저체중 청소년의 건강증진을 위한 데이터 마이닝 응용

- 서울시 고등학생을 중심으로 -

연세대학교 대학원

간 호 학 과

신 선 미

저체중 청소년의 건강증진을 위한 데이터마이닝 응용

- 서울시 고등학생을 중심으로 -

지도 김 인 숙 교수

이 논문을 석사 학위논문으로 제출함

2001년 12월 일

연세대학교 대학원

간 호 학 과

신 선 미

신선미의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2001년 12월 일

감사의 글

왜 저체중 청소년이 많을까? 라는 의문은 제가 석사과정동안 자주 생각했던 연구문제였습니다. 저체중 청소년을 주제로 석사논문을 생각하고 있던 중 김인숙 교수님이 주신 데이터 마이닝에 대한 아티클은 한 눈에 저를 사로잡았습니다. 그래서 부족하기만 한 연구능력에도 불구하고 저체중 청소년의 특성과 데이터 마이닝 기법 활용이 자연스럽게 제 석사논문의 주제로 결정되었습니다.

논문을 진행하면서 교수님들 도움이 많았습니다. 우선 데이터 마이닝을 처음 제시하여 주신 김인숙 교수님과 도전의식을 갖게 해주신 채영문 교수님, 연구 진행 단계 전반에 걸쳐 폭넓은 가르침을 주신 김주형 교수님께 감사드립니다. 또 국방연구원의 김종탁 교수님은 데이터 마이닝에 대해 아무 것도 모르면서 단지 용기만 있는 제게 청강을 허락해주시고, 끊임없이 일어나는 통계적 의문에 대해 중요한 방향을 제시해주신 고마운 분이었습니다.

데이터 마이닝 수업에서 만난 동료학우들의 도움도 컸습니다. 산업 시스템 공학과 오기열씨와 신형원씨는 마이닝 tool에 익숙해 질 수 있도록 도와주었고, 기술경영학과 최종문씨는 사뭇 생소하기만 한 모형 평가과정에서 자료와 시간을 아끼지 않았으며, 항상 도와 줄 것 같은 경영학과 임형채씨와의 만남도 기쁩니다. 또 고려대학교 정보통계학과 김수택씨는 수십차례 오간 이메일을 통해 제가 논문을 완성할 수 있도록 많은 해결책을 제시하곤 했습니다.

뿐만 아니라 힘을 북돋아 준 친구 미라, 힘들 때마다 찾게 되는 상남과 민희, 캐나다에 가서도 배려해주는 유정, 귀여운 무악학사 방 짝궁 경순도 잊을 수 없습니다. 끝으로, 이 논문을 기쁘게 받아줄 봉선, 제선, 문선오빠와 효숙언니에게 감사하며, 저체중 청소년에 대한 관심을 계속 가질 수 있도록 동기부여가 된 서울시 학교보건원 김종희 원장님과 연구의 시작부터 끝까지 함께 한 이희우 과장님께 진심으로 감사드립니다.

2002년 1월 신 선 미

차 례

그림차례

표 차 례

국 문 요 약

저체중 청소년의 건강증진을 위한 데이터 마이닝 응용

본 연구의 목적은 청소년의 건강문제 중 많은 범주를 차지하고 있는 저체중의 현황 및 일반적 특성을 파악한 후, 새로운 지식추출기법인 데이터 마이닝 기법을 활용하여 그 관련요인과 건강상태에 대한 특성을 파악하고 평가함으로서, 저체중 청소년의 건강증진 방향을 모색하는 것이다.

이에 연구방법은 첫째, 서울시 고등학교 1학년 학생 10,300명의 신체검사 자료 중 설문지를 비교적 충실히 작성한 5,188명에 대하여 연구 데이터 마트를 구축하였다. 둘째, 그중 BMI 18.5미만인 저체중 청소년 전수 1,180명과 BMI 18.5이상 23미만의 정상체중 청소년 전수 3,172명, 총 4,352명을 연구대상자로 추출한 후 두 집단의 일반적인 특성파악을 위해 빈도분석, 카이제곱 분석, 단순 로지스틱 회귀분석을 통해 그 두집단의 차이를 알아보았다. 그후, 전단계에서 얻어진 유용한 변수, 외국의 선행연구에서 밝혀진 저체중 관련 변수를 데이터 마이닝의 의사결정나무모형과 연관성 분석의 입력변수로 투입하였다.

또 가장 최적의 분류와 예측모형의 구축을 위해 교차타당성(Cross-validation) 방법을 통해 의사결정나무 분석중 CHAID, C4.5, CART 모형과 로지스틱회귀 모형을 평가하였으며, 그 중 가장 안정성과 예측력이 좋은 CART 모형으로 의사결정나무를 형성 후 제시하였다.

연구 결과는 다음과 같았다.

첫째, 연구 데이터 마트 5,188명 중 저체중 청소년의 분포(BMI 18.5 미만)는 22.74%를 차지함으로서, 현재 보건의료계의 주 관심인 비만(BMI 25이상)청소년의 6.96% 보다 더 많은 분포를 차지하고 있었다. 또 여학생보다 남학생이 1.52배, 강북보다는 강남에 1.15배 저체중 청소년이 많은 것으로 나타났다.

둘째, 정상체중 청소년보다 저체중 청소년이 더 많이 갖고 있는 신체 건강문제

는 척추측만증과 뇨당으로서, 척추측만증은 없을 때 보다 있을 때 1.7배, 뇨당검출 없을 때 보다 있을 때에 1.68배로 저체중이 많았다. 인구학적 측면에서는 어머니 교육정도가 대졸일 때, 국졸이하에 비해 0.52배로 저체중 청소년이 적음을 알 수 있었고, '종교가 없는 학생'에 비해 '기독교'인 학생이 1.14배로 저체중이 많음을 알 수 있었다.

셋째, CART 모형에 의한 의사결정 나무 분석에서는 성별, 종교, 지역, 척추측만증, 신생아때의 체중을 변인으로 하는 분류가지가 형성됨으로서, 신생아 체중을 제외한 나머지 요인은 일반통계와 유사한 결과를 보였다. 그러나, 각 잎마디의 정보를 통해 좀 더 자세한 저체중 청소년 분류특성을 알 수 있었다. 즉, 남학생 중에서, 종교가 없고, 어머니 교육이 전문대졸이하이며, 강남에 살고, 태어날 때 과체중이 아니었고, 특히 어머니가 국졸이하일 때 80%가 저체중임을 알 수 있었다. 또 여학생은 척추측만증이 없고, 어머니가 대학 또는 대학원을 졸업하였으며, 기독교 또는 천주교를 믿고, 신생아출생시 저체중일 때 저체중의 분포는 42.9%로서 전체 저체중 청소년의 분포인 27.5%보다 많음을 알 수 있었다.

넷째, 빈도상으로 남학생과 강남지역 학교에 저체중 청소년의 분포가 많은 것과는 달리, 연관성 분석에서는 건강문제와 관련이 있는 저체중은 남자보다는 여자에게, 강남보다는 강북학생에게, 건강문제 중에서는 척추측만증을 가진 학생에게 연관규칙이 많음을 알 수 있었다.

그러므로 저체중 청소년의 건강증진을 위한 접근은 척추측만증과 뇨당검출에 대한 신체건강에 대한 중재뿐 아니라, 어머니 교육정도, 출생시 체중도 저체중 청소년과는 서로 관련이 있는 관계이므로, 청소년 시절부터 바른 체중인식에 대한 준비가 다음세대 저체중 신생아 출산, 저체중 청소년을 예방할 수 있을 것이다.

또 저체중 청소년의 건강증진을 위한 대상자 선정은 강북 지역의 학교, 척추측만증 및 여학생을 목표집단으로 선정시 효율적일 것이다.

핵심되는 말 : 저체중, 데이터 마이닝, 건강증진

Ⅰ. 서 론

1. 연구배경 및 의의

청소년 중에서 비만이 많다고 하나 실제로는 저체중 청소년이 더 많다. 1999년도 서울시 학교 보건원이 고등학교 학생 7,342 명을 대상으로 실시한 종합 신체검사 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 가져할 필요성을 알게되었다(이희우와 신선미, 1999).

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%, 지난해 14.8%로 줄어 들은 반면 저체중 어린이는 97년 11.8%에서 98년 12.1%로 늘어나기 시작해 지난해에는 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(한겨레 신문, 2000, 1).

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(이희우와 신선미, 1999). 저체중이 건강에 미치는 영향을 살펴볼 때 쉽게 피로감을 느끼며, 질병에 대한 저항력 감소, 유년, 청년기의 성장 지연, 내분비 장애, 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(허갑범, 1999). 또한 노년기 건강에서도 저체중인 사람들이 정상체중인 사람들에 비해 고혈압으로 인한 사망과 발작의 위험이 높다는 사실이 발견되기도 했으며(로이터통신, 1999), 저체중의 연약한 여성은 다시 저체중 신생아를 출생하고, 이 신생아는 저체중 학생과 성인으로 발달함(Behrman, 1995)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 학교 보건행정가나 보건관계 교사들조차 저체중의 심각성과 잠재적인 건강문제에 대해 문제의식을 가지고 있지 않다.

저체중에 대하여 외국에서 수행된 연구 결과를 보면 무분별한 다이어트와 지나치게 날씬한 몸매를 선호하는 사회풍조, 여성의 흡연, 열량섭취부족과 식이조절로 인한 임신 중 낮은 체중증가를 저체중의 요인으로 분석하고 있으며, 그 외에도 성인남자의 실업률의 증가로 인한 본인 및 배우자의 부담증가가 그 사회에서 저체중의 건강문제가 높아진다는 결과가 분석되어 경제상태가 저체중과 관련 있으며, 여성뿐 아니라 남성의 건강도 저체중아 출생에 영향을 미침을 알 수 있었다 (WHO, 1992).

국내에서 실시된 저체중 연구에서는 신생아 출생지표와 관련된 연구(남철현: 1984, 한국인구보건연구원: 1985, 홍문식: 1994)에서 저체중 출생이 다루어 졌으나, 전국적인 규모의 저체중 출생 상황을 파악하기 위한 기본적인 통계자료도 없는 실정이다.

이희우와 신선미(2001)의 교직원의 건강상태에 관한 연구에서는 저체중 교직원은 스스로 인지하는 건강상태와 불편을 느끼는 신체증상이 많은 것으로 나타났다. 그러나 청소년을 대상으로 저체중의 현황 및 원인에 대한 연구는 아직 수행된 적이 없다.

2. 연구목적

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

본 연구의 구체적인 목적은 다음과 같다.

첫째, 저체중 청소년과 정상체중 청소년의 일반적 특성을 비교하여 파악한다.

둘째, 로지스틱 회귀분석을 통해 저체중 청소년의 변인을 파악한다.

셋째, 저체중 청소년의 특성파악과 예측을 위한 최적의 의사결정나무 모형을 구축하기 위해 로지스틱 회귀모형, CHAID, C4.5, CART 모형의 안정성과 예측력을 평가한다.

넷째, 모형 평가결과, 가장 정확도가 높은 의사결정나무 모형을 이용하여 저체중 청소년의 특성을 파악한다.

다섯째, 연관성 규칙을 통한 저체중 청소년의 관련성을 탐색한다.

Ⅱ. 이론적 배경

1. 저체중

1) 저체중의 현황

청소년 중에서 비만이 많다고 하나 실제로는 저체중 청소년이 더 많다. 1999년도 서울시 학교 보건원이 고등학교 학생 7,342 명을 대상으로 실시한 종합 신체검사 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 가져할 필요성을 알게되었다(이희우와 신선미, 1999).

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%, 지난해 14.8%로 줄어 들은 반면 저체중 어린이는 97년 11.8%에서 98년 12.1%로 늘어나기 시작해 지난해에는 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(한겨레 신문, 2000. 1).

1998년 실시된 국민건강영양조사의 결과를 보면 전반적으로 우리 나라는 정상이 68.7%, 저체중이 5.1%, 과체중이 23.9%, 비만이 2.4%로 비만, 저체중 모두 존재하고 있으며, 이것을 연령별, 성별로 보았을 때, 10-19세 여자의 저체중, 65세 이상 남자의 저체중 문제는 비만만이 건강문제가 아님을 보여주고 있다. 또 조사항목이 질병을 위주로 선택되어서 10세 이하의 경우 신체계측이 이루어지지 못해 정확한 평가가 안되었으나, 다른 자료를 볼 때 과거에 비해 현재 영양상태가 좋아져 체격이 좋아지고 있다는 것과는 대조적으로 아직도 저체중 문제는 있을 것으로 추정된다(조남한 외, 2001).

국내에서 실시된 저체중 연구에서는 신생아 출생지표와 관련된 연구(남철현: 1984, 홍문식: 1994)에서 저체중 출생이 다루어 졌으나, 전국적인 규모의 저체중 출생 상황을 파악하기 위한 기본적인 통계자료도 없는 실정이다.

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 학교 보건행정가나 보건관계 교사들조차 저체중의 심각성과 잠재적인 건강문제에 대해 문제의식을 가지고 있지 않다.

2) 저체중의 문제점

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(이희우와 신선미, 1999). 저체중이 건강에 미치는 영향을 살펴볼 때 쉽게 피로감을 느끼며, 질병에 대한 저항력 감소, 유년, 청년기의 성장 지연, 내분비 장애, 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(허갑범, 1999). 또한 노년기 건강에서도 저체중인 사람들이 정상체중인 사람들에 비해 고혈압으로 인한 사망과 발작의 위험이 높다는 사실이 발견되기도 했으며(로이터통신, 1999), 저체중의 연약한 여성은 다시 저체중 신생아를 출생하고, 이 신생아는 저체중 학생과 성인으로 발달함(Behrman, 1995)을 볼 때 청소년기의 저체중의 관리방안에 대한 중요성은 새롭게 인식해야 할 때이다.

국민건강영양조사의 생화학적 조사의 경우, 우리나라에서 문제가 되고 있는 칼슘, 철분, 비타민 A, 나트륨 등의 영양소에 대해서 생화학적 조사가 모두 이루어진 것은 아니고 철분과 관련된 헤모글로빈만이 조사되었으며, 식생활의 영향을 받

는 지방관련 인자들이 조사되었다. 헤모글로빈의 경우 전반적으로 정상인 경우가 91.4%이나, 젊은 여성이나 어린이의 경우는 정상보다 낮은 경우가 많아서 여기에 대한 대책이 따라야 할 것으로 보인다. 이런 생화학적 조사결과를 볼 때 영양부족과 영양과잉 문제가 우리나라에 공존하고 있다는 것을 잘 보여주고 있으며, 특정계층을 대상으로 영양상태를 살펴보아야 할 필요성을 보여주고 있다. 즉 식품섭취조사 결과를 종합해보면 평균적으로는 별 문제가 없어 보이나, 노인, 어린이, 젊은 여성과 같은 특정 연령, 성별, 집단에서는 문제를 보이는 집단이 있었고(문현경, 2001), 20세 미만 성장발육기의 소아 및 청소년의 경우, 체중 판정 기준은 설정되어 있으나 비만도를 산출하는 표준체중치를 어떠한 자료를 사용하는 가에 따라 비만 및 과체중에 대한 분류가 달라지는 것을 볼 수 있다(박용수, 2001).

저체중과 건강상태와의 관계를 살펴보면, 이희우와 신선미(2001)의 교직원의 건강상태에 관한 연구에서 저체중 교직원은 스스로 인지하는 건강상태가 정상체중이나 과체중 교직원보다 나쁜 것으로 나타났고, 불편을 느끼는 신체증상의 호소도 정상 정상체중에 비해 많은 것으로 나타났다. 그러나 우리나라에서는 저체중의 잠재적 건강문제에 대한 인식은 사회적 공감대를 얻지 못하고 있을뿐만 아니라, 보건의료 및 관련 연구에서도 신생아를 대상으로 극히 제한적으로 실시되고 있는 실정이다.

3) 저체중의 요인

저체중과 관련하여 외국에서 수행된 연구결과를 보면 1999년 Swenne의 연구결과, 저체중은 심전도 결과 QTc간격 지연에 있어 가장 중요한 독립 예측변인임을 밝혀냄으로서 저체중이 심장위험과 상관이 있음을 알 수 있었다.

가나의 도시 어린이에 대한 연구에서 출생시 저체중과 교육받지 못한 어머니가 청소년기의 저체중과 심한 영양결핍의 만연에 중요한 요인중 하나이다(Rikimaru, 1998)라고 하였다. 또한 식사의 공급, 생활 조건 정도가 저체중과 작

은 키를 설명한다(Paricio, 1998)고 하였으며, 류마티스열에 대한 위험도가 더 높은 환자는 나이와의 상관보다는 저신장, 저체중에서 더 많이 관찰되었고, 콩기름을 섭취할 때 그 위험이 감소하였다(Zaman, 1998)고 발표되었다.

건강상태에 대한 역학적 조사에서 볼 때 섬유성낭종 사망률에서 저체중이 독립예측요인이라고 제언하였고(Borowiz, 1996), 신경성 식욕부진증 환자가 저체중에서 체중이 회복된 후에도 대뇌의 회백질 양의 결핍은 계속됨으로서 구조적인 뇌의 변화가 비가역적임을 알 수 있었다(Katzman, 1997). 또 신경성 식욕부진증 환자를 몇 년뒤 다시 추적하여 사망률을 조사해본 결과 예측할 수 있는 변인에는 낮은 혈청알부민 수준과 저체중이 가장 중요한 변수였다(Herzog, 1997)고 하였다.

러시아의 학동기 학생을 대상으로 한 연구에서는 가족수가 많을수록, 아버지의 교육이 낮을수록 학생들의 총 지방섭취가 더 낮은 것으로 나타났으며(Martinchik, 1997), WHO(1998)는 저체중의 요인에 대하여 성인의 무분별한 다이어트와 지나치게 날씬한 몸매를 선호하는 사회풍조, 여성의 흡연, 열량섭취부족과 식이조절로 인한 임신 중 낮은 체중증가를 요인으로 분석하였다. 또 그 외에도 성인남자의 실업률의 증가로 인한 본인 및 배우자의 부담증가가 그 사회에서 저체중의 건강문제가 높아지게 하는 요인이라고 발표하여 경제상태가 저체중과 관련 있으며, 여성뿐 아니라 남성의 건강도 저체중아 출생에 영향을 미침을 알 수 있었다(WHO, 1992).

그러나 우리나라에서는 청소년을 대상으로 저체중의 원인에 대한 연구가 없는 것으로 알려져 있다.

2. 데이터 마이닝

데이터 마이닝은 대용량의 데이터로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화 함으로써 유용한 지식을 추출하는 일련의 과정들이라 정의할 수 있으며, 이는 하나의 분석기법을 의미하는 것이 아니라 여러 기법과 방법들의 적절한 조합으로 이루어진 일련의 과정(process)이라 말 할 수 있다(최종후, 1999). 또한 통계적 관점에서 간략하게 표현한다면 대용량 데이터에 대한 탐색적 데이터 분석이라고 할 수도 있을 것이다.

데이터 마이닝에 대한 또 다른 정의를 살펴보면 의사결정을 위한 응용프로그램으로 데이터 웨어하우스에서 패턴 및 상관관계를 자동적으로 발견해 내는 과정이다. 또 이 과정에서 대형 데이터베이스속에서 의미 있는 정보를 발견하기 위해 고급통계 기법과 기계적 학습을 이용한다. 그러나 데이터 마이닝은 다소 새로운 분야이고 지금도 발전되어 가는 과정에 있기 때문에 그 정의를 한 마디로 요약하기는 매우 어렵다(강현철외, 1999).

데이터 마이닝의 활용범위는 매우 광범위한데, 특히 대량의 데이터를 가지고 있으며 데이터 마이닝 기술의 잠재적 가치가 큰 도소매, 은행, 금융, 보험, 통신분야에서 그 활용가치가 인정되고 있다. 현재 정보통신회사나 보험회사들을 필두로 많은 기업에서 소비자에 대한 대규모의 정보가운데서 마케팅에 활용할 수 있는 유용한 정보의 도출을 위해 데이터 마이닝을 적극적으로 활용하고 있다.

데이터 마이닝과 관련된 타 학문의 연구를 살펴보면, 김수택(2001)은 거래고객의 속성자료를 토대로 신용의 정도를 점수화하는 신용평점(Credit Scoring)을 통해 우량고객 또는 불량고객으로 분류하였으며, 김미영(2000)은 신용평가모형에 이용되는 단일 모형인 로지스틱 회귀모형과 신경망 모형의 결과가 각각 모형에 따라 미래의 신용 위험에 대한 결과가 상이하게 나타나는 단점을 극복하고 신용공여에 대한 승인 여부를 분류하는데 있어서 정확도를 향상시키기 위한 다중 분류앙상블 모형을 제시하였다. 또한 독일 은행의 자료와 국내은행의 자료를 이용하여 단일 모형과 다중 분류앙상블 모형에 대한 결과들을 비교한 결과 신용위험이

불량인 신청고객을 분류하는데 있어서 예측된 확률을 이용한 다중 분류앙상블방법과 예측된 결과를 이용한 다중 분류앙상블 방법이 단일 분류모형에 비해서 더 우수한 분류 결과를 보여주었다.

데이터 마이닝과 관련된 보건의료계 연구를 살펴보면, Chae(2,000)의 건강보험에서의 지식경영을 위한 데이터미이닝 연구에서 건강 결과(Health outcome)를 예측하기 위하여 어떻게 데이터마이닝 알고리즘이 사용되는가를 실험하였다. 특히 이 연구에서는 로지스틱 회귀분석과 데이터 마이닝의 CHIAD(Chi-squared Automatic Interaction Detection)와 C5.0(a variant of 4.5)의 예측력을 비교하였는데, 그 결과 CHAID 알고리즘은 로지스틱 회귀분석보다 더 높은 예측력을 보였고, 세가지 비교에서 C5.0은 가장 낮은 예측력을 가지고 있었다. 또한 연관성규칙은 고혈압과 관련 있는 요인의 특성을 보여주었다.

용왕식(2,000)은 데이터 마이닝 기법을 할용하여 고양시에 거주하는 직장가입자 234,224명에 대한 보험료 부과자료를 가지고 통합의료보험료 부과체계모형을 개발하여 직장의료보험 가입자에게 확대 적용하는 방안을 제시하였다. 이처럼 보건의료분야에서도 데이터 마이닝을 이용한 연구가 시도되고 있으나 학교보건분야에서는 아직까지 데이터 마이닝을 이용한 연구가 실행되지 않았다.

1) 데이터 마이닝 분석의 종류

데이터 마이닝 분석기법 중 연관성 규칙(Association rule)은 데이터로부터 특정한 연관성을 발견하는 것으로 데이터 마이닝의 가장 일반적인 작업이라 할 수 있다. 일반적으로 연관성 분석은 수학과 통계학의 확률과 기대치에 대한 개념을 기반으로 하고 있는데, 이러한 연관성 규칙을 해석하는데 있어 원인과 결과의 직접적인 인과관계로 생각해서는 곤란하고 두 개 또는 그 이상 품목들 사이의 상호 관련성으로 해석해야 한다(강현철외, 1999).

데이터 마이닝 중 군집분석은 주어진 관찰치 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로서 데이터 전체의 구조에 대한

이해를 돕고자 하는 분석방법이다. 대용량 데이터에서 개개의 관찰치를 요약하는 것보다는 전체를 유사한 관찰치의 군집(cluster)으로 구분하여 복잡한 전체보다는 그를 잘 대표하는 군집들을 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 알아낼 수 있다.

데이터 마이닝 중 신경망분석은 뇌 신경생리학(neurophysiology)으로부터 영감을 얻어 시작되었다. 자료분석분야에서 신경망은 복잡한 구조를 가진 자료에서의 예측(prediction)문제를 해결하기 위해서 사용되는 유연한 비선형모형(nonlinear models)의 하나로 분류될 수 있다. 이처럼 신경생리학과의 유사성 때문에 일반적으로 다른 통계적 예측모형에 비해 보다 흥미롭게 받아들여지고 있다. 신경망은 은닉마디(hidden units)라고 불리는 독특한 구성요소에 의해서 일반적인 통계모형과 구별되어진다. 은닉마디는 인간의 신경세포를 모형화한 것으로써, 각 은닉마디는 입력변수들의 결합(combination)을 수신하여 목표변수에 전달한다(강은철, 1999). 다른 데이터 마이닝 분석방법과 신경망 분석을 비교해 볼 때, 회귀분석이나 의사결정나무는 분석의 결과를 비교적 쉽게 해석하고 이로부터 유용한 정보를 얻을 수 있다. 반면에 신경망은 매우 유연하기는 하지만 결과를 해석하는 것이 매우 어렵다. 즉, 로지스틱 회귀분석이 제공하는 것과 같은 계수들에 대한 간편한 해석이 불가능하여, 어떤 입력변수가 중요한지 또는 그것들이 어떻게 상호 작용하는지를 결정하기가 어렵다(강현철, 1999).

의사결정나무 분석은 알고자 하는 어떤 규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법으로 예측의 과정이 나무구조에 의한 추론규칙에 의해 표현되는 방법이다. 의사결정 나무구조는 마디라고 불리는 구성요소들로 이루어져 있고 목표변수인 뿌리마디로부터 끝마디에 이를 때까지 자식마디를 계속적으로 형성해 나감으로써 완성된다. 이처럼 의사결정나무형성을 위한 다양한 기준들이 제안되어 있는데 이는 하나의 부모마디로부터 자식마디들이 형성될 때 예측변수의 선택과 병합이 이루어지는 기준인 분리기준, 정지규칙, 가지치기 방법으로서 이들을 어떻게 결합하느냐에 따라 서로 다른 의사결정나무 형성 방법들이 만들어진다. 데이터 마이닝에서 의사결정나무는 탐색과 모형화의 특성을 지니며, 사전에 이상치를 검색하거나 분석에 필요한 변수를 찾아내고 분석모형에 포함되어야 할

교호효과를 찾아내는데 사용될 수 있고, 그 자체가 분류 또는 예측모형으로 사용될 수 있다(최종후, 1999). 주로 이용되는 의사결정 나무 모형의 알고리즘과 연관성 규칙에 대한 이론을 좀 더 자세히 살펴보면 다음과 같다.

2) 의사결정나무 분석의 알고리즘

(1) CHAID (Chi-squared Automatic Interaction Detection)

1975년 J.A Hadrian이 처음 만들어서 발표한 알고리즘이다. 이 알고리즘은 1963년에 J.A Morgan과 N.A Souquist에 의해 서술된 초기 자동 교호감지 장치인 AID로부터 내려온다. 자동교호감지라는 문구가 암시하듯이 CHAID는 변수간의 통계적 관계를 알아내기 위하여 사용된 것이다. 변수들간의 상관관계를 이용하여 의사결정나무를 만들어야 하므로 이 방법은 의사결정나무에서 분류를 위한 도구로 사용되게 되었다.

CHAID와 C4.5, CART 알고리즘과의 가장 큰 차이는 과다 적합하고 나서 가지치기를 하지 않고 과다 적합이 일어나기 전에 나무를 키워 가는 것을 중지한다는 점이다. 또 다른 차이는 CHAID는 범주형 변수에 국한되어 사용될 수 있다는 점이다. 연속형 변수는 범위로 나누어져서 상.중.하 등의 범주형 계급으로 대체되어야 한다.

CHAID에서는 분류에 있어서 통계적으로 유의한 차이를 주는 더 이상의 분할이 없을 때까지 나무를 계속 키워 나가며, 정확한 유의수준은 나무의 크기를 결정하는 값으로 사용된다. 현재 경향은 의미가 있는 곳 까지 분할을 계속하고나서 가지치기를 포함하는 방법을 연구하는 쪽으로 움직여가고 있으나, 어떤 연구자는 아직도 CHAID로 접근하는 것을 선호하고 있다(최국렬외, 2000).

(2) C4.5

C4.5는 호주 연구원, J Ross Quinlan에 의해 수년간 발전시키고 수정되어 온 의사결정나무 알고리즘이다. 초기 버전인 ID3(Iterative Dichotomize 3)는 1986년에 나왔는데 기계학습분야에 많은 영향을 주어왔고 몇몇 상업용 제품에 사용되고 있다.

C4.5의 선구자 격인 ID3는 잠재적인 분할자들을 비교하기 이해 엔트로피 지수인 $-\sum_{i=1}^{r} P(i)\log(P(i))$를 분리기준으로 사용한다. 분할평가 기준으로서 엔트로피 지수는 가지가 많이 형성된 의사결정나무인 경우에 값이 커지게 된다(조대현외, 2000).

(3) CART(Classification and Regression Trees)

CART 알고리즘은 의사결정나무를 만드는데 가장 많이 사용되는 방법 중의 하나다. 1984년 L. Breiman과 그의 동료들에 의해 발표된 이 방법은 기계학습(machine-learning)실험의 산물이라 할 수 있으며, 판매되고 있는 규칙유도 도구의 많은 것들이 CART를 약간 변형시킨 후 새롭고 거창한 이름을 만들어 명명한 것이라 할 수 있다.

CART 알고리즘은 잠재적인 분할자를 평가하는데 지니지수를 사용한다. 일반적으로 불순도의 높은 수치는 관측치의 집합 계층이 고른 분포를 포함한다는 것인 반면 낮은 수치는 한 계급이 구성원을 지배하고 있다는 것을 의미한다. 가장 좋은 분할은 관측치 집합의 불순도를 크게 감소시키는 것이다.

초기분할은 두 마디를 가지게 되는 데 이들 각각을 뿌리마디에서와 같은 방법으로 분할하게 된다. 한번 더 또 다른 분할자를 찾아내기 위하여 최초 분할자를 제외한 나머지 입력변수 모두를 조사한다. 만약 변수가 하나의 값을 취하면 더 이상 분할에 사용될 수 없으므로 이는 배제된다. 주어진 마디에서 불순도를 현저

히 줄여주는 분할이 발견될 수 없을 때 이를 잎마디(leaf-node)라고 부른다. 결국 잎마디만 남게 되고 여기서 의사결정나무는 완성되게 된다. 그러나 완성된 나무라고 해서 일반적으로 새로운 관측치의 집합을 분류하는데 최적의 역할을 한다고 볼 수 없다.

나무를 만들어가는 과정의 마지막에는 분석용 세트의 모든 관측치가 완성된 의사결정나무의 잎 중 하나에 할당되게 되며, 각 잎은 하나의 계급과 오류율을 할당 받게 된다(최국렬외, 2000).

3) 모형 평가 방법

모형구축과정에 있어서 정량적으로 유의성이 확인된 자료라 할 지라도, 과연 잘못된 것이 아닌지를 판단해야 하는 과정이 있어야 한다(Guoqiang, 1999). 그러므로 많은 연구자는 모형구축을 함에 있어 최적의 모형구축을 위한 평가를 몇가지 방식으로 하고 있다(Hongkyufo, 1996). 그 중 흔히 쉽게 사용할 수 있는 방법으로는 모형을 만드는데 사용했던 자료들을 다시 사용하는 방법으로 재투입법(Resubstitution Method)을 이용한 명목오류율(Apparent Error Rate)이 널리 이용되고 있으나 이 방법은 Unbias한 결과를 제시하지 못한다. 따라서 이러한 문제점을 보완할 수 있는 방법으로 교차타당성 방법(Cross-Validation Method), 잭 나이프 방법(Jackknife Method), 붓스트랩(Bootstrap Method)을 이용한 오류율 추정방법 등이 사용되고 있다(유지혜, 1999).

교차타당성 입증방법은 일반적으로 전체표본을 2개 집단으로 무작위로 나눈다. 이러한 집단들 중의 하나를 분석표본(analysis sample, train sample)이라고 하며 이 방법을 분할표본(split-sample)방법이라고도 부른다(여운승, 2000).

표본을 분석표본과 예비표본(validation sample)으로 나누는 일정한 기준이 있는 것은 아니다(Kun, 1999). 가장 흔히 이용되는 방법은 전체 표본에서 응답자의 반을 분석표본에 배정하고 나머지 반을 예비표본에 배정하는 것이다. 그러나 확고한 기준이 설정되어 있지는 않으며, 조사자에 따라서는 분석표본과 예비표본의

비율을 60대 40, 또는 75대 25로 유지하는 것이 바람직하다고 생각하기도 한다(최종후, 2000).

일반적으로 재투입법(Resubstitution method)은 분류규칙을 만드는데 사용된 자료와 분류규칙의 평가자료가 같기 때문에 정확도는 과다하게, 오류율은 과소하게 잡히기 마련이다. 이러한 문제점을 보와하기 위해 스톤(Stone, M. 1974)이나 가이서(Geisser, S. 1975)등에 의해 활발하게 연구되어온 방법이 교차 타당성 방법이다(여운승, 2000).

4) 소수 설명변수의 분류 모형

관심의 대상이 극소수인 상황에서 기존의 나무모형을 이용하여 우수한 분류모형을 도출하기는 무척 힘든 것이다. 이와같이 데이터 마이닝을 실제로 활용하는데 있어 관심 수준 구성비가 극소수인 경우 적용할 수 있는 단측 순수성(one-side purity)과 PRIM(Patient Rule Induction Method)이 있다(김용대외, 2001).

단측 순수성의 경우 분할단계 초기에 목표변수 수준 중 다수를 분리하는 분할이 연속적으로 일어난다. 이 분할 조건은 마치 다차원의 자료로부터 순수성이 높은 즉, 단일수준으로 이루어져 있는 초다면체(hyper cube)를 분리하는 분할이 연속적으로 일어나는 것을 말한다. 이와 같은 과정은 전체 자료 분포에서 발굴된 조건에 의해 껍질을 벗기는 것에 비유할 수 있어 Friedman과 Fisher(1997)는 이를 껍질벗기기 과정(Peeling process)이라 하였다.

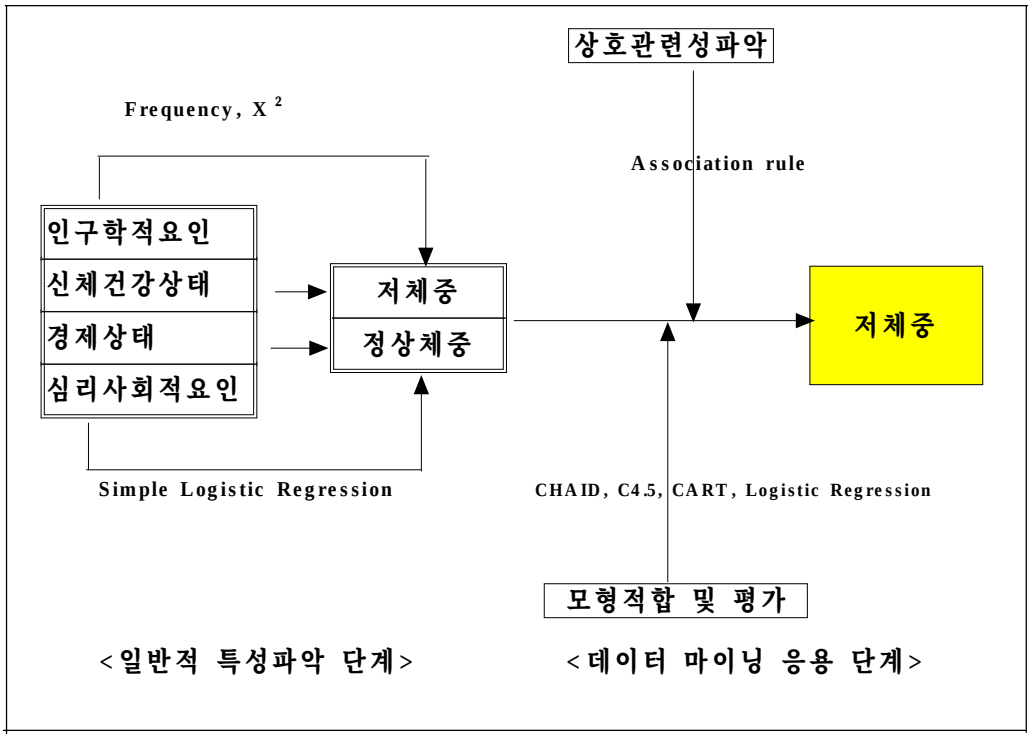
Friedman과 Fisher의 PRIM은 단측 순수성과는 달리 사용자가 입력한 비율만큼 각 변수의 극단 부분을 제거해 가면서 목표변수의 수준을 탐색해 간다. 결국 단측 순수성의 껍질벗기기 과정과 같은 탐색적 과정을 통해 관심대상의 수준을 설명하는 분할 조건(segmentation rule)을 발굴 할 수 있는 것이다(김용대외, 2001). 본 연구에서도 저체중에 영향미치는 관심있는 소수 설명변수의 수준을 탐색해 나가는 방법으로 2단계 분석절차를 걸쳐 다수의 극단 부분에 해당하는 체중과 키를 투입변수에 제거한 후 의사결정나무 모형을 구축하였다.

Ⅲ. 연 구 방 법

1. 연구 설계

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

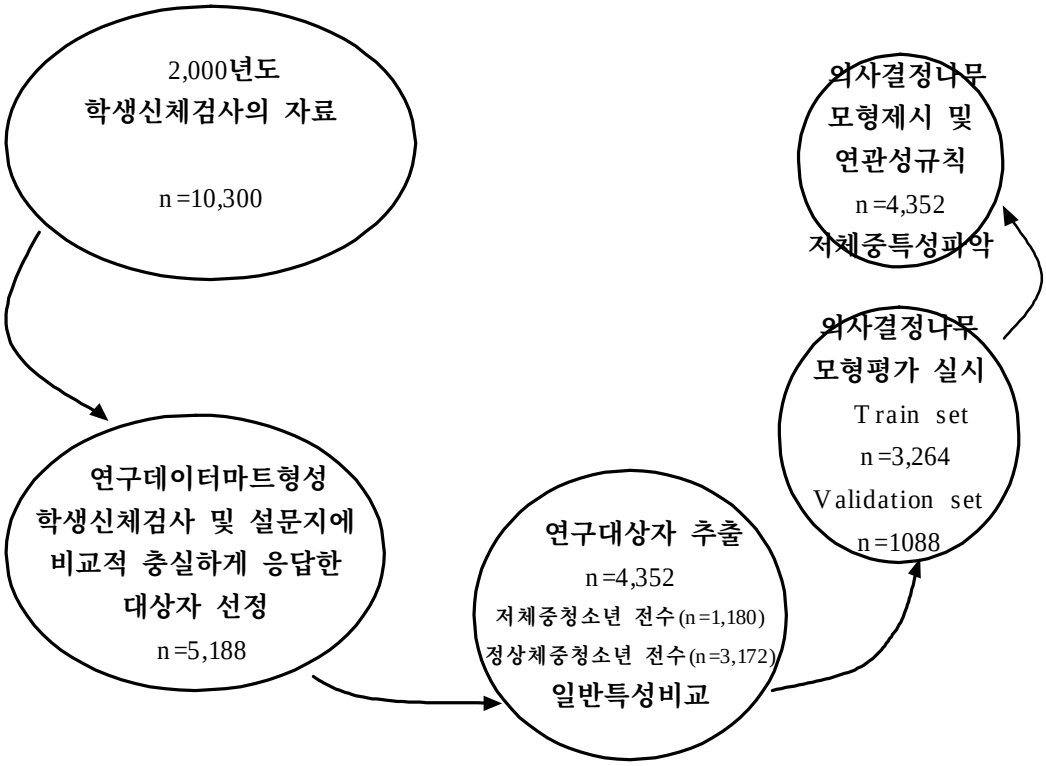
연구로서 연구모형은 다음과 같다(참조 그림 1).



<그림 1> 연구모형

2. 연구 대상

본 연구에서는 24개 고등학교에서 신체검사를 받은 1학년 학생 10,300명의 자료중, 인구,심리,사회학적 특성을 묻는 설문지를 비교적 성실하게 응답한 14개교 고등학교 1학년 학생 5,188명을 데이터 마트로 구축한 후 BMI 18.5미만인 저체중 청소년 전수 1,180명과 BMI 18.5이상 23미만의 정상체중 청소년 전수 3,172명, 총 4,352명을 연구대상으로 선정하였다. 그후 분석용(Train set)으로 3,264명(75%), 평가용 데이터(Validation set)로 1,088명(25%)를 분할 후 모형의 안정성과 예측력을 평가하였으며 그중 가장 설명력이 좋은 모형을 이용하여 최적의 저체중 분류 모형을 제시하였다. 또 연관성 규칙을 이용하여 저체중 청소년의 관련성을 탐색하였다.(참조 그림 2).



<그림 2> 연구 대상자 선정

3. 자료 수집 절차 및 분석항목의 선정

2,000년도 서울시 24개 고등학교에서 신체검사를 받은 1학년 학생 10,300명을 대상으로 조사된 신체검사 결과와 설문응답의 기존 자료 중 인구,심리,사회학적 특성을 묻는 설문지에 비교적 성실하게 응답한 14개교 고등학교 1학년 학생 5,188명의 자료를 데이터 마트로 구축한 후, 본 연구의 목적에 맞게 자료변환을 거쳐 분석항목을 선정하였으며, 분석항목은 인구학적 요인 4문항, 신체건강상태 17문항, 경제적 요인 2문항, 심리 사회학적 요인 38문항, 총 61문항으로 구성되었다.

1) 연구 대상자 데이터 마트 구축

데이터 마트란 특별한 영역에 중점을 두어 만든 데이터 웨어하우스(최종후, 1999)로, 본 연구를 위한 데이터 마트는 서울시 24개 고등학교에서 신체검사를 받은 학생의 원시데이터를 연구목적에 맞게 변환함으로써 구축되었다.

(1) 원시데이터의 구성

원시데이터는(source data)는 신체검사를 수검 받은 대상자의 체격검사 데이터, 혈압 및 체질검사 데이터, 소변검사 데이터, 흉부 x선 촬영 데이터, 심전도검사 데이터, 진찰 및 상담 데이터, 결과판정 데이터, 종합소견 및 조치상황 데이터, 판정의사 데이터, 인구학적 설문, 경제상태 설문, 사회심리학적 설문 데이터로 이루어졌다

(2) 변환(Transformation)

연구목적에 맞게 아래와 같은 방법으로 원시자료를 변환함으로서 연구데이터

마트를 생성하였다.

① 필드 값 변환

원시데이터 테이블의 특정 필드값을 정해진 변환식에 의해 대체(replace), 정정(update), 순위계산(scoring), default 값 부여 등을 행하였다. 예를 들어 폐결핵의 정상을 '0', 비정상을 '1'로 대체하는 등이 이에 속한다.

② 필드속성 변환

필드의 속성이 변경되어 목표테이블의 필드를 형성하는 프로세스로서 속성의 재정의(redefine), 변경(reformat), 필드명 변경(rename)등이다. 예를 들면 혈색소의 형태가 연속숫자인 것을 4등급으로 나누어 변경하는 것을 말한다.

③ 필드생성 변환

이 프로세스는 목표테이블에 새로운 필드를 생성하여 추가하는 프로세스로, 그 예로 BMI필드를 생성하여 추가하기 위해서 신장(m)의 제곱값을 체중으로 나누어 레코드별로 추가하였다.

2) 연구데이터 구성

본 연구데이터는 원시데이터 24개교 고등학교 1학년 학생 10,300명의 자료로부터 인구, 심리, 사회학적 요인을 알아보기 위한 설문지까지 비교적 충실히 응답한 14개교 고등학교 1학년 학생 5,188명을 대상으로 연구 데이터 마트를 만들었다. 그후 자료 변환 및 파생변수의 생성(Modification)을 거친 후 BMI 18.5미만인 저

체중 청소년 전수 1,180명과 BMI 18.5이상 23미만인 정상체중 청소년 전수 3,172명, 총 4,352명을 연구 대상자로 추출하였다.

3) 분석항목의 선정

저체중 청소년의 일반적 특성파악을 위해 사용된 기술통계량의 변수와 저체중 분류기준을 살펴보면 다음과 같다. 또 기술통계량(descritive statistics)에서 정상체중 청소년과 차이가 있는 변수, 외국의 선행연구의 결과 나타난 저체중 관련 주요 변수를 선별하여 데이터 마이닝의 모형 평가과정에 투입하였으며 이때 사용된 변수는 모형 선택방법에서 설명하였다(참조 표11).

(1) 저체중과 정상체중

본 연구에서는 국제 BMI 계산공식〔체중÷키(m)^2〕에 따라 각대상자의 BMI를 산출하였고, 대한비만학회기준에 의해 BMI 18.5미만은 저체중, 18.5이상 23미만은 정상체중으로 분류하였다(참조 표 1).

<표 1> BMI별 체중분류기준

BMI	대한비만학회 국제비만가 전문가 단체	세계보건기구(WHO)
18.5미만	저체중	저체중
18.5 23미만	정상	정상
23 25미만	과체중	정상
25 30미만	비만	과체중
30이상	비만	비만

(2) 인구학적 요인

인구학적 요인에는 성별, 부모교육, 종교에 대해 알아보았으며, 성별과 종교는 원시자료의 응답형태를 그대로 이용하였고, 부모교육은 원시자료에서는 무학, 국졸, 중졸, 고졸, 전문대졸, 대졸, 대학원졸의 7등급이었으나, 응답분포가 적은 무학과 국졸은 합하였고, 중졸과 고졸, 역시 응답분포가 적은 대학원졸은 대졸과 합하여 국졸이하, 고졸이하, 전문대졸, 대졸이상의 4등급으로 재분류하였다.

(3) 신체건강상태

저체중 청소년의 신체적 건강상태의 특징을 알아 보기 위해 객관적인 검사결과데이터 확보가 가능하고, 일반적으로 널리 행해지는 병리이학적 검사결과를 분석하였다. 또 지난 1년간 경험한 건강문제에 대한 대상자의 주관적 응답을 포함시켰으며, 그 개별항목은 다음과 같다.

① 요당

요당은 원시자료에서는 negative, trace(±), one positive(+), two positve(++), three positve(+++)로 측정되었으나, 본 연구에서는 음성(negative), 양성(+, ++, +++)으로 명목화하여 재분류하였다. 그 이유는 본 연구의 원시데이터는 일차 스크리닝 검사의 결과로서 양성으로 판명된 대상자는 그 정도에 상관없이 다시 정확한 재검사를 통하여 확진되어야 하는 특성 때문이다.

② 요단백

요단백도 요당과 마찬가지로 원시자료에서는 negative, trace(±), one

positive(+), two positve(++), three positve(+++)로 측정되었으나, 본 연구에서는 정확한 재검사를 통하여 확진되어야 하는 특성으로 인해 음성(negative), 양성(+, ++, +++)으로 명목화하여 재분류하였다.

③ 혈색소

혈색소의 원시자료는 연속변수로 측정되었으나, 본 연구에서는 임상 의학 기준에 따라 12g/dl 이상인 경우는 정상, 미만은 부족한 것으로 명목화 하였다.

④ 식전혈당

식전혈당의 원시자료는 연속변수로 측정되었으나, 본 연구에서는 임상 의학기준에 의해 126mg%이상은 고혈당, 미만인 경우 정상으로 범주화하였다.

⑤ 총콜레스테롤

총콜레스테롤도 혈색소, 식전혈당과 마찬가지로 원시자료에서는 연속변수로 측정되었으나, 본 연구에서는 임상의학 기준에 따라 250mg%을 기준으로 이상인 경우 고콜레스테롤증, 미만은 정상으로 분류하였다.

⑥ 폐결핵의증

폐결핵은 x-ray 검사상 밝혀진 것으로 활동성 결핵의 확진을 위해서는 세균검사등의 추후검사가 필요하나, 본 연구에서는 원시자료 형태 그대로 폐결핵 의증이 '없다', '있다'의 명목형 그대로 사용하였다.

⑦ 척추이상

척추이상 역시 x-ay 검사상 밝혀진 것으로 척추이상의 종류 및 정도에 대해서는 직접촬영등의 추후검진이 필요하나, 본 연구에서는 원시자료 형태 그대로 '없다', '있다'의 명목형 그대로 선정하였다.

⑧ 지난 1년간의 건강문제

저체중 청소년의 지난 1년간의 건강문제를 파악하기 위해 원시자료에서는 10가지 항목의 질환들(감기독감, 두통. 편두통, 구토.설사, 알레르기, 코피, 빈혈, 우울, 자살충동, 자살시도, 기타)에 중복 응답하도록 하였으며, 본 연구에서는 위 10가지 항목의 응답을 각각 '있다'와 '없다'로 명목화하여 분석하였다.

(4) 경제적 요인

저체중 청소년과 최근의 국가, 사회경제상태와의 관련여부를 파악하기 위해 조사한 것으로 원시자료 형태 그대로 다음과 같이 분류하였다.

① IMF초기파동과 현재 가계수입에 대한 질문

IMF 파동을 전후한 현재 가계수입과 저체중의 영향을 살펴보기위한 질문으로 본 연구의 원시자료에서는 '학생의 가계수입은 IMF파동시 보다 어떻게 변화했나요?'라는 질문을 했고, 대상자는 더나빠졌다, 똑같다, 더 나아졌다의 세가지로 응답하도록 하였다. 이에 본 연구 데이터마트에서는 원시자료 그대로 사용하였다.

② 주위친구들과 비교해본 가족수입

원시자료에서는 주위친구들과 비교해 본 가족수입정도에 대해 질문하였으며, 대상자는 적다, 거의 비슷하다, 많다로 응답하도록 하였고, 본 연구에서는 원시자료 그대로 사용하였다.

(5) 심리사회학적 요인

심리사회학적 요인에는 지역, 학교종류, 신생아 건강상태, 지난 1년간의 건강문제, 담배, 술, 부모.형제질환, 스트레스, 사회적지지, 자아 존중감, 우울감에 대해 조사했다.

① 지역 및 학교종류

대상자의 학교지역에 따른 특성을 알아보기 위하여, 강남과 강북으로 나누어 분류하였으며, 학교종류별 특성을 알아보기 위하여, 인문고와 비인문고로 분류하여 알아보았다.

② 태어났을 때의 건강상태

원시자료에서는 태어났을 때의 건강상태에 대해 질문하였고, 이에 대상자는 약했다고 한다, 정상이었다, 몸무게가 많이 나갔다로 응답하였다. 본 연구에서는 원시자료를 그대로 이용하였다.

③ 담 배

학생들의 흡연에 대한 솔직한 응답을 유도하기 위해 원시자료에서는 지난 한 달간 흡연습관[1]에 대해 조사하였으며, 대상자는 안피웠다, 피웠다, 매일 피웠다로 응답하도록 하였다. 본 연구에서도 원시자료와 똑 같이 사용하였다.

④ 술

학생들의 음주와 저체중과의 관계를 알기 위해 담배와 마찬가지로 지난 한달간 음주경험을 물었으며, 대상자는 안 마셨다, 가끔 마셨다, 꽤 자주 마셨다로 응답하였다. 본 연구에서도 원시자료와 똑같이 명목변수로 사용하였다.

⑤ 부모.형제질환

저체중 청소년의 가족력을 파악하기 위해 원시자료에서는 결핵, 간염, 당뇨병, 중풍, 심장질환, 고혈압, 알콜중독, 기타 의 8가지 항목에 중복 응답하도록 하였으며, 본 연구에서는 각각 항목의 응답을 '있다'와 '없다'로 명목화하여 분석하였다.

⑥ 스트레스

저체중 청소년의 스트레스 정도를 측정하기 위하여 스트레스 정도에 대해 물었으며 원시자료에서는 대단히 많이 느낀다, 많이 느낀다, 중간정도 느낀다, 약간 느낀다, 거의없다 로 응답하였으나, 본 연구에서는 응답분포가 적은 항목을 병합하여 '많다' '보통' '적다'로 순위화하여 처리하였다.

---

1) 현행 학교에서 금지되고 있는 흡연에 대한 질문에서 흡연 학생의 솔직한 응답을 기대하기 어렵다(이희우, 1999)는 선행연구 결과가 있음.

⑦ 가족지지

저체중 청소년의 가족지지 정도를 측정하기 위해 원시자료에서는 가족지지 5문항(부록 참조)에 대해 아니다, 보통이다, 그렇다의 3가지로 응답하도록 하였으나 본 연구에서는 무응답은 결측처리 후 5점-15점 척도로 계산하고 대상자의 남.녀별 사분위수를 달리 계산하여 사분위수 25%까지를 '아주적다', 50%까지를 '적다', 75%까지를 '약간 많다', 76에서 100%까지를 '많다'로 등간화하였다. 또 문항 신뢰도검정에서 cronbach α0.81 이었다.

<표 2> 가족지지의 사분위수 점수

구 분(해석)		남	여
평균점수		10.2	11.0
최저점수-최고점수		5.0 - 15.0	5.0 - 15.0
사분위수 점수	25%까지(아주적다)	5.0 - 8.0	5.0 - 9.0
	50%까지(적다)	8.1 - 10.0	9.1 - 11.0
	75%까지(약간 많다)	10.1 - 12.0	11.1 - 13.0
	100%까지(많다)	12.1 - 15.0	13.1 - 15.0

⑧ 자아 존중감

저체중 청소년의 자아 존중감을 알아보기 위해 원시자료에서는 자아존중감을 알기위한 10문항의 질문(참조 부록)에 대하여 전혀아니다, 아닌편이다, 그런편이다, 확실히 그렇다 라고 응답하였으나, 본 연구에서는 부정적인 질문은 긍정적인 질문으로 치환후 이에 대해 긍정적인 응답은 4점, 부정적인 응답은 1점을 부과하여, 총 10점에서 40점 척도로 점수화 하였다. 그 후 무응답인 경우 결측처리 후 남.녀별 사분위수를 산출하여 25%까지를 '아주적다', 50%까지를 '적다', 75%까지를 '약간

많다', 76에서 100%까지를 '많다'로 정하여 등간화하였다.

또 문항간 신뢰도 검정에서는 cronbach α0.81이었다.

<표 3> 자아존중감의 사분위 점수

구 분		남	여
평균점수		25.2	25.6
최저점수-최고점수		13.0 - 35.0	10.0 - 37.0
사분위수 점수	25%까지 (아주적다)	13.0 - 23.0	10.0 - 23.0
	50%까지 (적다)	23.1 - 25.0	23.1 - 26.0
	75%까지 (약간많다)	25.1 - 28.0	26.1 - 28.5
	100%까지(많다)	28.1 - 35.0	28.6 - 37.0

⑨ 우울감

저체중 청소년의 우울감을 알아 보기 위해 원시자료에서는 우울감을 알기위한 9문항의 질문(참조 부록)에 대하여 전혀아니다, 아닌편이다, 그런편이다, 확실히 그렇다 라고 응답하였으나, 본 연구에서는 무응답인 경우 결측치 처리후 9점에서 36점까지 점수화 하였다. 그후 대상자의 남.녀별 사분위수를 달리 계산하여 25%까지를 '아주적다', 50%까지를 '적다', 75%까지를 '약간 많다', 76에서 100%까지를 '많다'로 정하여 등간화 하였다. 또 문항 신뢰도검사의 cronbach α는 0.79이었다.

<표 4> 우울의 사분위 점수

구 분		남	여
평균점수		20.9	22.2
최저점수-최고점수		9.0 - 35.0	9.0 - 36.0
사분위수 점수	25%까지 (아주적다)	9.0 - 18.0	9.0 - 19.0
	50%까지 (적다)	18.1 - 21.0	19.1 - 22.0
	75%까지 (약간많다)	21.1 - 24.0	22.1 - 25.0
	100%까지 (많다)	24.1 - 35.0	25.1 - 36.0

4. 분석 방법

본 연구는 저체중 청소년의 현황 파악과 특성파악을 위해 다음과 같이 분석하였다.

1) 저체중 청소년과 정상체중 청소년의 일반 특성비교는 빈도와 백분율, X^2 을 통해 알아보았다.
2) 저체중 청소년의 분류 특성을 파악하기 위해 단순 로지스틱 회귀분석을 이용하였다.
3) 데이터 마이닝 기법을 이용한 저체중 분류 모형 구축에 앞서, Cross-validation 평가방법을 이용하여 Logistic 회귀모형과 3가지 의사결정나무 분석 모형 즉 CHAID, C4.5, CART 모형의 안정성과 예측력을 평가하였다.
4) 그 중 가장 정확률이 높은 CART 모형을 선택하여 저체중 청소년 분류모형을 구축후 저체중 청소년의 분류특성을 알아보았다.
5) 또 연관성 규칙을 통해 저체중 청소년의 관련 변인들 사이의 상호관련성을 파악하였다.

1) 분석 모형의 종류

(1) Logistic 회귀분석 모형

종속변수가 이분형이고 여러 가지 독립변수와의 관계를 알아보고자 하는 방법이 로지스틱 회귀분석이다. 그러므로 본 연구에서는 목표범주인 저체중과 독립변수인 인구학적 특성, 신체건강 상태, 경제적, 상황적 데이터와의 관계를 알아보기 위해 모형평가단계에서 설명변수를 모두 투입하는 다중 로지스틱 회귀모형을 구축후 평가하는 과정을 거쳤다.

(2) 의사결정나무분석(Decision tree) 모형

본 연구에서는 의사결정 나무분석을 위한 3가지 알고리즘, CART, C4.5, CHAID방법을 모두 적용하여 평가한 후 가장 최적의 모형으로 평가된 알고리즘을 이용하여 형성된 의사결정나무 분석을 제시하였다.

(3) 연관성 분석(Association rule)

데이터로부터 특정한 연관성을 발견하는 것은 데이터 마이닝의 가장 일반적인 작업이라 할 수 있다(최종후, 1999). 일반적으로 연관성 분석은 수학과 통계학의 확률과 기대치에 대한 개념을 기반으로 하고 있는데, 이러한 연관성 규칙을 해석하는데 있어 원인과 결과의 직접적인 인과관계로 생각해서는 곤란하고 두 개 또는 그 이상 품목들 사이의 상호관련성으로 해석해야 한다. 그러므로 본 연구에서는 여러 독립변수와 종속변수의 자료를 모두 연관성분석단계에 투입하여, 변수들간의 상호관련성을 발견하였다.

2) 모형평가

본 연구에서는 교차 타당성 입증(cross-validation)방법을 이용해, 우선 자료를 분석용과 평가용으로 3 : 1로 나눈 후 5회 반복 무작위 샘플링과 모형 적합 평가를 통해 모형의 안정성과 예측력을 평가하였다.

3) 분석 도구

본 연구에서 일반적 특성은 SAS 8.1을, 데이터 마이닝분석을 위해서는 SAS Enterprise minor 4.0을 이용하였다. 또 모형의 평가에서 민감도와 특이도는 각 모형의 오분류 행렬을 기초로 오분류표 작성후 공식에 의해 수작업으로 계산하여 산출하였고, 의사결정나무 분석의 이익도표는 의사결정나무의 잎마디 분포를 바탕으로, 공식에 의해 수작업으로 계산하여 제시하였다.

Ⅲ. 연 구 결 과

1. 연구 대상자의 일반적 특성

1) 저체중의 분포

본 연구에서는 24개교 남녀 고등학교 10,300명의 신체검사 수검자 자료 중 신체검사와 인구, 심리, 사회학적 특성을 알기 위한 설문지를 비교적 충실하게 응답한 14개 고등학교 1학년 학생 5,188명을 연구 데이터 마트로 만들었고, 그 중 BMI 18.5미만의 저체중 청소년 전수 1,180명과 BMI 18.5이상 23미만인 정상체중 청소년 전수 3,172명 총 4,352명을 연구대상으로 추출하였다.

연구 데이터 마트 5,188명 중 저체중 청소년의 분포를 보면 22.74%이고, 그중 남자는 27.94%, 여자는 21.05%를 차지함으로서, 남학생의 저체중 분포가 여학생보다 더 높음을 알 수 있었다(참조 표 5)

<표 5> 전체 대상자의 BMI등급 및 기준

BMI등급	BMI 기준	남	(%)	여	(%)	계	(%)
저체중	18.5 미만	356	(27.94)	824	(21.05)	1,180	(22.74)
정상	18.5 23미만	699	(54.87)	2,473	(63.18)	3,172	(61.14)
과체중	23.0 25미만	93	(7.30)	382	(9.76)	475	(9.16)
비만1	25.0 30미만	101	(7.93)	202	(5.16)	303	(5.84)
비만2	30.0 이상	25	(1.96)	33	(0.84)	58	(1.12)
계		1,274	(100.0)	3,914	(100.0)	5,188	(100.0)

2) 대상자의 인구학적 특성

본 연구대상자 4,352명의 학생 중 아버지의 교육정도는 저체중 청소년 그룹에서는 고졸이하가 제일 많았고, 정상체중 청소년에서는 고졸이하가 가장 적은 분포를 차지했으나, 통계학적으로는 유의하지 않았다. 어머니의 교육정도를 볼 때, 저체중 청소년군에서는 대졸이 가장 적은 분포를 보였고, 정상체중에서는 대졸이 가장 많았으나, 이는 통계학적으로 유의하였다. 종교를 살펴보면 저체중 청소년군에서는 기독교가 가장 많았고, 정상체중 청소년 군에서는 기독교가 가장 적었으나, 통계학적으로 유의하지 않았다(참조 표 6).

<표 6> 대상자의 인구학적 특성*

변 수	구 분	저 체 중		정 상 체 중		X^2-value	P-value
		실수	(%)	실수	(%)		
성별	남 자	356	(33.74)	699	(66.26)	30.97	< .001
	여 자	824	(24.99)	2,473	(75.01)		
	계	1,180	(27.11)	3,172	(72.89)		
부 교육정도	국졸이하	17	(22.67)	58	(77.33)	0.86	0.830
	고졸이하	324	(25.43)	950	(74.57)		
	전문대졸	8	(20.00)	32	(80.00)		
	대졸이상	174	(25.04)	521	(74.96)		
	계	523	(25.10)	1,561	(74.90)		
모 교육정도	국졸이하	36	(32.73)	74	(67.27)	9.80	0.020
	고졸이하	402	(25.69)	1,163	(74.31)		
	전문대졸	9	(37.50)	15	(62.50)		
	대졸이상	70	(20.29)	275	(79.71)		
	계	517	(25.29)	1,527	(74.71)		
종교	기독교	229	(27.23)	612	(72.77)	3.69	0.296
	불 교	63	(22.26)	220	(77.74)		
	천주교	68	(23.53)	221	(76.47)		
	기 타	187	(24.64)	572	(75.36)		
	계	547	(25.18)	1,625	(74.82)		

* 무응답제외

3) 대상자의 신체건강 특성

저체중 청소년과 정상체중 청소년의 신체적 건강상태 비교에서 볼 때 요당과 척추측만증은 저체중 청소년에게 많았다. 즉 요당이 음성인 경우의 저체중의 비율은 26.94%, 정상체중의 비율은 73.06%인 반면, 양성인 경우의 저체중의 비율은 38.24%, 정상체중의 비율은 61.76%를 차지하였다. X-RAY 검사에서 척추가 정상인 경우 저체중은 26.69%, 정상체중은 73.31%인 반면, 척추측만증인 경우 저체중은 38.32%, 정상체중은 61.68%를 차지하고 있었다.

또 요단백, 혈색소, 식전혈당, 콜레스테롤, 폐결핵, 일년동안의 건강문제경험 중 코피경험, 빈혈, 우울, 자살충동, 자살시도는 두 군간에 통계적 차이가 없었고, 일년동안 감기경험, 알레르기, 기타 경험한 건강문제에서는 정상체중 청소년이 더 많이 경험한 것으로 나타났다(참조 표 7).

<표 7> 대상자의 신체건강특성*

변 수	구 분	저 체 중		정 상 체 중		X^2-value	P-value
		실수	(%)	실수	(%)		
요당	음성	1,154	(26.94)	3,130	(73.06)	4.32	0.037
	양성	26	(38.24)	42	(61.76)		
	계	1,180	(27.11)	3,172	(72.89)		
요단백	음성	1,142	(26.94)	3,097	(73.06)	2.49	0.114
	양성	38	(33.63)	75	(66.37)		
	계	1,180	(27.11)	3172	(72.89)		
혈색소	정상	1,163	(27.12)	3,126	(72.88)	0.00	0.981
	빈혈	17	(26.98)	46	(73.02)		
	계	1,180	(27.11)	3172	(72.89)		
식전 혈당***	정상	1,179	(27.12)	3,169	(72.88)	0.00	0.920
	고혈당	1	(25.00)	3	(75.00)		
	계	1,180	(27.11)	3,172	(72.89)		
콜레스테롤	정상	1,177	(27.20)	315	(72.80)	2.60	0.100
	고콜레스테롤	3	(12.50)	21	(87.50)		
	계	1,180	(27.11)	3,172	(72.89)		
폐결핵***	정상	1,123	(1.12)	3,084	(73.27)	0.00	0.967
	의심	3	(27.27)	8	(72.73)		
	계	1,126	(26.73)	3,092	(73.27)		
척추	정상	1,123	(26.69)	3,084	(73.31)	7.15	0.007
	측만증	41	(38.32)	66	(61.68)		
	계	1,164	(26.98)	3,150	(73.02)		
출생시체중	저체중	86	(28.01)	221	(71.99)	2.04	0.359
	정상	349	(24.96)	1,049	(75.04)		
	과체중	50	(22.73)	170	(77.27)		
	계	485	(25.19)	1,440	(74.81)		

<표 7> 대상자의 신체건강특성<계속>

변 수	구 분		저 체 중		정 상 체 중		X^2-value	P-value
			실수	(%)	실수	(%)		
지년 1년간 건강문제**	감기/독감	무	825	(28.35)	2,085	(71.65)		
		유	355	(24.62)	1,087	(75.38)	6.79	0.009
		계	1,180	(27.11)	3,172	(31.72)		
	두통/편두통	무	912	(28.45)	2,294	(71.55)		
		유	268	(23.45)	875	(76.55)	11.77	0.008
		계	1,180	(27.11)	3,172	(72.89)		
	구토/설사	무	1,040	(27.45)	2,749	(72.55)		
		유	140	(24.96)	421	(75.04)	2.28	0.510
		계	1,180	(27.11)	3,172	(72.89)		
	알레르기	무	1,113	(27.61)	2,918	(72.39)		
		유	67	(20.87)	254	(79.13)	6.83	0.009
		계	1,180	(27.11)	3,172	(72.89)		
	코피	무	1,089	(27.35)	2,892	(72.65)		
		유	91	(24.53)	280	(75.47)	1.37	0.241
		계	1,180	(27.11)	3,172	(72.89)		
	빈혈	무	1,004	(27.23)	2,683	(72.77)		
		유	176	(26.55)	487	(73.45)	0.87	0.830
		계	1,180	(27.11)	3,172	(72.89)		
	우울	무	1,092	(27.53)	2,874	(72.47)		
		유	88	(22.80)	298	(77.20)	3.99	0.045
		계	1,180	(27.11)	3,172	(72.89)		
	자살충동	무	1,152	(27.32)	3,064	(72.68)		
		유	28	(20.59)	108	(79.41)	3.02	0.080
		계	1,180	(27.11)	3,172	(72.89)		
	자살시도	무	1,173	(27.10)	3,155	(72.90)	0.05	0.820
		유	7	(29.17)	17	(70.83)		
		계	1,180	(27.11)	3,172	(72.89)		
	기타	무	1,149	(27.40)	3,045	(72.60)	4.65	0.030
		유	31	(19.62)	127	(80.38)		
		계	1,180	(27.11)	3,172	(72.89)		

* 무응답 제외
** 중복응답
*** Fisher 정확도 검정 결과

4) 대상자의 경제적 특성

IMF파동시와 현재 가계 수입비교에 대한 질문에서 저체중 청소년은 '더 좋아졌다'고 응답한 비율이 가장 많은 반면, 정상체중 청소년은 '똑같다'고 응답한 비율이 가장 많음을 알 수 있었다. 또 친구들과 비교해본 가족수입에서는 저체중 청소년은 '더 많다'에 응답한 비율이 가장 높은 반면, 정상체중 청소년은 '비슷하다'고 응답한 비율이 가장 높음을 알 수 있다. 그러나 두 질문 모두 통계학적으로 유의하지는 않았다(참조 표 8).

<표 8> 대상자의 경제적 특성*

변 수	구 분	저 체 중		정 상 체 중		X^2-value	P-value
		실수	(%)	실수	(%)		
IMF 파동시와 현재가계수입 비교	더 나빠졌다	195	(26.21)	549	(73.79)	2.72	0.255
	똑같다	294	(24.06)	928	(75.94)		
	더 나아졌다	53	(29.12)	129	(70.88)		
	계	542	(25.23)	1,606	(74.77)		
주위 친구들과 비교해본 가족수입	더 나쁘다	130	(25.59)	378	(74.41)	0.12	0.938
	거의비슷하다	403	(25.22)	1,195	(74.78)		
	많다	11	(27.50)	29	(72.50)		
	계	544	(25.35)	1,602	(74.65)		

* 무응답 제외

5) 심리사회학적 특성

학교지역별 비교에서 볼 때 강남지역의 학교에 저체중 청소년의 분포가 많음을 알 수 있었고, 인문고와 비인문고와의 비교, 담배, 술, 부모형제의 결핵, 간염, 고혈압, 알콜중독에는 두 군간에 서로 차이가 없었다. 또 부모.형제질환 중 당뇨는 정상체중 청소년군에서 더 많음을 알 수 있었다. 즉 당뇨가 없을 때 저체중 청소년의 비율은 27.40%이고 정상체중의 비율은 72.60%인 반면, 당뇨가 있을 경우의 저체중 비율은 17.83%, 정상체중의 비율은 82.17%로서 정상체중 청소년에게 당뇨의 가족력이 더 많음을 알 수 있었다(참조 표 9).

또 스트레스와 우울감, 가족지지, 자아존중감 모두 통계학적으로 유의하지 않았다.

<표 9> 대상자의 심리 사회학적 특성*

변 수	구 분	저 체 중		정 상 체 중		X^2-value	P-value
		실수	(%)	실수	(%)		
지 역	강북	629	(25.87)	1,802	(74.13)	4.28	0.030
	강남	551	(28.68)	1,370	(71.32)		
	계	1,180	(27.11)	3,172	(72.89)		
학교종류	비인문고	476	(27.09)	1,281	(72.91)	0.00	0.970
	인문고	704	(27.13)	1,891	(72.87)		
	계	1,180	(27.11)	3,172	(72.89)		
담배	안피웠다	413	(24.79)	1,253	(75.21)	5.55	0.135
	가끔 피웠다	33	(30.28)	76	(69.72)		
	매일 피웠다	40	(28.99)	98	(71.01)		
	계	487	(25.44)	1,427	(74.56)		
술	안마셨다	380	(25.80)	1,093	(74.20)	1.14	0.766
	가끔 마셨다	96	(27.12)	258	(72.88)		
	꽤 자주 마셨다	7	(21.88)	25	(78.13)		
	계	483	(26.01)	1,376	(73.99)		
부모.형제 질환**	결핵 없다	1,172	(27.10)	3,153	(72.90)	0.08	0.768
	결핵 있다	8	(29.63)	19	(70.37)		
	결핵 계	1,180	(27.11)	3,172	(72.89)		
	간염 없다	1,166	(27.18)	3,124	(72.82)	0.91	0.632
	간염 있다	14	(22.95)	47	(77.05)		
	간염 계	1,180	(27.11)	3,172	(72.89)		
	당뇨병 없다	1,157	(27.40)	3,066	(72.60)	5.79	0.016
	당뇨병 있다	23	(17.83)	106	(82.17)		
	당뇨병 계	1,180	(27.11)	3,172	(72.89)		
	중풍 없다	1,175	(27.12)	3,157	(72.88)	0.04	0.831
	중풍 있다	5	(25.00)	15	(75.00)		
	중풍 계	1,180	(27.11)	3,172	(72.89)		
	심장질환 없다	1,173	(27.17)	3,144	(72.83)	0.90	0.341
	심장질환 있다	7	(20.00)	28	(80.00)		
	심장질환 계	1,180	(27.11)	3,172	(72.89)		
	고혈압 없다	1,141	(27.25)	3,046	(72.75)	1.04	0.305
	고혈압 있다	39	(23.64)	126	(76.36)		
	고혈압 계	1,180	(27.11)	3,172	(72.89)		

<표 9> 대상자의 심리 사회학적 특성<계속>*

변 수	구 분		저 체 중		정 상 체 중		X^2-value	P-value
			실수	(%)	실수	(%)		
부모.형제 질환**	알콜 중독***	없다	1,179	(27.16)	3,162	(72.84)	1.81	0.178
		있다	1	(9.09)	10	(90.91)		
		계	1,180	(27.11)	3,172	(72.89)		
	기타	없다	1,136	(26.10)	3,060	(72.93)	0.51	0.771
		있다	43	(28.67)	107	(71.33)		
		계	1,179	(27.11)	3,167	(72.89)		
스트레스	많이 느낀다		377	(24.54)	1,159	(75.46)	2.44	0.294
	중간 정도		140	(28.00)	360	(72.00)		
	거의 없다		29	(24.37)	90	(75.63)		
	계		546	(25.34)	1,609	(74.66)		
가족지지	아주 적다		89	(24.79)	270	(75.21)	3.62	0.304
	적다		147	(27.95)	379	(72.05)		
	약간 많다		131	(22.98)	439	(77.02)		
	많다		169	(25.15)	503	(74.85)		
	계		536	(25.20)	1,591	(74.80)		
지난 한달동안의 우울감	아주 적다		144	(26.92)	391	(73.08)	1.28	0.730
	적다		131	(24.49)	404	(75.51)		
	약간 많다		143	(24.24)	447	(75.76)		
	많다		109	(24.94)	328	(75.06)		
	계		527	(25.15)	1,570	(74.87)		
자아존중감	아주 적다		129	(26.93)	350	(73.07)	3.38	0.335
	적다		145	(27.51)	382	(72.49)		
	약간 많다		105	(25.18)	312	(74.82)		
	많다		145	(23.20)	480	(76.80)		
	계		524	(25.59)	1,524	(74.41)		

* 무응답 제외
** 중복응답
*** Fisher 정확도 검정 결과

6) 저체중 청소년의 변인분석

저체중과 여러 독립변수와의 관계를 알아보기 위해 단순 로지스틱 회귀분석을 실시해본 결과 통계학적으로도 유의한 변수를 살펴보면, '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 1.14배로 저체중이 많았고, 뇨당은 '없다'에 비해 '있다'가 1.68배로 저체중이 많았다. 척추측만증의 경우도 '없다'에 비해 '있다'가 1.70배로 저체중이 많았다. 또 알러지는 있을 때 보다 없을 때에 오히려 1.44배, 일년동안 경험한 신체건강 특성 중 우울은 '있다'에 비해 '없다'가 1.28배로 저체중이 많았고, 자살충동은 '있다'에 비해 '없다'일때 오히려 1.45배로 저체중이 많은 경향이 있었다(P-value 0.08). 또 자살시도도 '있다'에 비해 '없다'가 1.54배로 저체중이 많았다.

학교지역으로 볼 때 '강북'에 비해 '강남'이 1.15배로 저체중이 더 많았고, 아버지 당뇨가 '있다'에 비해 '없다'가 1.73배로 저체중이 많았다.

그러나, 가족지지, 자아존중감 그리고 한달동안의 우울감에서는 통계학적으로 저체중과 정상체중군간에 차이가 없었다(참조 표 10).

<표 10> 단순 logistic 회귀분석 결과

구분	변 수	기 준	구 분	Parameter Estimate	Odds ratio	Pr>Chisq	95% wald confidece limtis
인구 특성	성별	(여)	남	0.21	1.52	<.001	1.31-1.77
	부(父)교육	(국졸이하)	대졸이상	0.09	1.13	0.470	0.64-2.01
	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 건강 특성	신생아체중	(과체중)	정상	-0.01	1.13	0.880	0.87-1.97
	뇨당	(없다)	있다	0.25	1.68	0.030	0.36-0.97
	척추측만증	(없다)	있다	0.26	1.70	0.000	1.14-2.53
	알러지	(있다)	없다	0.18	1.44	0.000	1.09-1.90
	지난 일년간 우울	(있다)	없다	0.12	1.28	0.040	1.00-1.64
	자살충동	(있다)	없다	0.18	1.45	0.080	0.95-2.20
	자살시도	(있다)	없다	0.21	1.54	0.030	1.03-2.30
심리 사회 특성	학교지역	(강북)	강남	0.07	1.15	0.030	1.00-1.31
	부(父)당뇨	(있다)	없다	0.27	1.73	0.010	1.10-2.74
	가족지지	(많다)	아주적다	-0.02	0.98	0.830	0.73-1.31
	자아존중감	(많다)	아주 적다	0.06	1.22	0.460	0.92-1.60
	지난한달간 우울감	(많다)	아주 적다	-0.08	0.74	0.390	0.50-1.09

2. 저체중 청소년 분류를 위한 모형 선택 방법

저체중 청소년의 특성탐색과 분류, 예측을 위한 최적의 데이터 마이닝 모형 선택을 위해 본 연구에서는 보건의료계에서 관련 변인을 규명하기 위해 널리 사용되는 Logistic regression 모형과 데이터 마이닝의 의사결정 나무분석의 3가지 모형(CHAID, C4.5, CART)을 비교하여 평가하였다.

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 변인으로 제시된 변수이었다(참조 표 11).

<표 11> 모형평가에 투입된 변수

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

* 구분점수는 표 2, 3, 4 참조

2) 모형 평가의 기준

실제로 모형 분석을 수행함에 있어서 여러 가지 기준과 규칙을 어떻게 지정하느냐에 따라서 서로 다른 의사결정 나무를 얻게 된다(강현철, 1998).

그러므로 본 연구에서는 모형의 평가에 있어 각각의 모형 모두 사후확률[2]이 0.5 [3]로 동일하다고 가정하였다. 로지스틱회귀모형에서는 입력변수를 다중으로 투입하였고, 의사결정나문분석에서는 각 모형의 가지분리기준을 각 모형 모두 Default로 지정하여 동일한 정지기준을 설정하였다. 그 이유는 대부분의 경우에 디폴트 옵션들은 훌륭하게 작동하기 때문이며(강현철, 1998), 동일한 조건에서 모형을 비교하고 검토하여 가장 적절한 의사결정나무를 얻을 수 있어야 하기 때문이다.

즉 끝마디에 포함될 관측개체의 최소개수는 1, 지정된 분리기준에 의해 부모마디가 자식마디로 분리되기 위해 요구되는 관측개체의 수[4]는 43으로 이 수보다 부족하면 더 이상 분리가 일어나지 않는다. 자식마디가 형성될 때 최대의 분리개수는 이지분리를 기본으로 하였다. 마지막으로 뿌리마디로부터 끝마디까지의 깊이가 7로 이수보다 큰 경우엔 더 이상 분리를 수행하지 않았다.

또 지나치게 많은 마디를 갖는 Tree는 새로운 자료에 적용 시 예측오차가 매우 클 가능성이 있기 때문에 적절한 부나무(Subtree)를 선택하기 위해 Tree 형성 후 최적의 평가 타당도를 갖는 부나무를 형성하였으며, Cross-validation 평가에 의해 오류율을 나타내도록 하였다.

---

2) 각 대상학생이 저체중이 될 확률
3) 일반적으로 분류의 기준으로 삼는 사후확률의 경계는 '1/목표범주의 범주개수'이 보통이다.
4) Data의 전체 관측개체의 수÷100

3) Cross-validation을 이용한 모형 평가방법

모형의 평가결과를 비교하기 위해 각 모형마다 동일한 변수를 투입한 후 교차타당성 방법(Cross-validation test)으로 모형의 안정성과 예측력을 평가하였다.

이를 위해 4,352명의 연구대상자 중 75%를 분석용(Train), 25%를 평가용(Validation)으로 무작위 Sampling을 통해 분할 한 후 3가지 적중률인 정분류율(Acuracy ratio), 민감도(Sensitivity), 특이도(Specificity)를 산출하였다. 또 모형의 안정성을 위해 1회에 그치지 않고, 표본추출과 평가를 5회 반복하여 각 자료에 대하여 평균을 비교하였다.

3가지 적중률의 공식을 살펴보면 다음과 같다(참조 표 12).

<표 12> 정분류율, 민감도, 특이도 공식

$$\text{정분류율} = \frac{(\text{관찰} = 1 \cap \text{예측} = 1)\text{의빈도} + (\text{관찰} = 2 \cap \text{예측} = 2)\text{의빈도}}{\text{전체빈도}}$$

$$\text{민감도} = \frac{(\text{관찰} = 1 \cap \text{예측} = 1)\text{의빈도}}{(\text{관찰} = 1)\text{의빈도}}$$

$$\text{특이도} = \frac{(\text{관찰} = 2 \cap \text{예측} = 2)\text{의빈도}}{(\text{관찰} = 2)\text{의빈도}}$$

각각의 개념을 살표보면, 정분류율(Accuracy ratio)은 저체중을 저체중으로 판별하고 정상체중을 정상체중으로 판별할 확률이고, 민감도(Sensitivity)는 관찰된 전체 저체중 청소년 중에서 제대로 저체중으로 판별할 확률, 특이도(Specificity)는 관찰된 전체 정상체중 청소년 중에서 제대로 정상체중을 판별할 확률이다.

4) 모형 평가 결과 비교

모형 평가 결과 5회에 걸쳐 실시한 Cross-validation 평가에서 구축용과 평가용의 결과가 큰 차이를 보이지 않음으로서, 모형구축과 예측력 평가에 있어 안정적임을 알 수 있었다(참조 표 12, 13, 14, 15, 16).

분석용에서 정분류율이 가장 높은 모형은 CART모형이었고, 민감도가 가장 좋은 모형은 CHAID모형, 특이도가 가장 좋은 모형은 CART모형임을 알 수 있었다(참조 표 15). 또 평가용에서도 CART모형이 가장 정분류율이 높게나왔으며, 분석용과 평가용 모두 가장 정분류율이 낮은 모형은 Logistic regression모형임을 알 수 있었다.

<표 13> 각 모형의 정분류율 평가

모형의 종류	CHAID		C4.5		CART		Logistic regression	
	분석용	평가용	분석용	평가용	분석용	평가용	분석용	평가용
set 1	85.39	86.86	85.20	86.86	85.42	87.13	76.79	78.77
set 2	85.97	84.65	85.60	84.60	86.18	84.65	78.01	75.46
set 3	85.29	87.32	85.20	87.32	85.11	87.41	77.43	77.58
set 4	85.60	84.93	86.37	86.37	86.21	86.24	77.95	75.83
set 5	86.31	85.02	86.15	84.74	86.40	85.11	77.46	76.94
모형의 평균	85.71	85.76	85.70	85.98	85.86	86.11	77.53	76.92

<표 14> 각 모형의 민감도 평가

모형의 종류	CHAID		C4.5		CART		Logistic regression	
	분석용	평가용	분석용	평가용	분석용	평가용	분석용	평가용
set 1	68.58	73.29	68.92	73.63	69.03	73.29	22.07	23.02
set 2	70.32	68.15	71.36	69.43	71.02	68.79	23.79	22.84
set 3	68.92	73.63	67.45	71.92	68.36	73.29	24.10	24.72
set 4	70.27	69.87	69.93	68.21	69.25	68.54	26.72	24.75
set 5	69.37	69.51	69.37	68.52	69.03	68.20	21.89	23.85
모형의 평균	69.49	70.89	69.41	70.34	69.34	70.53	23.71	23.84

<표 15> 각 모형의 특이도 평가

모형의 종류	CHAID		C4.5		CART		Logistic regression	
	분석용	평가용	분석용	평가용	분석용	평가용	분석용	평가용
set 1	91.67	91.83	91.29	91.71	91.54	92.21	97.10	97.16
set 2	91.62	91.34	90.74	90.83	91.66	91.09	95.58	97.52
set 3	91.41	92.34	91.84	92.96	91.37	92.59	97.35	97.45
set 4	91.24	90.71	92.41	91.73	92.46	91.73	99.53	97.51
set 5	92.51	91.06	92.30	91.06	92.76	91.70	97.19	97.20
모형의 평균	91.69	91.46	91.72	91.66	91.96	91.86	97.35	97.37

<표 16> 각 모형의 평균 비교

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

3. CART 알고리즘에 의한 의사결정 나무 모형

Cross-validation결과 모형의 정분률이 가장높은 모형은 CART모형이었다. 이에 본 연구에서는 CART모형에 따른 의사결정나무 모형을 제시하였다.

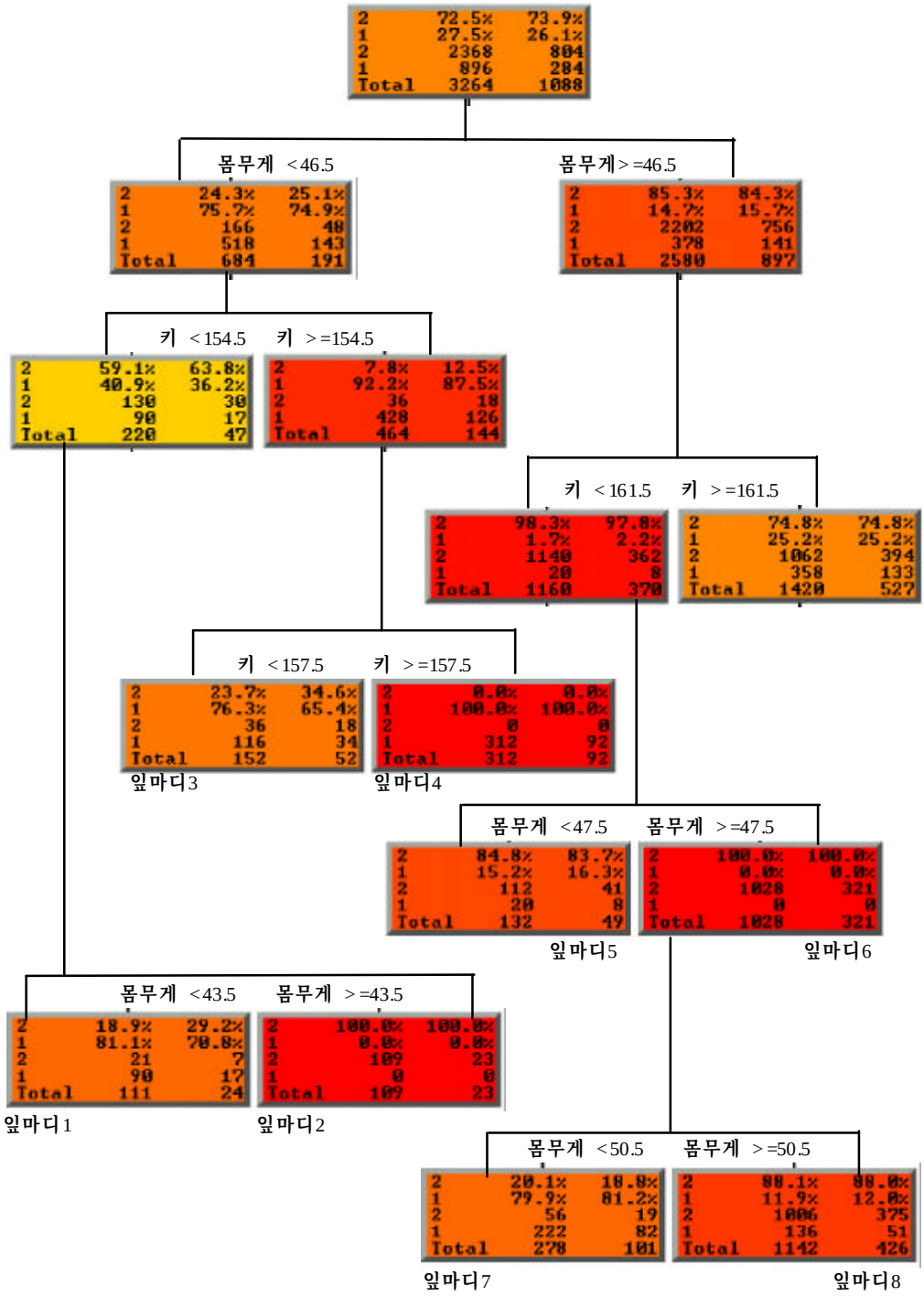
제시된 모형은 모형 평가와 똑 같은 분류기준으로 구축하였다. 즉 투입된 변수, 모형구축의 기준은 모두 모형 평가과정과 동일하며, 모형구축의 효율성을 위해 모형 평가때와 역시 마찬가지로 분석용(Train set)으로 3,264명(75%), 평가용 데이터(Validation set)로 1,088명(25%)를 분할하였다. 그러므로 마디내 왼쪽 값은 분석용 데이터에 대한 결과이며, 오른쪽 값은 평가용 데이터에 대한 결과이다.

그러나 관심의 대상이 극소수인 상황에서 나무모형을 이용하여 우수한 분류모형을 도출하기는 무척 힘든 것이다(김용대외, 2000). 본 연구에서도 그 예외는 아니어서, 저체중 청소년의 가장 많은 분류규칙은 체중이었고, 그다음은 키임을 알 수 있었다(참조 그림3). 그러나 이는 본 연구에서 알아내기 위한 관심의 대상은 아니다. 즉 본 연구의 목적은 저체중에 영향미치는 작지만 중요한 변인이 무엇인가를 밝히는 것이므로, 소수이지만 중요한 관심의 대상을 설명하는 조건을 알아내기 위하여 키와 체중을 변수로 투입한 1단계 분석과 키와 체중을 변수 투입에서 제외한 2단계 분석을 수행하였다(참조 그림 3, 4, 5).

1) 1단계 분석

1단계에서 분석된 의사결정모형에서, 잎마디 1과 잎마디2를 보면 키 154.5cm 미만에서 체중이 43.5kg미만인 경우 저체중은 81.1%이었으며, 체중이 43.5kg이상이면 저체중은 0%이었다.

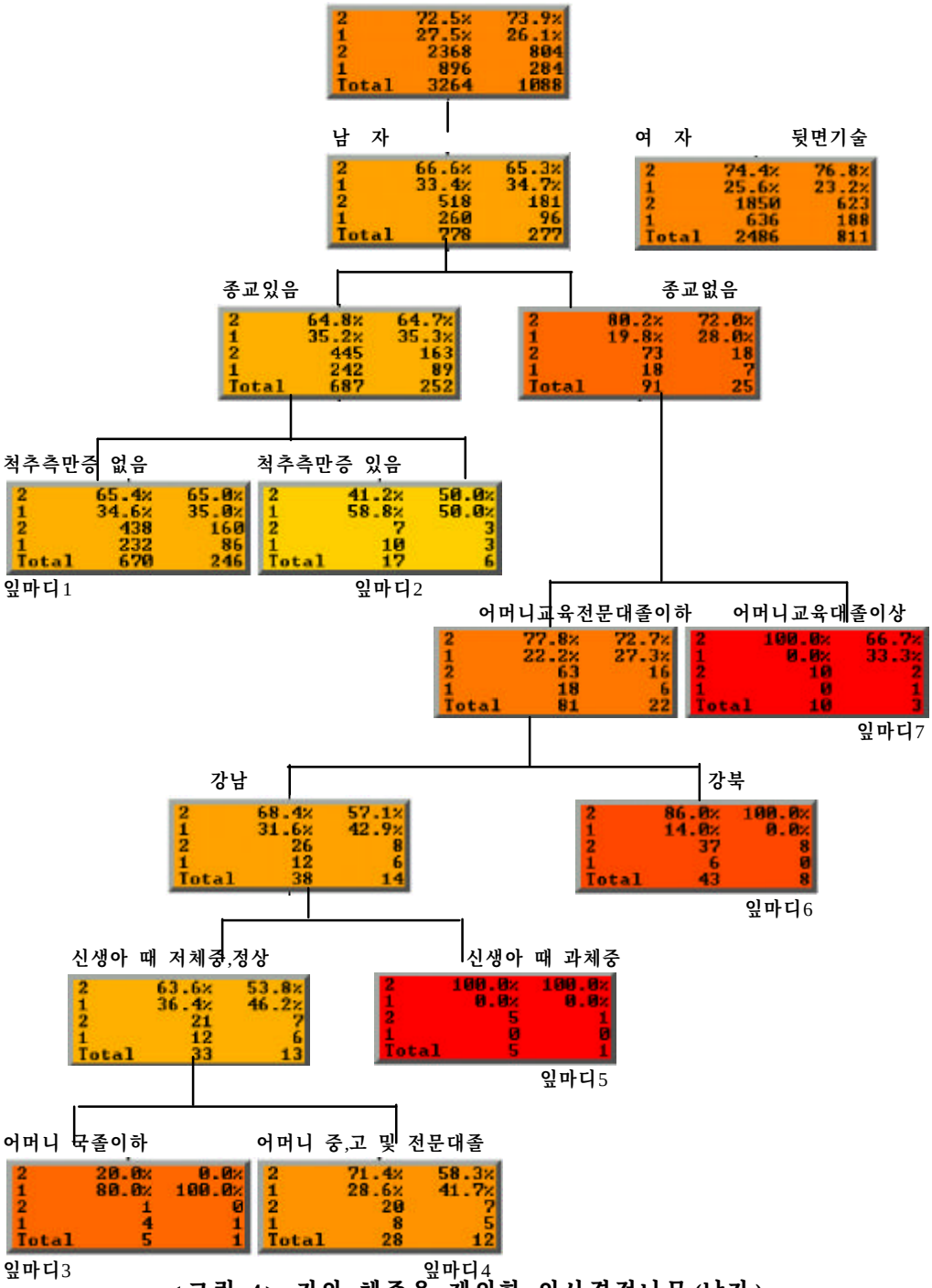
잎마디 4를 보면, 키 157.5cm 이상에서 체중 46.5kg 미만은 100% 모두 저체중이었다. 잎마디 6을 보면, 키가 161.5cm미만에서 체중 47.5kg이상이면 저체중은 0%이었다. 잎마디 7과 잎마디 8을 보면, 키가 161.5cm이상에서 체중이 46.5kg이상 50.5kg미만인 경우 저체중은 79.9%이었으며, 50.5kg이상에서 저체중은 11.9%이다(참조 그림 3).



<그림 3> 키와 체중을 포함한 의사결정나무

2) 2단계 분석

본 연구에서는 저체중의 분류기준을 밝힘에 있어 저체중에 영향을 미치는 극소수이지만, 중요한 요인을 알아보기 위해 2단계 분석을 시도하였다. 즉 데이터 마이닝을 실제로 활용하기 위해서는 소수의 관심대상을 분류하는 모형을 도출하여야 할 것이고, 이를 위해 본 연구에서는 다수를 차지하는 키와 체중 변수를 통제한 상태에서 2차 모형을 구축하였다(참조 그림 4, 5).



<그림 4> 키와 체중을 제외한 의사결정나무(남자)

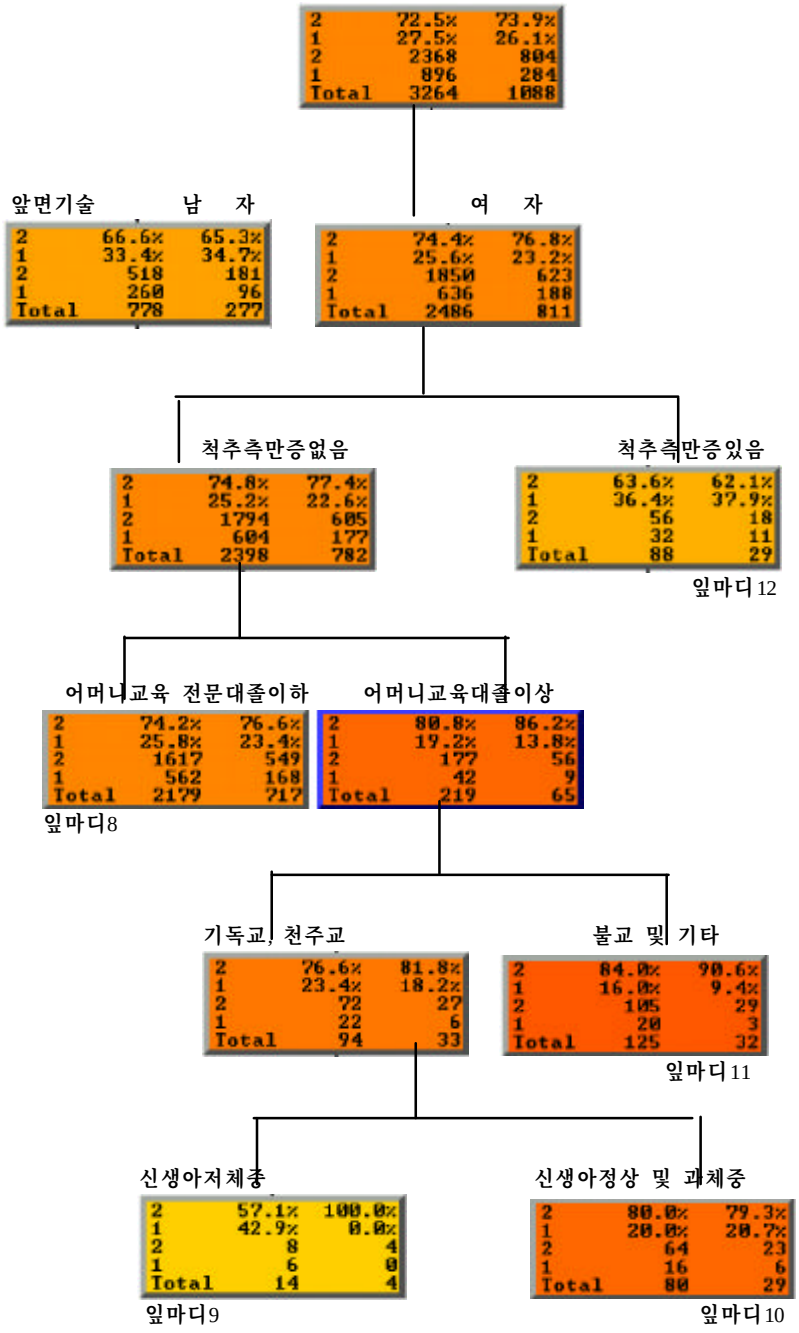


그림 5 키와 체중을 제외한 의사결정나무(여자)

전체 4,352명을 3대 1로 분석용(train set)과 평가용(validation set)으로 나누었으다.

분석용 3,264명중에 저체중은 896명으로 27.45%(base line gain %)였다(참조 그림 4). 이중 남자 770명 중에서 저체중은 33.4%로서 여자2,486명중 저체중 25.6%보다 높았다.

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 대졸이상인경우의 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 국졸이하경우 현재 저체중 청소년은 80%로서 어머니교육이 중,고등학교. 및 전문대졸인경우의 청소년 저체중 28.6%보다 높았다.

여자에서 보면 척주측만증이 있을 때 저체중은 36.4%로서 척추측만증이 없을 때의 25.2%보다 높았다(참조 그림 5). 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 저체중 은 42.9%로서 신생아때 정상체중 및 과체중인경우의 현재 저체중 20.0%보다 높았다.

또 각 잎마디에는 전체 집단을 파악하거나, 전체집단을 세분화(Segmentation)하는 정보를 알 수 있는데, 이를 일목요연하게 정리하기 위하여 표 17을 제시하였다.

표17 CART모형에 의한 마디내 이익도표

Leaf No.	Leaf Node Analysis						Cumulative analysis					
	Node :n	Node %	Resp :n	Resp %	Gain %	Index %	Node :n	Node :%	Resp :n	Resp :%	Gain :%	Index :%
3	5	0.15	4	0.45	80.00	291.4	5	0.15	4	0.45	80	291.4
2	17	0.52	10	1.12	58.82	214.3	22	0.67	14	1.56	63.6	231.8
9	14	0.43	6	0.67	42.86	151.1	36	1.1	20	2.23	55.6	202.4
12	88	2.7	32	3.57	36.36	132.5	124	3.8	52	5.8	41.9	152.8
1	670	20.5	232	25.9	34.63	126.2	794	24.33	284	31.7	35.8	130.3
4	28	0.86	8	0.89	28.57	104.1	822	25.18	292	32.59	35.5	129.4
8	2179	66.8	562	62.7	25.79	93.95	3001	91.94	854	95.31	28.5	103.7
10	80	2.45	16	1.79	20	72.86	3081	94.39	870	97.1	28.2	102.9
11	125	3.83	20	2.23	16	58.29	3206	98.22	890	99.33	27.8	101.1
6	43	1.32	6	0.67	13.95	50.82	3249	99.54	896	100	27.6	100.5
5	5	0.15	0	0	0	0	3254	99.69	896	100	27.5	100.3
7	10	0.31	0	0	0	0	3264	100	896	100	27.5	100

<표 17>에서 Gain(%)값은 마디내 목표 범주의 개체의 수/마디내 총 개체의 수로서 해당 잎마디에서의 저체중의 비율을 의미한다. Index(%)는 잎마디에서의 목표범주의 비율/전체에서의 목표범주의 비율로서 Index(%)가 높을수록 전체 저체중 청소년의 비율에 비해 해당 잎마디에서의 저체중 청소년의 비율이 높음을 의미한다.

gain%와 index%를 높은 마디부터 살펴 보면, 잎마디 3번은 남자 중에서 종교가 없고, 강남지역의 학교를 다니고 있고, 신생아 때 저체중 또는 정상으로 태어나고, 어머니교육이 국졸인 경우 현재 저체중의 gain값은 80%이고, index값은 291.44%로 가장 많았다. 즉 전체청소년 중 저체중 청소년의 27.45%(baseline gain %)보다 3번 잎마디의 대상자의 저체중 비율은 80.0%(gain %)이고, 이 80%의 gain%를 27.45%의 baseline gain%로 나누어 준 값이 index%로서 291.44%(기준에 비해 2.91배)이었다.

다음으로 잎마디 2번은 남학생 중 종교가 있고 척추측만증이 있는 경우 저체중의 gain%는 58.82%이고 index 값은 214.28%(기준에 비해 2.14배)였다.

세번째로는 잎마디 9번으로 여자이면서 척추측만증이 없고 어머니교육이 대졸이상이면서 기독교 또는 천주교이고 신생아때 저체중으로 태어난 경우 gain값은 42.86%, index값은 151.14%였다. 하지만 상위 3번,2번,9번 잎마디의 누적index는 202.4%이나 누적마디빈도 1.1%이고, 누적response%는 2.23%에 불과하였다.

상위 4번째 마디는 잎마디 번호 12번으로 여자 중 척추측만증이 있을때 저체중의 gain값은 36.36%, index값은 132.46%이었다.

상위 5번째 마디는 잎마디 번호 1번으로 남자이고, 종교가 있고, 척추측만증이 없을때 저체중 gain값은 34.63%, index값은 126.16% 이었다.

이익도표에서 상위 4개 잎마디 끝만 열거해보면 3번 어머니 국졸이하, 2번 남자 척추측만증있음, 9번 신생아때 저체중, 12번 여자척추즉만증 있음이었다. 이익도표에서 상위 4개 마디의 누적 gain %는 41.94%이었고 누적 index%는 152.79%이었다. 즉 CART모형에서 나타난 저체중 청소년의 분류특성은 어머니 교육, 척추측만증, 종교, 신생아때의 건강상태임을 알 수 있었으며, 이중 어머니교육상태, 척추즉만증, 종교는 일반통계분석과 일치하고 있었고, 신생아때의 건강상태는 의사결정나무 분석에서 새롭게 발견된 패턴이었다. 이는 데이터 마이닝 분석이 일반통계분석보다 가급적 정보를 버리지 않고 많이 사용하여 좀더 많은 정보를 준다는 특징때문 일 것으로 생각한다.

마디내 이익도표를 구하기 위한 공식은 다음과 같다(참조 표 18)

<표 18> 이익도표 공식

잎마디(n) = 잎마디에 속하는 전체 개체의 수

$$\text{잎마디:\%} = \frac{\text{잎마디에속하는개체의수}}{\text{전체개체의수}}$$

Resp(n)= 잎마디에 속하는 목표 범주의 개수

$$Resp:(\%) = \frac{\text{잎마디에속하는목표범주의개체의수}}{\text{전체에서목표범주의개체의수}}$$

$$\text{Gain}(\%)\frac{\text{마디내목표범주의개체수}}{\text{마디내총개체수의수}}$$

$$\text{Index}\ (\%) = \frac{\text{잎마디에서의목표범주의비율}}{\text{전체에서의목표범주의비율}}$$

4. 연관성 규칙(Association rule)을통한 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[5]을 살펴보아야 할 것이다. 이런 확률을 지지도(Support)라고 한다.

또 저체중 대상자와 다른 변수가 동시에 있을 조건부 확률[6]을 살펴보아야 한다. 이런 확률을 신뢰도(Confidence)라고 한다.

향상도(Lift)[7]는 1에 가까우면 독립에 가까운 사건, 1보다 크면 연관관계로 판단하면 될 것이다. 따라서 의미있는 연관성 규칙이 되려면 향상도 값이 1이상이 되어야 할 것이다.

이에 따라 본 연구에서는 연관성 규칙의 자료를 신뢰도, 지지도, 향상도가 높은 순으로 정렬시킨 후 그중 가장 높은 순위를 차지한 25개의 연관성 규칙을 제시하였다(참조 표 21).

---

5) $Pr(A \cap B) / N$
6) $Pr(A \cap B) / P(A)$
7) $Pr(A \cap B) / Pr(A)Pr(B)$

<표 19> 저체중일때의 연관성규칙(Association rule)

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

연관성 규칙을 통한 관련성 탐색결과를 보면 강북지역의 학교를 다닐 때, 출생시 체중이 정상이었으며, 여자이고, 저체중일 확률은 연구대상 전체 학생 중 10.53%를 차지하고있었고, 저체중이면서 출생시 체중이 정상이고 여자일 때 강북지역 학교에 다닐 확률은 100%이다. 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 309명의 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 선정 후 건강중재방안을 제시한다면, 시간과 경비 면에서 가장 효율적인 건강증진 전략이 가능할 것이다.

또 연관성 규칙을 의사결정나무 분석과 비교해 볼 때, 의사결정나무분석 역시 잎마디에서의 Gain 값과 Index값을 가지고, 기준값에 비해 얼마나 많은 분포를 보이고 있는지를 알 수 있었다. 하지만, Gain값이 높은 잎마디의 대부분은 Response %, 즉 해당대상자 수가 적다는 것에 비교해 볼 때, 연관성 규칙은 커다란 집단의 대상에서, 현실적이며, 효과적인 목표집단을 찾을 수 있다는 장점을 가지고 있었다.

이처럼 저체중 청소년은 가족지지, 스트레스, 자아존중감, 우울감, 출생시 체중 등과 같은 변수와 서로 관련이 있었고, 이는 강북지역 또는 척추측만증과 신뢰도가 높음으로 저체중 청소년의 건강증진 프로그램 수행에 있어, 강북지역 청소년과 척추측만증을 가진 학생을 대상으로 할 때 효과적일 것임을 알 수 있었다.

Ⅳ. 고 찰

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(Rikimaru, 1998)라고 하였다. 본 연구결과에서도 출생시 저체중은 과체중에 비해 1.13배로 저체중이 많았으나, 이는 일반통계 부분에서는 통계학적으로 유의하지 않았다(P value 0.88). 그러나 의사결정나무분석의 CART 모형에서는 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(1998)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 청소년 보다, 저체중 청소년이 오히려 긍정적 경제상태로 응답하는 경향을 보였고, 또 강남학생에게 저체중 학생이 더 많은 것으로 볼 때, 외국의 선행연구와는 다소 다른 결과를 보여주는 듯 했다. 그러나 데이터 마이닝의 연관성규칙에서는 저체중과 관련된 신뢰도가 높은 상위 25개의 연관성 규칙에서 저체중학생과 강남지역 학생과의 연관성 규칙이 하나도 없는 반면, 강북학생과의 연관성 규칙은 12개를 차지하는 것으로 보아, 경제적 상태와 저체중은 서로 연관이 있는 것으로 추정되어지므로, 추후 연구에서는 삶의 수준정도와 저체중과의 관련 등의 심도있는 연구가 필요할 것이다.

WHO(1998)의 '여성뿐 아니라 남성의 건강도 저체중아 출생에 영향을 미침을 알 수 있다'는 발표와 본 연구의 결과를 비교해 볼 때, 본 연구의 저체중 청소년

보다 정상체중 청소년의 부모가 더 당뇨병의 병력이 많음을 알 수 있었고, 기타 다른 가족력과 저체중과는 통계학적으로 서로 관련이 없었다.

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. 반면, 알레르기, 감기, 두통, 우울에 있어서는 저체중 청소년이 정상체중 청소년보다 덜 경험한 것으로 분석되었다.

데이터 마이닝의 연관성 분석에서는 부족한 가족지지, 과도의 스트레스, 부족한 자아존중감, 과도의 우울감 등의 심리사회적 요인을 가지고 있는 저체중 청소년은 강남보다는 강북과 또 신체적 건강문제 중 척추측만증과 서로 신뢰도가 높음을 알 수 있었다. 이런 결과는 이태화와 신선미(2001)의 연구에서 고등학생의 흡연, 음주, 식습관, 운동, 스트레스 관리등의 건강증진행위 실천의 지역간 비교에 있어 강남의 학생이 강북의 학생보다 전반적인 건강증진 행위 실천을 더 잘하고 있다(P =0.00)는 결과와도 그 맥락을 같이 하는 것으로 여겨진다. 그러므로 저체중 청소년의 건강증진 프로그램 수행에 있어, 전체집단을 획일적으로 접근하는 것 보다 강북지역 청소년과 척추측만증을 가진 학생을 대상으로 프로그램을 강화할 때 효과적일 것임을 알 수 있었다.

저체중에 관한 국낸 연구를 보면, 저체중 출생에 대한 연구는 다루어 졌으나(조남훈, 1997), 신생아 외의 저체중에 관한 연구는 극히 제한적으로 실시되었고, 특히 청소년기의 저체중에 관한 연구는 없어서 본 연구의 결과를 국내연구와 비교하기는 힘들다. 다만, 1998년 실시된 국민건강영양조사의 결과를 보면 전반적으로 우리 나라는 체중에서 정상이 68.7%, 저체중이 5.1%, 과체중이 23.9%, 비만이 2.4%로 비만과 저체중이 모두 존재하고 있음을 알 수 있었으나, 성별, 연령별 세분화된 분석의 정보는 불충분하였다. 또 조사항목이 질병을 위주로 선택되어서, 인구, 심리, 사회학적 측면을 고려하지 않음을 알 수 있었다.

그러나, 본 연구를 진행하면서 숙고가 필요했는데, 그 첫 번째는 성장발육기의

청소년의 경우 본 연구에서 사용한 '아시아 표준 체중 기준'이 적합한 것인가의 의문이었다(최중명, 2001). 즉 WHO나 임상비만학회는 각기 다른 표준 체중 기준치를 제시하고 있으며, 외국의 저널에서도 연구자마다 각기 다른 기준으로 체중의 등급을 분류하고 있어서 비만도를 산출하는 표준체중치를 어떠한 자료를 사용하는 가에 따라 저체중, 정상체중 및 과체중에 대한 분류가 달라지는 것을 볼 수 있었고, 이는 연구결과에도 큰 영양을 미치고 있었다. 그러므로 BMI 기준을 17.5로 낮춰 '심한 저체중군'을 경계체중군을 제외한 '정상체중군'과의 비교를 시도한다면, 본 연구보다 좀더 많은 결정적 변인을 도출할 수 있을 것으로 추정된다.

둘째, 본 연구는 데이터 마이닝 기법을 활용하여 최적의 저체중 분류규칙을 시도하였다. 그 중 의사결정나무분석은 연구결과를 사용하는 의사결정자가 쉽게 이해할 수 있고, 시각적 효과로 해석이 용이하다는 점, 일반통계에서 의의가 없는 정보를 버리지 않고 가급적으로 많이 사용한 분류를 보여준다는 장점이 있는 반면, 분리기준, 정지규칙, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무가 형성되어 연구자의 주의가 요구되었다. 따라서 실제로 단 한번만에 분석이 끝나지 않고, 연구자는 필요한 과정을 반복적으로 수행하여 실제 적용에 따르는 문제점을 비교 검토하여야 하는 노력이 필요하다.

또 보건의료계의 소수를 차지하지만, 중요한 변인을 밝혀내기 위한 연구에서는 다수를 차지하는 변인에 의해 가려 밝힐 수 없다는 단점도 있었으며, 이는 연구자의 분석 및 통계적 통찰력과도 직결되는 문제였다.

데이터 마이닝의 이런 문제에 대해 Buja와 lee(1999)는 이와 같이 소수의 관심이 대상이 되는 부류를 찾아내기 위한 단측 순수성(One side purity)나무모형 생성방법을 소개하였고, Friedman과 Fisher의 PRIM(Patient Rule Induction Method)에서는 단측 순수성과는 달리 사용자가 입력한 비율만큼 각 변수의 극단 부분을 제거해 가면서 목표변수의 수준을 탐색해 가는 방법을 소개하였다. 본 연구도 그 예외가 아니어서 저체중에 영향을 미치는 소수 변인을 밝히기 위해 2단계에 걸쳐 분석을 시도하였다.

또 데이터 마이닝 분석에 있어 또 생각해야 할 조건은 다른 어떤 통계방식보다 사전계획과 양질의 데이터 준비가 무엇보다 중요하다는 것이다. 오류가 있는

데이터에는 아무리 훌륭한 방법을 적용한다고 해도 얻어진 결과가 만족스럽지 않을 수 있으며, 데이터 정제하는 시간과 경비가 전통적 통계방식 보다 더욱 많이 들기 때문이다(김지은, 2001).

이런 면에서 볼 때 본 연구의 제한점은 첫째, 데이터 마이닝을 염두에 두고 장기간에 걸쳐 계획된 데이터가 아니고, 둘째, 집단검진의 설문에 의해 대상자의 성의있는 응답이 다소 제한 되어있을 수 있으며, 무응답의 비율이 높다라는 점을 들 수 있다. 그러나 본 연구의 데이터는 강북과 강남의 지역의 안배, 학교별 종류, 심리.사회적 변수의 고려, 학교인구의 표본 대표성을 고려한 14개 학교의 큰 Sample size, 데이터 마트내의 저체중과 정상체중의 전수 조사라는 점에서 데이터의 질(quality)은 높을 것으로 생각한다.

또 본 연구결과 어머니 교육, 척추측만증, 뇨당, 종교등의 관련변인을 밝힐 수 있었고, 연관성 규칙에서는 스트레스, 우울감, 자아존중감, 가족지지등의 사회심리적 변수와의 관련성을 탐색할 수 있었다는 점, 국내 연구에서는 없는 저체중의 변인을 밝히기 위한 시도라는 점, 학교보건학에서 아직 이용한 적이 없는 데이터 마이닝을 응용한 방법이라는 점, 전통적 통계방법과 데이터 마이닝을 비교하여 유용성을 비교해 보았다는 점에서 작은 의의를 찾고자 하며, 본 연구에서 아직 밝혀지지 않은 변인들 즉 날씬함에 대한 사회적 압박과 영양상태와 식이습관에 대해서는 후행 연구자의 몫으로 남기고자 한다.

Ⅴ. 결　론

본 연구는 우리나라 청소년 중 많은 분포를 차지하고 있는 저체중의 현황과 특성을 파악하여 건강증진 방향을 모색하고자, 빈도분석, 차이검정, 로지스틱회귀분석, 데이터 마이닝의 의사결정나무분석, 연관성분석을 통해 저체중 청소년의 건강관련 특성을 파악하였다.

그 결과는 다음과 같았다.

첫째, 저체중 청소년의 분포(BMI 18.5 미만)는 전체 연구 데이터 마트 5,188명 중 22.74%를 차지함으로서, 현재 보건의료계의 주 관심인 비만(BMI 25이상)청소년의 6.96% 보다 더 많은 분포를 차지하고 있었다.

둘째, 여학생보다 남학생에게, 강북보다는 강남지역의 학교에 저체중 청소년이 각각 1.52배, 1.15배 더 많은 빈도를 차지했으나, 건강문제를 가진 저체중 청소년과의 관련성을 알기위한 연관성 분석에서는 남학생보다는 여학생이, 강남보다는 강북이 더 많은 연관규칙이 있음을 알 수 있었다.

셋째, 정상체중 청소년보다 저체중 청소년이 더 많이 갖고 있는 신체 건강문제는 척추측만증과 뇨당으로서, 척추측만증은 없을 때 보다 있을 때 1.7배, 뇨당검출이 없을 때 보다 있을 때에 1.68배로 저체중이 많았다.

넷째, 인구학적 측면에서는 어머니 교육정도가 대졸일 때, 국졸이하에 비해 0.52배로 저체중 청소년이 적음을 알 수 있었고, '종교가 없는 학생'에 비해 '기독교'인 학생이 1.14배로 저체중이 많음을 알 수 있었다.

다섯째, 신생아때의 체중이 저체중 청소년을 분류할 수 있는 특성임을 의사결정나무 분석을 통해 알 수 있었다.

그러므로 향후 저체중 청소년의 건강증진을 위한 미시적 접근은 척추측만증, 뇨당검출에 대한 신체건강문제에 대한 관리가 필요할 것이며, 스트레스, 자아존중

감, 우울감, 가족지지 등의 건강문제를 가진 저체중 청소년은 여학생과 강북지역에 더 많았으므로 저체중 청소년의 건강증진을 도모함에 있어, 여학생과 강북청소년에게 좀더 집중적인 관심과 연구 근거에 기반을 둔 구별된 건강증진전략이 필요할 것이다.

또 어머니의 교육정도와 출생시의 건강상태는 저체중 청소년에 영향 미치는 변인이므로 건강한 체중인식에 대한 보건교육이나 홍보 등의 거시적 건강증진 정책을 통한 중재가 요구된다.

본 연구의 제한점 및 제안점을 살펴보면, 첫째, 저체중(기준-BMI 18.5미만)과 정상체중(기준-BMI 18.5이상 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 데이터 마이닝의 의사결정분석은 로지스틱 회귀모형보다 저체중 분류에 있어 정확률이 높았고, 여러 특성이 결합된 규칙을 도출할 수 있는 등의 장점을 가지고 있다. 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.

셋째, 본 연구에서 날씬함에 대한 사회적 압박과 영양상태 및 식이습관과 저체중과의 관계에 대한 분석이 누락되었으므로, 이에 대해서는 후행 연구가 필요할 것이다.

참고문헌

강현철, 서두성, 최종후(1998), Enterprise Miner의 의사결정나무분석 알고리즘, SUSI-K 발표논문집 제 9권, 169-186.

강현철, 최종후외(1999), 데이터마이닝-방법론 및 활용, 자유아카데미.

국민건강.영양조사 심층분석 결과 발표(2001), 한국 지질.동맥경화학회지 Vol. 11, 3, 2001:408-410.

김미영(2000), 다중 분류앙상블 모형을 위한 개인 신용 평가 시스템에 관한 연구, 연세대학교 대학원 응용통계학과 석사학위 논문.

김수택(2001), 신용평점모형의 평가와 기각 추론의 적용, 고려대학교 대학원 정보통계학과 석사학위 논문.

김용대, 최대우(2000), 단측 순수성에 의한 나무모형의 성장에 대하여, 한국지능정보시스템학회, 제 7권 제 1호, pp 17~25.

김지은(2001), 데이터 마이닝을 성공적으로 구현하기 위해 고려해야 할 요건, http://dept.stat.pusan.ac.kr/info/DataMining/1-3.htm.

남철현 등(1984): 인구.보건지표 및 통계. 한국인구보건연구원, p.24.

로이터통신, 1999년 12월 10일기사 http://dailynews.yahoo.com/headlines/hl

문현경(2001), 국민건강영양조사로 본 한국인의 영양상태, 한국 지질.동명경화학회지, Vol. 11, No. 3, 2001.

박용수(2001), 고지혈증의 역학적 특성 및 심혈관 질환의 위험요인 분석, 한국 지질.동맥경화학회지 Vol. 11, 3, 2001:402-404.

여운승(2000), 사회과학과 마케팅을 위한 다변량 행동조사, 민영사, pp 314-390.

용왕식(2000), 의료보험 지식경영시스템 구축방안에 관한 연구, 연세대학교 보건대학원.

유지혜(1999), 판별분석에서 분류나무 모형과 판별함수의 비교연구, 연세대학교 대학원 응용통계학과 석사학위 논문.

이태화, 신선미(2001), 서울시 고등학생들의 건강증진행위 실천에 관한 연구, 지역

사회간호학회지 제 12권 2호;459-467.

이희우, 신선미(1999), 서울시 12개 고등학교 1학년 종합검진 통계조사 분석, 서울특별시 학교보건원 연보 제 29호:81-90.

이희우, 신선미(1999), 고1 종합검진 결과 이상장에 대한 추후관리사업분석, 서울특별시 학교보건원 연보 제 29호:91-95.

이희우, 신선미(1999), 고등학교 1학년 학생의 건강관련 의식 및 행위에 대한 조사, 서울특별시 학교보건원 연보 제 29호:96-106.

이희우, 장철환, 신선미(2001), 교직원 건강상태와 생활습관에 관한 연구, 서울상록과학학술재단지.

조남한(2001), 당뇨병의 역학적 특성 및 위험요인 분석, 한국 지질.동맥경화학회지 Vol. 11, 3, 2001:396-401.

최국렬 조대현외(2000), 데이터 마이닝 이론과 실습-보건의료 데이터 중심, 청구문화사.

최중명(2001), 성별 연령별 체위 표준치 분석 및 신체계측자료의 표준화 연구, 한국 지질.동맥경화학회지, Vol. 11, No. 3, 2001:405-407.

한겨레 신문 2000년 1월 25일 기사. http://www.hani.co.kr.

허갑범(1999) 한국인 제2형 당뇨병의 병인론적 이형성, 대한당뇨병학회지, 제23권 1호, 62.

홍문식 등(1985), 전국출산력 및 가족보건실태조사, 한국보건사회연구원, P 187.

Behrman RE, Kliegman RM, Arvin AM (eds)(1995), Nelson textbook of pediatriacs, 15th ed. W.B. Saunders Company, Philadelphia, pp.169-172.

Borowitz D(1996), The interrelationship of nutrition and pulmonary function in patients with cystic fibrosis, Curr Opin Pulm med Nov;2(6):457-461.

Brejman, L., Fredman, J.H., Olshen, R.A., and Stone. C. J.(1983) Classification and regression trees, Wadsworth International, Belment, CA.

BuJa, A., and Lee, Y-S(1999), Data mining criteria for tree-based regression

and classification, http://www research. att. com/~andreas/papers /trees.ps.,gz.

Chae YM(2,000) Data Mining for Knowledge Management in a Health Insurance Domain, International Journal of Informatics 000(2001)000-00.

Friedman, J.H., and Fesher, N. I.(1997), Bump hunting in high-demensional data. Tech. report, Dept. of Statistics, Stanford University.

Guoqiang Zhang, Michael Y. Hu, B. Eddy Patuwo, Daniel C. Indro(1999), Artificial neural networks in bankrupcy prediction; General framework and cross-validation analysis, European Journal of Operational Research 116(1999) 16-32.

Herzog W(1997), Medical findings and predictors of long-term physical outcome in anorexia nervosa: a prospective, 12-year follow-up study, Psychol Med Mar;27(2):269-279.

Hongkyufo and Ingoo Han(1996),Integration of Case-Based Forecasting, Neural Network, and Discriminant Analysis for Bankruptcy Prediction, Pergamon PII S0957-4174(96)00056-5.

Katzman DK(1997), A longitudinal magnetic resonance imaging study of brain changes in adolencents with anorexia nervosa, Arch Pediatr Adolesc Med Aug;151(8):793-797.

Kun Chang Lee(1999), Hybrid neural network models for bankruptcy predictions, Decision Support System 18(1996) 63-72.

Martinchik AN(1997), Monitoring food comsumption and nutritional status of Moscow schoolchildren in 1992-1994. 2. Anthropometric evaluation of nutritional status, effect of social factor, on the character and status of nurtition, Vopr Pitan ;1: 3-9.

Paricio JM(1998), Health examination of children from the democratic Sahara Republic(North West Africa) on Vacation in Spain, An Esp pediatr jul;49(1):33-38.

Rikimaru T(1998), Risk factors for the prevalence of malnutrition among urban children in Ghana, J Nutr Sci Vitaminol(Tokyo) June;44(3):391-407.

Swenne I(1999), Heart risk associated with weight loss in anorexia nervosa and eating disorders : risk factors for QTc interval prolongation and dispersion, Acta paediatr, Mar;88(3):304-309.

Word Health Organization(1992) : Low birth weight - a tabulation of available information, Geneva, 1992.

Zaman MM(1998), Nutritional factors associated with rheumatic fever, J Trop Pediatr Jun:44(3):142-147.

ABSTRACT

Applying Data Mining Techniques to Promote the Health of Underweight Adolescents

Shin, Sun mi

Dept. of Nursing

The Graduate School

Yonsei University

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 first study made a data mart of 5,188 adolescents, who had answered questionnaires well among 10,300 10th grade students in Seoul, 2000.

The second study sampled (n=4352) 1,180 underweight(BMI <18.5) and 3,172 average weight(18.5<=BMI<23) adolescents, and investigated the differences between two groups by frequency, chi-square, and simple logistic regression analysis for general characteristics. Related variables derived from previous step and risk factors known by previous overseas studies were input in a decision tree and an association rule of SAS E-Miner.

Further more, to find the most suitable and predictable model, CHAID, C4.5, CART, and logistic regression model were assessed through a cross-validation method. The most predictable model was CART.

The results of the study were as follows:

Underweight adolescents(BMI<18.5) consisted of 22.74% of the 5,188, and the obese adolescents(BMI>=25) only consisted of 6.96% of the 5,188. Surprisingly from these numbers, the obese adolescents currently have more medical interest than the

underweight adolescents.

In simple logistic regression which has target variables of underweight to normal weight adolescents, the odds ratio of male high school students was 1.52 times over the females, and the odds ratio in the living region south of the Han-river was 1.15 times over the north. Scoliosis odds ratio was 1.7 times over non-scoliosis, and odds ratio of urine sugar positive cases was 1.68 times over the negative cases. In demographics, however, the odds ratio of a student's mother's education as a university graduate was 0.52 times over elementary graduates. Odds ratio of christians was 1.14 times over non-christians.

In decision tree analysis by CART model, there was a classification tree formated by gender, religion, living region, scoliosis, and neonatal weight as risk factors. It was the same as general statistics except for neonatal weight. However, there were more detailed classification characteristics in low body weight adolescents through each leaf. Underweight adolescents were 100% in males, without religion, living south of the Han-river, low or normal neonatal weight, and with a mother's education of elementary graduate or below.

Finally in frequency, the underweight rate was higher 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 the Han-river, and scoliosis.

Therefore, approaches for health promotion of underweight adolescents is not only intervention of physical health, but also the education of proper weight perception to prevent low birth weight and underweight adolescents because a mother's education and child's low birth weight are related to underweight adolescents. In conclusion, we suggest to select these target groups: females, students who are north of the Han-river in Seoul, and adolescents with scoliosis.

Key words : underweight, data mining, health promotion

부록-연구도구

신체검사 결과기록지 및 설문지

학교명 :　　고등학교 :

성　명 :　　성　별 : 남　여

■ 신체검사결과

계측 측정자 기록

신 장 :　　cm

체 중 :　　kg

병리 검사자 기록

요　　당 :　+　++　+++　+++

요 단 백 :　+　++　+++　+++

혈 색 소 :　　g/dl

식 전 혈 당 :　　mg/dl

총콜레스테롤 :　　mg/dl

X선 판독자 기록

폐결핵 의증 :

척 추 이 상 :

■ 설 문 지

1. 얼마전 우리나라는 국제금융위기(IMF파동)가 있었습니다. 학생의 가계수입은 IMF파동시보다 어떻게 변화했나요?

① 더 나빠졌다

② 똑같다

③ 더 나아졌다

2. 주위친구들과 비교 할 때 가족의 수입은 어느정도입니까?

① 적다

② 거의 비슷하다

③ 많다

3. 학생의 종교는 무엇입니까?

① 기독교

② 불교

③ 천주교

④ 무교

⑤ 기타

4. 부모님의 교육정도는 어느정도인지 아래 네모안에서 해당번호를 적으세요.

부 _____ 모 _____

① 무학	② 국졸	③ 중졸
④ 고졸	⑤ 전문대졸	⑥ 대졸
⑦ 대학원 졸		

5. 학생이 태어났을 때(신생아)의 건강상태는?

① 약했다고 한다

② 정상이었다

③ 몸무게가 많이 나갔다고 한다

6. 학생에게 지난 1년부터 현재까지 있었던 건강문제에 대해 있는 대로 고르세요

① 감기, 독감

② 두통, 알레르기

③ 구토, 설사

④ 알레르기

⑤ 코피

⑥ 빈혈

⑦ 우울

⑧ 자살충동

⑨ 자살시도

⑩ 기타

7. 부모, 형제 중 다음과 같은 질환을 앓었거나 또는 앓고 있는 분이 있으면 있는 대로 고르세요

① 결핵

② 간염

③ 당뇨병

④ 중풍

⑤ 심장질환

⑥ 고혈압

⑦ 알콜중독

⑧ 기타

8. 지난 한 달간 학생은 담배를?

① 안 피웠다

② 가끔 피웠다

③ 매일 피웠다

9. 지난 한달간 학생은 술을?

① 안 마셨다

② 가끔 마셨다

③ 꽤 자주 마셨다

10. 학생은 평상시 생활 중에서 스트레스는 어느 정도 느끼시나요?

① 대단히 많이 느낀다

② 많이 느낀다

③ 중간정도 느낀다

④ 거의 없다

11. 다음의 물음에 대해 학생들이 느끼는 대로 빠짐없이 모두 체크하세요.

나는...	아니다	보통이다	그렇다
1. 괴롭거나 어려운 일이 생기면 가족들에게 도움을 얻을 수 있다.			
2. 가족들은 공동의 관심사에 대해 대화를 잘 나누고 함께 공동문제를 잘 해결해 나간다			
3. 가족들은 내가 새로운 일이나 제안을 할 때 이를 따라주고 도와준다			
4. 가족들은 서로에게 애정을 잘 표시하고 분노, 슬픔, 사랑등과 같은 나의 감정을 잘 받아준다			
5. 가족들은 함께 여가를 보내는 것에 대해 만족한다.			

12. 다음 각 문항에 해당되는 곳에 빠짐없이 체크하십시오.

나는...	전혀 아니다	아닌 편이다	그런 편이다	확실히 그렇다
1 내가 적어도 다른 사람만큼은 가치 있는 사람이라고 생각한다.				
2 내가 좋은 성품을 많이 갖고 있다고 생각한다.				
3 대체로 나는 내가 실패자라는 느낌이 쉽게 들곤 한다.				
4 대부분의 다른 사람만큼 일을 잘 할 수 있다.				
5 자랑할 것이 많지 않은 것 같다.				
6 나 자신에 대해 긍정적인 태도를 갖고 있다.				
7 대체로 나는 나 자신에 대해 만족하고 있다.				
8 내가 좀더 나은 사람이었으면 좋겠다.				
9 나는 내가 쓸모없는 사람이라고 확실히 느낄때가 있다.				
10 나는 가끔 내가 전혀 좋은 점이 없다고 생각한다.				

약 한달 전부터 돌이켜 볼 때 나는...	전혀 아니다	아닌 편이다	그런 편이다	매우 그런 편이다
1. 걱정 때문에 잠을 많이 못 잔다				
2. 계속해서 긴장감을 느낀 적이 있다				
3. 어려움을 잘 극복할 수 없다고 느낀 적이 있다				
4. 우울하고 불행하다고 느낀 적이 있다				
5. 자신감이 없었던 적이 있었다				
6. 나를 쓸모 없는 사람이라고 생각한 적이 있다				
7. 매사가 한꺼번에 나를 힘들게 한다고 느낀 적이 있다				
8. 매사에 올바른 결정을 잘 할 수 있다고 느꼈다				
9. 때때로 신경이 너무 쇠약해져서 아무 것도 할 수 없다				

☻ 끝까지 응답해 주셔서 감사합니다.