대한의료정보학회지
제8권, 제3호 Volume8, Number3

데이터 마이닝 기법을 활용한 저체중 청소년의 특성탐색과 건강증진방안에 관한 연구

- 서울시 고등학생을 중심으로 -

신선미, 김인숙[1], 채영문[2], 김주형[1], 이희우[3]

연세대학교 보건학과 박사과정 연세대학교 대학원 간호학과[1], 연세대학교보건대학원[2], 서울시학교보건원[3]

A Study of Underweight Adolescents Characteristics and Health Promotion by Applying Data Mining Techniques

Sun Mi Shin, In Sook Kim[1], Young Moon Chae[2], Joo Hyung Kim[1], Hee Woo Lee[3]

Graduate School of Public Health, Yonsei University
College of Nursing, Yonsei University[1]
Graduate School of Health Science and Management, Yonsei University[2]
Seoul School Health Center[3]

Abstract

The purpose of this paper is to describe general characteristics of underweight adolescents and to search for ways to promote the health of underweight adolescents through assessing health related factors by using data mining techniques. The study sampled(n=4352) 1,180 underweight(BMI<18.5) and 3,172 average weight (18.5<=BMI<23) adolescents, 10th grade students in Seoul, 2000, and investigated the differences between two groups. Related variables were input in a decision tree and an association rule of SAS E-Miner. The most predictable model was CART. In frequency, the proportion of underweight adolescents was higher on the south of the Han-river than on the northern side; but in association rule, associated variables with high support rate and confidence rate were females, north of Han-river, and scoliosis. Therefore, approaches for health promotion of underweight adolescents are not only intervention of physical health, but also the education of proper weight perception to prevent low birth weight and underweight adolescents because mother's education and child's low birth weight are related to underweight adolescents. In conclusion, the following sample groups in Seoul are suggested: female adolescents with scoliosis on the north of Han-river in Seoul. ***(Journal of Korean Society of Medical Informatics 8-3,61~69, 2002)***

Keyword : Underweight, Data Mining, Health Promotion

교신저자: 신선미, 서울특별시 동대문구 휘경동 가든아파트 611
전화: 011-9732-5177 E-mail: healthteam@hanmail.net

I. 서 론

1. 연구배경 및 의의

청소년 중에서 비만이 많다고 하나[1] 실제로는 저체중 청소년이 더 많다. 1999년도 서울시 학교보건원이 고등학교 학생 7,342 명을 대상으로 실시한 종합신체검사 결과, 중등도 비만과 고도비만이 각각 2.7%, 0.5%인 반면 저체중과 심한 저체중은 18.8%, 3.4%로 밝혀져 오히려 비만보다 저체중 학생에 대해 집중적인 관심을 가져야 함을 알게 되었다[2].

경남 창원시에서도 지난 97년부터 3년간 시내 초등학교 학생 8,100명을 대상으로 어린이 비만도를 조사한 결과 지난 97년 비만 학생 비율이 27.5%였으나 98년 26.5%, 99년 14.8%로 줄어 들은 반면 저체중 어린이는 97년 11.8%에서 98년 12.1%로 늘어나기 시작해 99년에는 30.3%로 크게 늘어남을 알 수 있었다. 특히 '심한 저체중'으로 판정된 어린이가 전체의 5.0%로 나타나 지난 97년 1.4%, 98년 1.5%에 비해 3 4배 가량 늘어남을 알 수 있었다[3].

저체중이란 보통 '자기 신장에서 100을 뺀 수치에 0.9를 곱한 값을 표준체중'이라고 볼 때 표준체중의 90%미만을 말하며, 80%미만인 경우 심한 저체중을 의미한다. 저체중이 건강에 미치는 영향은 피로감, 저항력 감소, 유년, 청년기의 성장 지연, 내분비 장애 등 뿐 아니라 종종 다른 심각한 질병의 원인이 되기도 한다. 이런 현상은 젊었을 때에는 건강에 별 이상이 없는 경우도 많으나 중년이 되면 당뇨병, 고혈압 등 성인병에 걸릴 확률이 정상체중에 비해 훨씬 높아진다[4]. 또한 노년기 건강에서도 저체중인 사람들이 정상체중인 사람들에 비해 고혈압으로 인한 사망과 발작의 위험이 높다는 사실이 발견되기도 했으며[5], 저체중의 연약한 여성은 다시 저체중 신생아를 출생하고, 이 신생아는 저체중 학생과 성인으로 성장함[6]을 볼 때 청소년기의 저체중의 관리방안에 대한 중요성은 새롭게 인식해야 할 때이다.

이렇듯 학교보건측면에서 비만보다 더 많은 범주의 건강문제가 저체중임에도 불구하고 자신이 비만하다고 걱정하는 청소년은 많아도 체중이 적게 나가는 것을 걱정하는 청소년은 많지 않음을 우리 주변에서 쉽게 볼 수 있으며, 대부분의 학교보건행정가나 보건관련 교사들조차 저체중의 심각성과 잠재적인 건강문제에 대해 문제의식을 가지고 있지 않다.

2 목적

본 연구에서는 청소년의 건강문제 중 많은 범주를 차지하고 있는 저체중의 현황 및 일반적 특성을 파악한 후, 새로운 지식추출기법인 데이터 마이닝 기법을 활용하여 저체중 관련요인과 건강상태에 대한 특성을 파악함으로서 저체중 청소년의 건강증진 방향을 모색하고자 한다.

본 연구의 구체적인 목적은 다음과 같다.

첫째, Chi-square test와 단순 로지스틱 회귀분석을 통해 저체중 청소년의 일반적 특성 및 관련변수를 파악한다.

둘째, 최적의 저체중 청소년 분류모형을 구축하기 위해 다중 로지스틱 회귀모형, CHAID[A], C4.5[B], CART[C] 모형의 안정성과 예측력을 평가한 후 그중 가장 정확도가 높은 모형을 이용하여 저체중 청소년의 특성을 파악한다.

셋째, 연관성 규칙을 통해 저체중 청소년의 관련성을 탐색한다.

II. 재료와 방법

1. 연구설계

본 연구는 저체중 청소년의 현황을 파악하고, 정상체중 청소년과 저체중 청소년의 일반적 특성을 비교한 후, 데이터 마이닝 기법을 이용하여 저체중 청소년의 분류 특성과 관련성을 파악하기 위한 탐색연구이다.

A. Chi-squared Automatic interaction Detection : 변수간의 통계적 관계를 알아내기 위해 변수들간의 상관관계를 이용하여 의사결정나무를 만듬.
B. C45 : ID3(Iterative Dichotomizer 3)의 후기 버전, 잠재적인 분할자들을 비교하기 위해 엔트로피지수를 분리기준으로 사용한다.
C. Classification and Regression trees : 가장좋은 분할인 관측치의 불순도를 크게 감소시키기 위해 지니지수와 분산을 사용한다.

2 연구대상

본 연구에서는 24개 고등학교에서 신체검사를 받은 1학년 학생 10,300명의 자료 중, 인구, 심리, 사회학적 특성을 묻는 설문지를 비교적 성실하게 응답한 14개 고등학교 1학년 학생 5,188명을 데이터 마트로 구축한 후 BMI(Body Mass Index kg/m²) 18.5미만인 저체중 청소년 전수 1,180명과 BMI 18.5이상 23미만의 정상체중 청소년 전수 3,172명, 총 4,352명을 연구대상으로 선정하였다. 그 후 분석용(Train set)으로 3,264명(75%), 평가용 데이터(Validation set)로 1,088명(25%)를 분할 후 모형의 안정성과 예측력을 평가하였으며, 그 중 가장 설명력이 좋은 모형을 이용하여 최적의 저체중 분류모형을 제시하였다. 또 연관성 규칙을 이용하여 저체중 청소년의 관련성을 탐색하였다.

3 분석방법

SAS(version 8.1)와 SAS Enterprise Miner(version 3.0)를 사용하여 다음과 같이 분석하였다.

1) 저체중 청소년과 정상체중 청소년의 일반 특성 비교는 빈도와 백분율, X^2을 통해 알아보았다.

2) 저체중 청소년의 관련변수를 파악하기 단순로지스틱회귀모형을 이용하여 알아보았다. 또 이중 유의한 변수는 의사결정나무모형의 입력변수로 투입하였다.

3) 데이터마이닝기법을 이용한 저체중 분류 모형 구축에 앞서, 5 fold Cross-validation 평가방법을 이용하여 다중 로지스틱회귀모형, CHAID, C4.5, CART모형의 안정성과 예측력을 평가하였다.

4) 그 중 가장 정확률이 높은 CART 모형을 선택하여 저체중 청소년 분류모형을 구축 후 저체중 청소년의 특성을 알아보았다.

5) 또 연관성 규칙을 통해 저체중 청소년의 관련 변인들 사이의 상호관련성을 파악하였다.

III. 연구결과

1. 연구대상자의 일반적 특성

1) 저체중의 분포

연구 데이터 마트 5,188명 중 저체중 청소년의 분포(BMI 18.5 미만)는 22.74%이었고, 그 중 남자는 27.94%, 여자는 21.05%를 차지 함으로서, 남학생의 저체중 분포가 여학생보다 더 많았다. 반면 비만 청소년(BMI 25 이상)은 6.96%를 차지하였고 그 중 남

Table 1. Characteristics of study subjects

특 성	변 수	구 분	저 체 중		정 상 체 중		X^2-value	P-value
			실수	(%)	실수	(%)		
인구학적특성	성별	남자	356	(33.74)	699	(66.26)		
		여자	824	(24.99)	2,473	(75.01)	30.97	<.001
		계	1,180	(27.11)	3,172	(72.89)		
	모	국졸이하	36	(32.73)	74	(67.27)		
	교육	고졸이하	402	(25.69)	1,163	(74.31)		
	정도	전문대졸	9	(37.50)	15	(62.50)	9.80	0.020
		대졸이상	70	(20.29)	275	(79.71)		
		계	517	(25.29)	1,527	(74.71)		
신체건강특성	요당	음성	1,154	(26.94)	3,130	(73.06)		
		양성	26	(38.24)	42	(61.76)	4.32	0.037
		계	1,180	(27.11)	3,172	(72.89)		
	척추	정상	1,123	(26.69)	3,084	(73.31)		
		측만증	41	(38.32)	66	(61.68)	7.15	0.007
		계	1,164	(26.98)	3,150	(73.02)		

* 무응답 제외

자는 9.89%, 여자는 6.0%였다. 이는 저체중 분포와 마찬가지로 남학생의 비만분포가 여학생보다 더 높았다.

2) 연구대상자의 특성

정상체중군과 차이가 있는 저체중군의 특성을 살펴보면 여자보다는 남자에게, 어머니 교육수준이 대졸일때보다는 대졸이 아닐 때 저체중 청소년이 많았다. 저체중 청소년과 정상체중 청소년의 신체적 건강상태를 비교해보니 저체중 청소년에게 요당과 척추측만증이 많았다. 즉 요당이 음성인 경우 저체중의 비율은 26.94%인 반면 양성인 경우의 저체중의 비율은 38.24%이었고 척추가 정상인 경우 저체중은 26.69%인 반면 척추측만증인 경우 저체중은 38.32%이었다(Table 1). 그외의 요단백, 혈색소, 식전혈당, 총콜레스테롤, 폐결핵, 건강문제 등의 신체건강상태, 경제적 요인, 학교지역, 건강행위, 스트레스, 가족지지, 우울감, 자아존중감 등의 문항에서는 두 군간에 통계학적 차이가 없었다.

3) 저체중과 관련된 변수파악

저체중과 여러 독립변수와의 관계를 알아보고, 데이터 마이닝의 투입변수를 선별하기 위해 단순 로지스틱 회귀분석을 실시했다. 그 결과 '여학생'에 비해 '남학생'이 1.52배 만큼 저체중이 많았으며, 어머니 교육이 '국졸이하'에 비해 '대졸이상'인 경우 0.52배로 저체중이 적었다.

'종교가 없는 학생'에 비해 '기독교'인 학생이 1.14배, 뇨당은 없을 때보다 있을 때 1.68배, 척추측만증이 없을 때보다 있을 때 1.7배, 강북지역 학교보다 강남지역 학교에 1.15배로 저체중이 많았다.

그러나, 가족지지, 자아존중감 그리고 한달동안의 우울감은 통계학적으로 의미가 없었다(Table 2).

2. 저체중 청소년 분류를 위한 데이터 마이닝 모형 선택

1) 모형평가에 투입된 변수

모형평가에 이용된 변수 중 목표범주는 '저체중'이고 설명변수는 대상자의 키와 체중 및 일반적 특성 파악단계에서 통계적 차이가 유의한 변수와 선행연구에서 관련 변인이었던 변수였다(Table 3).

Table 3. Input variables in model assessment

변수명	범 주
체중그룹	BMI 기준에 따른 저체중/정상체중
성별	남/ 여
체중	대상자 본인의 체중
키	대상자 본인의 키
어머니 교육상태	국졸이하/ 고졸이하/ 전문대졸/ 대졸이상
종교	기독교/ 천주교/ 불교/ 무교 또는 기타
뇨당	신체검사에서 뇨당 있다/ 없다
신생아 체중	태어날 때 저체중/ 정상체중/ 과체중
척추측만증	X-Ray검사에서 척추측만증 있다/ 없다
자살시도	지난 1년간 자살 시도 한적 있다/ 없다
학교 지역구분	학교 위치지역 강북/ 강남
스트레스 및 신체건강 상태	스트레스 유/ 무, 감기 유/무
가족지지	아주 적다/ 적다/ 약간 많다/ 많다
우울감	아주적다/ 적다/ 약간 많다/ 많다
자아존중감	아주 적다/ 적다/ 약간 많다/ 많다

Table 2. Results of logistic regression analysis

구 분	변 수	기 준	구 분	Parameter Estimate	Odds ratio	Pr>Chisq	95% wald confidence limits
인구	성별	(여)	남	0.21	1.52	<.001	1.31-1.77
특성	모(母)교육	(국졸이하)	대졸이상	-0.45	0.52	0.003	0.32-0.84
	종교	(없음)	기독교	0.14	1.14	0.050	0.91-1.43
신체	뇨당	(없다)	있다	0.25	1.68	0.030	0.36-0.97
건강	척추측만증	(없다)	있다	0.26	1.70	0.000	1.14-2.53
심리	학교지역	(강북)	강남	0.07	1.15	0.030	1.00-1.31
사회	가족지지	(많다)	아주적다	-0.02	0.98	0.830	0.73-1.31
특성	자아존중감	(많다)	아주 적다	0.06	1.22	0.460	0.92-1.60
	지난한달간우울감	(많다)	아주 적다	-0.08	0.74	0.390	0.50-1.09

Table 4. Means of 5-fold cross validations

모형의 종류	CHAID		C4.5		CART		Logistic regression	
	분석용	평가용	분석용	평가용	분석용	평가용	분석용	평가용
평균 정분류율	85.71	85.76	85.70	85.98	85.86	86.11	77.53	76.92
평균 민감도	69.49	70.89	69.41	70.34	69.34	70.53	23.71	23.84
평균 특이도	91.69	91.46	91.72	91.66	91.96	91.86	97.35	97.37

2) 모형평가 결과

모형 평가 결과 5회에 걸쳐 실시한 Cross-validation 평가에서 각 회마다 분석용과 평가용의 결과가 큰 차이를 보이지 않음으로서, 모형구축과 예측력 평가에 있어 안정적임을 알 수 있었고, 5회 Cross-validation 평가의 평균을 비교해 볼 때 분석용에서 정분류율이 가장 높은 모형은 CART모형이었고, 민감도가 가장 좋은 모형은 CHAID모형, 특이도가 가장 좋은 모형은 CART모형임을 알 수 있었다(Table 4). 또 평가용에서도 CART모형이 가장 정분류율이 높게 나왔으며, 분석용과 평가용 모두 가장 정분류율이 낮은 모형은 Logistic regression모형임을 알 수 있었다(Table 4).

3. CART알고리즘에 의한 의사결정 나무 모형

Cross-validation결과 모형의 정분률이 가장 높은 모형은 CART모형이었다. 이에 본 연구에서는 CART 모형의 의사결정나무분석을 제시하였다. 제시된 모형은 모형 평가 때와 똑같은 분류기준과 분할기준(분석용 75%, 평가용25%)이었으나 키와 체중은 투입변수에서 제외했다. 그 이유는 변수 중 소수이지만 중요한 관심의 대상인 변수가 키와 체중변수로 인해 가려지는 것을 방지하기 위함이다[7]. 그림에서 마디내 왼쪽 값은 분석용 데이터에 대한 결과이고, 오른쪽 값은 평가용 데이터에 대한 결과이다(Fig 1, Fig 2).

분석용 3,264명중에 저체중은 896명으로 27.45%(base line gain %)였다. 이중 남자 770명 중에서 저체중은 33.4%로서 여자 2,486명 중 저체중 25.6%보다 높았다(Fig 1).

남자 중 종교가 있는 경우 저체중은 35.2%로서 종교가 없을 때의 저체중 19.8%보다 높았다. 또 종교있는 남자 중에서 척추측만증이 있을 때 저체중 58.8%로서 척추측만증이 없을 때의 저체중 34.6%보다 높았다. 종교 없는 남자의 경우 어머니 교육이 전문대졸 이하인 경우 저체중 22.2%로서 대졸이상인 경우의 저체중 0%보다 높았다. 또한 남자 중 종교가 없고 어머니 교육이 전문대졸 이하이면서 강남지역의 학교인 경우 저체중 31.6%로서 강북학교 저체중 14.0%보다 높았다. 계속해서 강남의 가지를 따라가 보면 신생아 때 저체중 또는 정상으로 태어난 경우 현재 저체중 36.4%로 과체중으로 태어난 경우의 현재 저체중 0%보다 높았다. 어머니 교육상태가 국졸 이하 경우 현재 저체중 청소년은 80%로서 어머니교육이 중,고등학교 및 전문대졸인 경우의 청소년 저체중 28.6%보다 높았다.

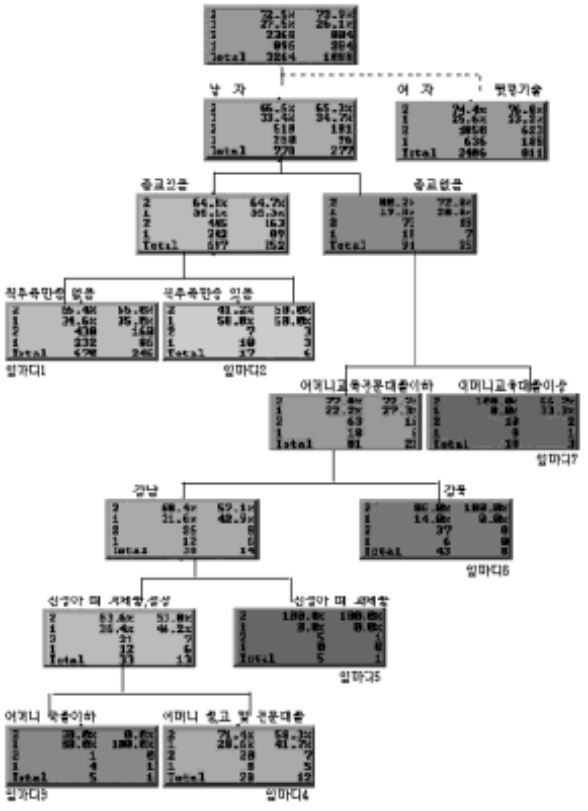

Fig 1. Decision tree in CART model(Male)

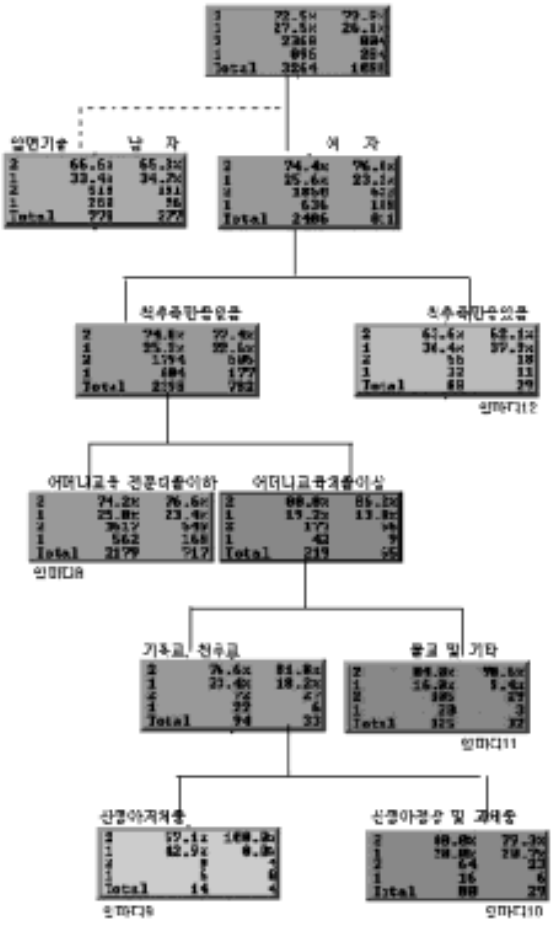

Fig 2. Decision tree in CART model(Female)

여자에서 보면 척주측만증이 있을 때 저체중은 36.4%로서 척추측만증이 없을 때의 25.2%보다 높았다(Fig 2). 또한 여자에서 척추측만증이 없을때 어머니 교육을 보면 전문대졸이하가 저체중 25.8%로서 대졸이상 저체중 19.2%보다 높았다. 이어서 여자에게 척추측만증이 없고, 어머니 교육이 대졸이상이며, 종교가 기독교 또는 천주교일때 저제충 청소년은 23.4%로 불교 및 기타 16.0%보다 높았다. 또 종교가 기독교 또는 천주교이면서 신생아 출생시 저체중인 경우 현재 저체중은 42.9%로서 신생아때 정상체중 및 과체중인 경우의 현재 저체중 20.0%보다 높았다.

4. 연관성 규칙을 통한 관련성 탐색

본 연구결과 도출된 연관성 규칙은 약 1,000여개에 이르렀다. 그러나 이러한 연관성 규칙은 모두 유용하다고 볼 수 는 없으므로, 먼저 의미있는 연관성 판단을 위해서는 전체대상자 중에서 저체중 대상자와 다른 변수가 동시에 있을 사건의 확률($Pr(A \cap B)/N$)을 살펴보아야 할 것이다[8]. 이런 확률을 지지도(Support)라고 한다.

또 저체중 대상자와 다른 변수가 동시에 있을 조건부 확률($Pr(A \cap B)/P(A)$)을 살펴보아야 한다. 이런 확률을 신뢰도(Confidence)라고 한다[8]. 향상도(Lift : $Pr(A \cap B)/Pr(A)Pr(B)$)는 1에 가까우면 독립에 가까운 사건, 1보다 크면 연관관계로 판단하면 될 것이다. 따라서 의미 있는 연관성 규칙이 되려면 향상도 값이 1이상이 되어야 할 것이다[8].

이에 따라 본 연구에서는 연관성 규칙의 자료를 신뢰도, 지지도, 향상도가 높은 순으로 정렬시킨 후 그 중 가장 높은 순위를 차지한 25개의 연관성 규칙을 제시하였다(Table 5).

연관성 규칙을 통한 관련성을 탐색해보면 현재 저체중이면서 출생시 체중이 정상이었고, 여자의 조건을 가진 청소년이 강북의 학교를 다닐 확률은 전체 학생 중 10.53%를 차지하고 있었고, 이런 조건을 가졌을 때 100%가 강북에서 학교를 다니고 있었다. 또 향상도 값이 1.55배로서 양의 연관관계가 있음을 알 수 있다.

특히 강북에 사는 학생 중 저체중이면서, 가족지지가 적고, 여자인 경우는 전체 대상자중 5.04%에 해당하나, 가족지지가 적은 여자 청소년의 100%가 모두 강북지역의 학교를 다니고 있음을 알 수 있다. 또 강북에 살면서 스트레스가 많은 여자 저체중 청소년은 11.42%에 불과했으나, 스트레스가 많은 여자 저체중 청소년의 99.67%가 강북에 살고 있었다. 즉 스트레스가 많은 여자 저체중 청소년은 약 309명이고 이 309명이 거의 다(99.67%)가 강북에 살고 있음을 알 수 있다.

또 여자이고, 척추 측만증을 가지고 있으며, 저체중 청소년이 강북지역에서 학교를 다닐 확률은 15.41%이나, 99.04%의 신뢰도를 가지고 있었다.

이처럼 연관성 규칙에서는 어느 조건을 가진 대상자가 어느 특성을 가지고 있는가를 볼 수 있었다. 그러므로 향후 건강증진 전략수행에 있어 지지도와 신뢰도가 높은 집단을 목표집단으로 삼으면 효율적일 것이다. 특히, 신뢰도가 높은 집단을 목표집단으로

Table 5. Association rule in underweight adolescents

Confidence	Support	Lift	Count	Rule
100	10.53	1.55	285	저체중 & 출생시체중정상 & 여자 ==> 강북
100	7.57	1.55	205	저체중 & 비인문고 & 여자 ==> 강북
100	5.73	1.55	155	저체중 & 부 대졸 & 여자 ==> 강북
100	5.04	1.55	137	저체중 & 가족지지 적음 & 여자 ==> 강북
99.67	11.42	1.54	309	저체중 & 스트레스 많다 & 여자 ==> 강북
99.54	8.17	1.54	221	저체중 & 일년간 두통경험 유 & 여자 ==> 강북
99.41	6.28	1.54	170	저체중 & 인문고 & 1년간 감기 경험 유 ==> 강북
99.31	5.36	1.53	145	저체중 & 자아존중감 많음 & 여자 ==> 강북
99.27	5.02	1.53	136	저체중 & 인문고 & 일년간 두통경험 유 ==> 강북
99.19	9.13	1.53	247	저체중 & 모 고졸 & 여자 ==> 강북
99.09	16.26	1.53	440	저체중 & 여자 ==> 강북
99.04	15.41	1.53	417	저체중 & 척추측만증 유 & 여자 ==> 강북
98.58	7.72	1.52	209	저체중 & 부 고졸 & 여자 ==> 강북
98.56	10.12	1.52	274	저체중 & 우울감 많다 & 여자 ==> 강북
98.45	7.06	1.52	191	저체중 & 스트레스 많다 & 인문고 ==> 강북
98.32	8.68	1.52	235	저체중 & 인문고 & 여자 ==> 강북
98.15	5.91	1.45	160	저체중 & 출생시체중정상 & 부 고졸 ==> 척추측만증 유
98.14	5.87	1.52	159	저체중 & 인문고 & 출생시체중정상 ==> 강북
97.53	5.84	1.44	158	저체중 & 1년간 감기 경험 유 & 부 고졸 ==> 척추측만증 유
97.42	6.98	1.44	189	저체중 & 출생시체중정상 & 모 고졸 ==> 척추측만증 유
97.40	5.54	1.44	150	저체중 & 일년간 두통경험 유 & 모 고졸 ==> 척추측만증 유
97.38	6.87	1.44	186	저체중 & 1년간 감기 경험 유 & 모 고졸 ==> 척추측만증 유
97.27	9.24	1.44	250	저체중 & 부 고졸 ==> 척추측만증 유
97.00	5.98	1.43	162	저체중 & 우울감 많다 & 부 고졸 ==> 척추측만증 유
96.96	8.28	1.43	224	저체중 & 강북 & 부 고졸 ==> 척추측만증 유

선정 후 건강중재방안을 제시한다면, 시간과 경비 면에서 가장 효율적인 건강증진 전략이 가능할 것이다.

Ⅳ. 고찰 및 결론

본 연구결과 저체중의 관련변인을 살펴보면, 어머니 교육정도, 종교, 학교지역, 신생아 때의 체중, 척추측만증, 뇨당검출 등이었는데, 이런 결과는 외국의 선행연구결과와 다소 유사하기도 하고, 다소 차이가 있기도 했다.

가나의 도시 어린이에 대한 연구에서 출생시 저체중과 교육받지 못한 어머니가 청소년기의 저체중의 중요한 요인이다[9]라고 하였다. 본 연구결과에서도 출생시 저체중은 과체중에 비해 1.13배로 저체중이 많았으나, 이는 일반 통계 부분에서는 통계학적으로 유의하지 않았다(p-value 0.88). 그러나 의사결정나무분석의 CART모형에서는 신생아때의 체중을 중심으로 그 가지가 나뉨으로서 저체중 청소년이 신생아 때의 체중과도 관련이 있는 경향이 있음을 알 수 있었다. 이는 데이터 마이닝 분석이 일반통계분석보다 가급적 정보를 버리지 않고 많이 사용하여 좀더 자세한 정보를 준다는 특징 때문일 것으로 생각한다. 어머니의 교육정도는 본 연구에서도 국졸이하의 어머니 보다 대졸이상의 어머니일 때 0.52배로 저체중 청소년이 적음을 알 수 있었다.

또 Paricio[10]는 식사의 공급, 생활 조건 정도, 성인남자의 실업률 증가로 인한 본인 및 배우자의 부담증가가 그 사회에서 저체중의 건강문제가 높아지게 하는 요인이라고 발표하였으나, 본 연구의 결과는 정상체중 청소년보다, 저체중 청소년이 오히려 긍정적 경제상태로 응답하는 경향을 보였고, 또 강남학생에게 저체중 학생이 더 많은 것으로 볼 때, 외국의 선행연

구와는 다소 다른 결과를 보여주는 듯 했다. 그러나 데이터 마이닝의 연관성규칙에서는 저체중과 관련된 신뢰도가 높은 상위 25개의 연관성 규칙에서 저체중 학생과 강남지역 학생과의 연관성 규칙이 하나도 없는 반면, 우울감, 스트레스, 낮은 자아존중감을 가진 저체중 청소년과 강북학생과의 연관성 규칙은 12개를 차지하는 것으로 보아, 경제적 수준과 저체중은 어떤 연관이 있는 것으로 추정되어진다. 그러므로 추후 연구에서는 삶의 수준정도와 저체중과의 관련 등의 심도있는 연구가 필요할 것이다.

문헌고찰에서 볼 수 없었으나, 본 연구에서의 변인으로 나온 것을 살펴보면, 종교, 뇨당, 척추측만증을 들 수 있었다. 즉 저체중 청소년의 종교는 기독교가 많았고, 정상체중 청소년에게는 기독교가 가장 적은 분포를 보였다. 저체중 청소년의 신체건강상태에서는 뇨당이 정상체중 청소년보다 많이 검출되었고, 척추측만증의 유병율이 정상체중 청소년보다 높았다.

데이터 마이닝의 연관성 분석에서는 두통, 감기, 척추측만증등의 건강문제와 부족한 가족지지, 과도의 스트레스, 낮은 자아존중감, 과도의 우울감 등의 심리사회적 요인에 대한 문제를 가지고 있는 저체중 청소년은 강남보다는 강북에 많은 것을 볼 때 지역간 건강수준의 차이를 의심할 수 있었다. 이런 결과는 이태화와 신선미[11]의 연구에서 고등학생의 흡연, 음주, 식습관, 운동, 스트레스 관리 등의 건강증진행위 실천의 지역간 비교에 있어 강남의 학생이 강북의 학생보다 전반적인 건강증진 행위 실천을 더 잘하고 있다는 결과와도 그 맥락을 같이 하는 것으로 여겨진다. 이런 결과는 우리나라 학교보건관리시스템의 개선을 제안할 수 있는데 왜냐하면 현재 모든 지역의 학교와 학생에게 실시하고 있는 건강검사와 이상자 관리는 지역차이와 개인수준차이를 고려하지 않고 많은 수의 학생을 짧은시간내에 형식적 방식으로 관리하고 있기 때문이다[2]. 그러므로 저체중 청소년의 건강증진프로그램의 접근에 있어 전체학생을 대상으로한 획일적 접근보다는 강북지역 청소년과 척추측만증을 가진 학생을 목표집단으로 선정하여 수행할 때 비용-효율적일 것임을 알 수 있다.

본 연구의 제한점 및 제안점을 살펴보면, 첫째, 저체중(기준-BMI 18.5미만)과 정상체중(기준-BMI 18.5 이상 23미만)사이의 경계체중군으로 인해 저체중 청소년의 구별에 있어 오분류율(miss classification)이 높음을 알 수 있었다. 그러므로 우리나라의 특성과 성장기 청소년의 특성에 적합한 체중분류기준의 개발이 필요하다고 본다. 또 향후 연구에서는 경계체중을 통제한 후 연구를 수행한다면, 좀더 결정적인 저체중의 변인을 찾아낼 수 있을 것이다.

둘째, 본 연구의 데이터는 데이터 마이닝을 염두에 두고 장기간에 걸쳐 계획된 것이 아니고, 집단 설문으로 인한 응답자의 성의여부, 또 높은 무응답 비율 등의 제한점이 있다. 그러나 본 연구결과 동일한 분류기준으로 분석하였을 때 데이터 마이닝의 의사결정분석은 로지스틱 회귀모형보다 저체중 분류에 있어 정확률이 높았고, 여러 특성이 결합된 규칙을 도출할 수 있는 등의 장점을 가지고 있다. 그 이유는 로지스틱 회귀모형은 변수들간의 관계가 복잡한 비선형성(nonlinearity)을 가지는 경우에는 예측의 유용성 측면에서는 한계가 있고, 각 입력변수의 영향이 다른 입력변수에 종속되어 있지 않다고 가정하고 있어 유용한 교호작용(intreraction)을 탐색하는 것은 실제적으로 어렵기 때문이다[12]. 이에 비해 의사결정나무와 같은 분석방법은 유용한 입력변수나 교호효과 또는 비선형성을 자동적으로 찾아내는 알고리즘이라고 할 수 있기 때문이다. 그러므로 충분한 사전 계획 즉, 명확한 문제 찾기, 양질의 데이터 마트 구축, 프로젝트 경험이 풍부한 인력이 전제된다면, 보건계에서 데이터마이닝 활용은 단지 조각 정보로 굴러다니는 데이터를 지식으로 바꿀 수 있는 좋은 기회이며, 방법일 것이다.

참고문헌

1. Report on national health and nutrition survey, Korea lipid and Atherosclerosis. 2001;11(3):408-410
2. Lee. H. W, Shin. S. M. A study of health-related behavior in 10th grade students of 12 high schools located in Seoul, yearbook of School Health, 1999;29:
3. Hankyoreh newspaper. 2000.125
4. Huh K. B. Pathogenetic Heterogeneity of Type 2 Diabetes Melitus in Korea The Journal of Korean Diabetes Association. 1999;23(1):62-69
5. the Reuters News. 1999.12.10
6. Behrman RE, Kliegman RM, Arvin AM, Nelson textbook of pediatrics, 15th ed. W.B. Saunders Company, Philadelphia, 1995:169-172
7. Kim, Y.D. Growth of Tree Model by one side purity. Korean intelligent information system society, 2000;7(1):17-25
8. Choi K. R. Theory and Practice of Datamining, Chonggu , Seoul, 2000:116-143
9. Rikimaru T. Risk factors for the prevalence of malnutrition among urban children in Ghana. J Nutr Sci Vitaminol(Tokyo), 1998;44(3):391-407
10. Paricio JM. Health examination of children from the democratic Sahara Republic(North West Africa) on Vacation in Spain. An Esp pediatr, 1998;49(1):33-38
11. Lee T. H, Shin. S. M. Health Promotion Behavior and Related Psychosocial variables among High School Students in Seoul. The journal of Korean Community Nursing, 1998;11(1):459-467
12. Kang H. H. Datamining Free Academy, Seoul, 2000;1:153-16