 2, \cdots, c$$

여기서 w_{kj}는 k번째 출력노드와 j번째 은닉노드의 가중치, b_1은 bias, f_k는 활성화/선택함수가 된다. 여기서도 활성함수를 로지스틱 함수로 사용하여 계산된 Y_k는 0~1값이 되며, 임계치 θ 를 기준으로 나뉘어진다.

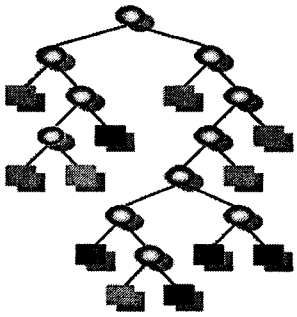
$$Y_k = \begin{cases} 1, & y_k > \theta \\ 0, & otherwise \end{cases}$$

최적의 임계치를 구하기 위하여 은닉노드와 출력노드의 가중치를 조정하게 된다.

의사결정나무기법은 다양한 알고리즘에 의해 분리가 이루어지며 이런 과정은 나무구조로 표현된다. 이러한 나무구조는 여러 가지 마디(node)라고 불리는 구성요소들로 이루어져 있으며, 나무구조가 시작되는 뿌리마디, 하나의 마디로부터 분리되어 나간 두 개 이상의 마디들인 자식마디, 자식마디의 상위마디인 부모마디, 각 나무줄기의 끝에 위치하는 끝마디 등 있다. 이러한 마디들의 분리기준은 어떤 입력변수를 사용하고 그 변수의 어떤 값을 기준으로 분리하는 것이 목표변수를 가장 잘 구별할 수 있는지에 초점을 두며, 몇 가지 알고리즘에 의해 분리기준이 정해지게 된다. (허명회 등, 2003)

CART(Classification and Regression Tree)는 설명변수들과 목표변수로 이루어진 자료들에서 설명변수들의 특성에 따라 자료들을 이지분류(binary split)하여, 2개의 하위노드를 생산하는 과정

을 반복하여 자료들을 목표변수의 값이 유사한 부분집합으로 만드는 방법으로 <그림 3>과 같이 나타난다.



<그림 3> CART의 분류방식

CART의 알고리즘은 마디의 순수함을 나타내는 지니지수(Gini Index)에 의해 분리여부를 결정하게 된다. 특정 변수에 의해 집단이 구분되면, 구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리하며 집단이 순수할수록 지니지수의 값이 작아지며 확률 또한 작아지게 된다. 지니지수는 다음과 같다.

$$\sum_{i=1}^{r} P(i)(1-P(i))$$

여기서 r은 목표변수의 범주의 수이며, $P(i)$는 주어진 자료 중 i범주에 분류될 확률을 나타낸다.

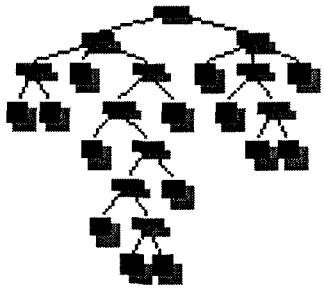
CART는 지니지수를 가장 감소시켜 주는 예측변수와 그 변수의 최적분리를 자식노드로 선택하는데, 지니지수의 감소량은 다음과 같이 계산한다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R$$

여기서 n은 부모노드의 관측치 수, n_L과 n_R은 각각 자식노드의 수를 의미한다. 즉, 자식노드로 분리되었을 때 불순도가 가장 작도록 자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)G_L + P(R)G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R$$

C5.0 알고리즘은 정보이론(Information Theory)에 따른 엔트로피(Entropy)개념을 이용하여 마디의 정보량에 따른 엔트로피 지수에 의해 분리가 되며 <그림 4>와 같이 나타나게 된다.



<그림 4> C5.0의 분류방식

분리된 마디에서의 집단의 정보량이 부족 할수록 엔트로피 지수는 커지게 되며, 많은 정보량을 가질수록 엔트로피 지수는 작아지게 되어 부족한 것이 없는 완전한 정보를 얻었음을 뜻하며, 엔트로피 지수는 다음과 같다.

$$-\sum_{i=1}^{r} p(i)(log\ (p(i))$$

여기서 r은 목표변수의 범주의 수이며, $p(i)$는 주어진 자료 중 i범주에 분류될 확률을 나타낸다.

만일, 자료 D가 변수 Y에 의하여 k개의 부그룹으로 분할되고, 각 부그룹의 상대적 빈도를 $p_1, p_2, \cdots, p_k$라고 하면, 자료 D의 엔트로피는 다음과 같이 계산된다.

$$info(D) = -\sum_{l=1}^{k} p_l log\ p_l$$

또한, 자료 D가 변수 X_i값을 바탕으로 자료를 분할하여 $D_1, D_2, \cdots, D_n$을 얻는다고 하자. $|D_j|$을 1개의 분할 자료 D의 크기라고 하면, 변수 X_i값을 바탕으로 분할된 $D_1, D_2, \cdots, D_n$의 엔트로피는 다음과 같이 계산된다.

$$infox_i(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} info(D_j)$$

일반적으로, 변수 X_i값을 바탕으로 자료 D를 분할하여 나온 정보와 단순히 자료 D를 구별하는데 필요한 정보의 차이가 나게 되는데, 이러한 정보의 차이량을 정보의 이득(Gain)이라 하며,

$$Gain(X_i) = info(D) - infox_i(D)$$

로 정의한다.

$Gain(X_i)$이 큰 순서대로 어떤 변수 X_i가 더 많은 정보를 가지고 있는지를 판단할 수 있으며, $Gain(X_i)$이 큰 값의 X_i부터 자료 D를 분할하게 된다.

이득(Gain)기준이 이론적으로는 명확하지만, 많은 수의 범주를 갖는 예측변수를 선호하는 편향이 내제되어 있어서 실제로 이득기준을 그대로 사용하지 않는다. 따라서, 이득 비율을 사용하여 실제적인 변수선택의 기준으로 한다.

$$GainRatio(X_i) = \frac{Gain(X_i)}{SplitInfox_i(D)}$$

여기서

$$SplitInfox_i(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} log(\frac{|D_j|}{|D|})$$

로 정의된다.

$SplitInfox_i(D)$는 자료 D를 단순히 $|D_1|,\cdots,|D_n|$에 비례하게 임의 분할하였을 때의 엔트로피를 나타낸다. 따라서, $GainRatio(X_i)$는 자료 D를 X_i로 분할함으로써 발생한 이득의 상대량을 의미한다.

간단히 요약하면 $Gain(X_i)$는 절대 비교, $GainRatio(X_i)$는 상대비교를 나타낸다. C5.0은 $GainRatio$를 이용하여, $GainRatio(X_i)$가 최대화되는 점에서 데이터의 분할을 선택한다. (허명회 등, 2003)

4. BMI의 주요인에 대한 데이터 마이닝 분석

2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 종합건강검진센터에 건강검진을 받은 20대 이상의 35,671명의 건강검진자료를 전체 데이터로 이용하였다. 종속변수를 BMI로 하고, 독립변수로 흡연력, 음주력, 운동습관, 결혼상태, 성별, 직업, 총 콜레스테롤, HDL 콜레스테롤, LDL 콜레스테롤의 변수를 이용하였다.

공정한 모형 평가를 위하여 전체 데이터를 각각 반으로 나누어 만든 Train Data와 Test Data를 이용하여 클레멘타인 프로그램을 사용하여 데이터 마이닝을 실시하였다. Train Data를 이용하여 각각의 분석방법으로 만들어진 모형을 Test Data에 적용하여 공정한 모형 평가를 실시하였으며, 데이터 마이닝 기법 중 신경망기법과 의사결정 나무분석을 이용하였다.

먼저 Train Data에서 비만집단과 비만이 아닌 집단의 비율차이에 대하여 균형을 맞추어 신경망 분석과 의사결정나무분석을 실시하였다. (허명회 등, 2003)

<그림 5~그림 10>은 Train Data와 Test Data에서의 분류율을 나타내고 있다.

$N-BMI 이분화

BMI 이분화		비만	정상	합계
비만	빈도	7527	2851	10378
	행 %	72.528	27.472	100
	열 %	64.621	31.368	50.048
정상	빈도	4121	6238	10359
	행 %	39.782	60.218	100
	열 %	35.379	68.632	49.954
합계	빈도	11648	9089	20737
	행 %	56.155	43.845	100
	열 %	100	100	100

<그림 5> Neural Network Train Data

$N-BMI 이분화

BMI 이분화		비만	정상	합계
비만	빈도	5174	2080	7254
	행 %	71.326	28.674	100
	열 %	54.172	25.124	40.684
정상	빈도	4377	6199	10576
	행 %	41.386	58.614	100
	열 %	45.828	74.876	59.316
합계	빈도	9551	8279	17830
	행 %	56.356	43.644	100
	열 %	100	100	100

<그림 6> Neural Network Test Data

$R-BMI 이분화				
BMI 이분화		비만	정상	합계
비만	빈도	6823	3492	10315
	행 %	66.146	33.854	100
	열 %	65.373	34.112	49.894
정상	빈도	3614	6745	10359
	행 %	34.888	65.112	100
	열 %	34.627	65.888	50.106
합계	빈도	10437	10237	20674
	행 %	50.517	49.483	100
	열 %	100	100	100

<그림 7> CART Train Data

$R-BMI 이분화				
BMI 이분화		비만	정상	합계
비만	빈도	4763	2491	7254
	행 %	65.660	34.340	100
	열 %	55.996	26.716	40.694
정상	빈도	3743	6833	10576
	행 %	35.391	64.609	100
	열 %	44.004	73.284	59.316
합계	빈도	8506	9324	17830
	행 %	50.526	49.474	100
	열 %	100	100	100

<그림 8> CART Test Data

$C-BMI 이분화				
BMI 이분화		비만	정상	합계
비만	빈도	7635	2750	10385
	행 %	73.519	26.481	100
	열 %	68.397	29.746	50.063
정상	빈도	3864	6495	10359
	행 %	37.301	62.699	100
	열 %	33.603	70.254	49.937
합계	빈도	11499	9245	20744
	행 %	55.410	44.590	100
	열 %	100	100	100

<그림 9> C 5.0 Train Data

$C-BMI 이분화				
BMI 이분화		비만	정상	합계
비만	빈도	5069	2185	7254
	행 %	69.879	30.121	100
	열 %	54.800	25.466	40.684
정상	빈도	4181	6395	10576
	행 %	39.533	60.467	100
	열 %	45.200	74.534	59.316
합계	빈도	9250	8580	17830
	행 %	54.706	45.294	100
	열 %	100	100	100

<그림 10> C 5.0 Test Data

<표 2> 데이터 마이닝 기법들의 총 정확도

	Train	Test	비고
Neural Network	66.374	64.97	
CART	65.629	65.135	
C5.0	**68.109**	**65.173**	← 최적의 모형

<표 2>에 따르면, Train Data에서 총 정확도는 Neural Network 66.374%, CART 65.629%, C5.0 68.109%로 나타났다. Train Data에서 만들어진 모형을 이용하여 Test Data에 적용한 결과 총 정확도는 Neural Network 64.97%, CART 65.135%, C5.0 65.173%로 나타났다. Train Data에서는 C5.0의 총 정확도가 가장 높게 나타나며, Test Data에서도 C5.0의 총 정확도가 높다는 것을 알 수 있다. 따라서, C5.0을 이용하여 BMI의 주요인에 대하여 살펴보았다.

C5.0 기법을 통하여 BMI의 주요인을 판별 해 본 결과, "**중성지방 > 성별 > 나이, HDL 콜레스테롤**"이 다른 요인들보다 높은 중요도를 차지하고 있는 것을 알 수 있다. C5.0기법에서의 BMI의 주요인에 대한 Decision Tree를 살펴보면 <그림 11>과 같다.

① 중성지방 = 위험그룹 (2,344)
  ② 성별 = 여자 (396)
    ③ 흡연력 = 비 흡연자 (372)
      ④ 운동습관 = 안한다 ⇒ 비만 (234, 0.702)
  ② 성별 = 남자 ⇒ 비만 (1,948, 0.771)
① 중성지방 = 경계그룹 (2,662)
  ② 성별 = 여자 (623)
    ③ 나이 = 50대 ⇒ 비만 (243, 0.626)
      나이 = 60대 ⇒ 비만 (2130, 0.754)
      나이 = 70대 이상 ⇒ 비만 (19, 0.632)
  ② 성별 = 남자 ⇒ 비만 (2,039, 0.708)
① 중성지방 = 정상그룹 (2,662)
  ② 성별 = 여자 (623)
    ③ HDL 콜레스테롤 = 경계그룹 ⇒ 비만 (550, 0.753)
      HDL 콜레스테롤 = 위험그룹 ⇒ 비만 (48, 0.771)
  ② 성별 = 남자 ⇒ 비만 (2,039, 0.708)
    ③ HDL 콜레스테롤 = 경계그룹 ⇒ 비만 (550, 0.753)
      HDL 콜레스테롤 = 위험그룹 ⇒ 비만 (48, 0.771)

<그림 11> BMI의 주요인에 대한 C 5.0의 Decision Tree

위 결과에서 보면 BMI분류에서 가장 주요한 요인으로 먼저 중성지방이 나타났다. 다음으로 성별로 분화가 되고, 나이와 HDL 콜레스테롤에 따른 분화가 세 번째로 이루어진다. 먼저 중성지방의 위험그룹에서 여성인 경우 흡연을 하지 않더라도 운동을 하지 않는다면 비만으로 판정이 되며, 남성인 경우 비만으로 판정이 된다. 중성지방의 경계그룹에서 여성인 경우 나이가 많으면 비만으로 판정이 되며, 남성인 경우 비만으로 판정이 된다. 중성지방의 정상그룹에서 여성인 경우와 남성의 경우 동일하게 HDL 콜레스테롤이 위험그룹과 경계그룹이 되면 비만으로 판정이 된다.

BMI의 주요인에 대한 대응위험도 (Odds Ratio)를 계산하여 개별적인 영향력에 대하여 알아보았다.

<표 3> BMI의 주요인에 대한 개별적인 Odds Ratios

위험요인		Odds Ratio	95%신뢰구간	
			하한	상한
중성지방	경계그룹 이상	3.215	3.054	3.386
	정상그룹			
성별	남자	2.538	2.246	2.655
	여자			
HDL 콜레스테롤	경계그룹 이상	2.613	2.483	2.750
	정상그룹			
나이	50대 미만	2.016	1.856	2.190
	50대 이상			

BMI의 주요인에 대한 Odds Ratio를 계산한 <표 3>의 결과에 의하면, 중성지방에서 경계그룹 이상과 정상그룹에 대한 Odds Ratio는 3.215로 나타나 경계그룹이상이 정상그룹보다 비만이 3.2배 높게 나타난다. 성별에서 남성과 여성에 대한 Odds Ratio는 2.538로 나타나 남성이 여성보다 비만이 2.5배 높게 나타난다. HDL 콜레스테롤에서 경계그룹이상과 정상그룹에 대한 Odds Ratio는 2.613으로 나타나 경계그룹이상이 정상그룹보다 비만이 2.6배 정도 높게 나타난다. 나이에서 50대 이상과 50대 미만에 대한 Odds Ratio는 2.016으로 나타나 50대 이상이 50대 미만보다 비만이 2배 정도 높게 나타난다.

Odds Ratio를 살펴보면 중성지방이 가장 높고, 다음으로 HDL 콜레스테롤, 성별, 나이 순으로 나타났다.

5. 결론

2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 종합건강검진센터에 건강검진을 받은 20대 이상 35,671명을 전체 데이터로 이용하여, BMI의 주요인에 대한 데이터 마이닝을 실시하였다. 전체 데이터를 Train Data와 Test Data로 나누어 공정한 모형 평가가 가능하게 하였다. Train Data와 Test Data에서 총 정확도가 가장 높은 C 5.0을 최종 모형으로 선택하여 BMI에 가장 영향을 많이 주는 요인을 살펴본 결과 가장 먼저 중성지방이 선택되었으며, 그 다음으로 성별과 HDL 콜레스테롤, 나이가 세 번째로 선택되었다. 중성지방의 위험그룹은 다시 성별로 나누어지며 여성인 경우 비 흡연자이지만 운동을 하지 않는 경우 비만으로 판정이 되며, 남성인 경우 비만으로 판정이 된다. 또한, 중성지방의 경계그룹은 다시 성별로 나누어지며 여성인 경우 나이가 많은 경우 비만으로 판정이 되며, 남성인 경우 비만으로 판정이 된다. 중성지방의 정상그룹에서는 다시 성별로 나누어지며, 남성이나 여성이나 모두 HDL 콜레스테롤이 경계그룹이상이면 비만으로 판정된다.

본 연구에서는 대규모 데이터를 바탕으로 데이터 마이닝 기법을 활용하였는데, 데이터 마이닝 기법은 확률적 접근이기 때문에 충분치 못한 데이터인 경우 오류가 점차 커질 수 있다. 또한, 비만의 주요인에 대하여 보다 단순화된 결론을 도출하기 위하여 BMI를 이분화시켜 분석을 실시하였다. 비만은 유전적, 사회 환경적, 경제적인 요소 등 많은 요소들이 복합적으로 연관되어 있다. 이러한 비만에 대한 정확한 분석을 위해서는 구체적인 평소 식생활습관이나 부모나 가족의 비만 정도 등에 대한 설문내용의 추가와 수정이 필요하다. 추가되고 수정된 설문내용을 바탕으로 한 데이터를 이용한다면 더 정확한 분류와 함께 결정적인 비만의 요인을 찾아낼 수 있을 것이다.

참고 문헌

[1] M. Baglion, U. Ferrara, A. Romei, S. Ruggier, F. Turini (2003). Preprocessing and Mining Web log Data for Web Personalization *Proc. of 8th Natl' Conf. of the Italian Association for Artificial Intelligence (AI*IA 2003)*, Paris (to be held September 2003), Italy.

[2] WHO (1997) Preventing and managing the global epidemic, *Report of a WHO Consultation on Obesity*, Geneva, 3-5 Jun. 1997

[3] WHO, West Pacific Region. The Asia-Pacific Perspective: *Refining Obesity and its Treatment*. IOTF. Feb. 2000.

[4] 이성원 (2001). Logistic modelling for receiver operation characteristic curves with neural networks, Ph.D, 영남대학교

[5] 오희숙, 천병렬, 감신, 예민혜, 강윤식, 김건엽, 이영숙, 박기수, 손재희, 이상원, 안문영 (2000). 농촌지역 주민들의 고혈압 발생 위험요인:1년간 전향적 추적 조사. *예방의학회지 Vol.33 No.2 p.231-238*

[6] 이성희 (2001). 비만이 고혈압 발생에 미치는 영향에 관한 후향적 코호트 연구. Ph.D 서울대학교.

[7] 최병권 (2004). 데이터 마이닝 기법을 이용한 제조업 부도예측 주요 변수 선택, 서울대학교

[8] 탁양주, 이영성, 이진석, 강재현 (2004). 최근 국내 비만 연구의 경향:1984년부터 2002년까지. *대한비만학회지 Vol.13 No.1 p.1-8*

[9] 허명회, 이용구 (2003). *데이터 마이닝 모델링과 사례*, SPSS 아카데미 p.29, 144-178

[2004년 12월 접수, 2005년 3월 채택]